THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

A holistic view on transcriptional regulatory networks in *S. cerevisiae*: Implications and utilization

David Bergenholm

Department of Biology and Biological Engineering

CHALMERS UNIVERSITY OF SWEDEN

Gothenburg, Sweden 2020

A holistic view on transcriptional regulatory networks in *S. cerevisiae*: Implications and utilization

David Bergenholm

Gothenburg, Sweden 2020

ISBN **978-91-7905-211-9**

Löpnummer 4678

Doktorsavhandling vid Chalmers tekniska högskola

Ny serie ISSN0346-718X

Division of System and Synthetic Biology

Department of Biology and Biological Engineering

Chalmers University of Technology

SE-41296

Gothenburg, Sweden

Telephone + 46(0) 31 772 1000

Cover: Schematic representation of this thesis

Printed by Chalmers Reproservice

Gothenburg, Sweden 2020

A holistic view on transcriptional regulatory networks in *S. cerevisiae*: Implications and utilization

David Bergenholm

Department of Biology and Biological Engineering

Chalmers University of Technology

ABSTRACT

Life; perhaps it is bold to start an abstract with this powerful word, but this is where I will start. My research is at the heart of life. How can a single human cell proliferate to become bones, eyes, fingers and, finally, a human being? How can different cells containing the same set of DNA be so versatile? The answer lies within the regulation of genes. To build upon our understanding of gene regulation, I have studied gene transcription and especially transcription factors in a holistic, systems biology way using the model organism *Saccharomyces cerevisiae*. Translation from *S. cerevisiae* to humans will help us get both a fundamental understanding of the networks and engineer better cell factories.

Transcription factors play an essential role in transcription as they function to activate and suppress genes in response to stimuli. The transcription factors form transcriptional regulatory networks (TRNs), with intricate cross-talk and overlapping functions balancing the ability of the cells to react to stimuli but at the same time remain as steady as possible. This is a fine-tuned machinery that has a built-in safety feature of self-regulation if the system is perturbed in any way. We study the TRNs with state-of-the-art methods for transcription factor-DNA interaction: Chromatin Immunoprecipitation with exonuclease treatment or ChIP-exo for short. This method provides us with all the DNA interactions of a selected transcription factor at the nucleotide level and to what degree these interactions occurs.

To study these transcriptional regulatory networks, we put the yeast cells under nutrient starvation in fermentation systems. The fermentation system used is the chemostat, which enables a tight control on the environmental parameters, ensures a steady-state in the culture, and allows for high reproducibility. Ensuring that the cell culture is identical in-between runs is important since we can't study all transcription factors at the same time.

In this thesis, I present studies on transcription factors both individually, or as part of a bigger whole. We investigate stress response, NADPH generation, control over lipid and amino acid metabolism and the glycolytic pathway. Thanks to the different metabolic conditions used to study the transcription factors, we can both determine a core set of genes and genes that are specific for different conditions. We also employ statistical methods and regression models to understand and predict regulatory pathways. While doing so we discover novel functions and modularity and expand the transcriptional regulatory network for all studied transcription factors. We also constructed a multi-paralleled miniaturized chemostat-system to study these transcription factors in a high-throughput fashion. Finally, we have developed a toolbox for analysis of transcription factor data, including visual representation of the DNA binding, comparison of gene transcription and transcription binding between conditions and statistical methods for identifying regulatory pathways that can be used both for a fundamental understanding of TRNs and for better cell factory engineering.

List of Publications

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

I. **Bergenholm D**, Liu G, Hansson D, & Nielsen J (2019) Construction of mini-chemostats for high-throughput strain characterization. *Biotechnology and Bioengineering* 116(5):1029-1038.

II. Liu G, **Bergenholm D**, & Nielsen J (2016) Genome-wide mapping of binding sites reveals multiple biological functions of the transcription factor Cst6p in *Saccharomyces cerevisiae. mBio* 7(3):e00559- 00516.

III. Börlin CS, **Bergenholm D**, Holland P, & Nielsen J (2019) A bioinformatic pipeline to analyze ChIP-exo datasets. *Biology Methods and Protocols* 4(1):bpz011.

IV. **Bergenholm D***, Liu G*, Holland P, & Nielsen J (2018) Reconstruction of a global transcriptional regulatory network for control of lipid metabolism in yeast by using chromatin immunoprecipitation with lambda exonuclease digestion. *mSystems* 3(4): e00215-17.

V. Ouyang L, Holland P, Lu H, **Bergenholm D**, & Nielsen J (2018) Integrated analysis of the yeast NADPH-regulator Stb5 reveals distinct differences in NADPH requirements and regulation in different states of yeast metabolism. *FEMS Yeast Research* 18(8):foy091.

VI. Holland P, **Bergenholm D**, Borlin CS, Liu G, & Nielsen J (2019) Predictive models of eukaryotic transcriptional regulation reveals changes in transcription factor roles and promoter usage between metabolic conditions. *Nucleic Acids Research* 47(10):4986-5000.

VII. **Bergenholm D**, Börlin CS, Holland P, Nielsen J. 2019 T-rEx: A *Saccharomyces cerevisiae* transcription factor explorer. *Manuscript*

VIII. **Bergenholm D***, Dabirian Y*, Ferreira R*, Siewers V, David F, Nielsen J, Rational gRNA design based on transcription factor binding data. *Manuscript*

*Contributed Equally

Additional publications not included in this thesis.

IX. **Jullesson D**, David F, Pfleger B, & Nielsen J (2015) Impact of synthetic biology and metabolic engineering on industrial production of fine chemicals. *Biotechnology Advances* 33(7):1395-1402.

X. **Bergenholm D***, Gossing M*, Wei Y, Siewers V, & Nielsen J (2018) Modulation of saturation and chain length of fatty acids in *Saccharomyces cerevisiae* for production of cocoa butter-like lipids. *Biotechnology and Bioengineering* 115(4):932-942.

XI. Wei Y, Gossing M, **Bergenholm D**, Siewers V, & Nielsen J (2017) Increasing cocoa butter-like lipid production of *Saccharomyces cerevisiae* by expression of selected cocoa genes. *AMB Express* 7(1):34.

XII. Wei Y, **Bergenholm D**, Gossing M, Siewers V, & Nielsen J (2018) Expression of cocoa genes in *Saccharomyces cerevisiae* improves cocoa butter production. *Microbial Cell Factories* 17(1):11.

XIII. Börlin CS, Cvetesic N, Holland P, **Bergenholm D**, Siewers V, Lenhard B & Nielsen J (2019) *Saccharomyces cerevisiae* displays a stable transcription start site landscape in multiple conditions. *FEMS Yeast Research* 19(2):foy128.

XIV. Rajkumar AS, Liu G, **Bergenholm D**, Arsovska D, Kristensen M, Nielsen J, Jensen M. K, Keasling J. D (2016) Engineering of synthetic, stress-responsive yeast promoters. *Nucleic Acids Research* 44(17):e136.

Contribution summary

I. Conceptualized the study, carried out the experiments, analyzed the data and wrote the manuscript.

II. Participated in the conceptualization of the study, carried out parts of experiments, analyzed the ChIP-exo data and wrote parts of the manuscript.

III. Participated in the conceptualization of the study, wrote parts of the scripts, analyzed the data and wrote parts of the manuscript.

IV. Together with Co-author: Conceptualized the study, carried out the experiments, analyzed the data and wrote the manuscript.

V. Participated in the conceptualization of the study, carried out fermentation and RNA-seq experiments, analyzed parts of the RNA-seq data and wrote parts of the manuscript.

VI. Participated in the conceptualization of the study, carried out the experiments, analyzed the parts of the ChIP-exo data and wrote parts of the manuscript.

VII. Conceptualized the study, wrote the scripts, analyzed the data and wrote the manuscript.

VIII. Together with Co-authors: Conceptualized the study, carried out parts of the experiments, analyzed the data and wrote parts of the manuscript.

IX. Conceptualized the review, carried out literature search and wrote the manuscript.

X. Together with Co-author: Conceptualized the study, carried out the experiments, analyzed the data and wrote the manuscript.

XI. Participated in the conceptualization of the study, carried out parts of the experiments, analyzed the data and wrote parts of the manuscript.

XII. Participated in the conceptualization of the study, carried out parts of the experiments and wrote parts of the manuscript.

XIII. Participated in the conceptualization of the study, carried out experiments on fermentation and RNA-seq, analyzed parts of the RNA-seq data and wrote parts of the manuscript.

XIV. Participated in the conceptualization of the study, carried out experiments on fermentation and ChIP-qPCR, analyzed parts of the data and wrote parts 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 March 2014 and January 2020 at the Division of Systems and Synthetic Biology (SysBio) under the supervision of Jens Nielsen and co-supervised by Verena Siewers. This thesis was examined by Christer Larsson. This thesis was funded by The Novo Nordisk Foundation Center For Biosustainability and the Knut and Alice Wallenberg Foundation

David Bergenholm

January 2020

CONTENTS

To family and friends,

for making me the best person I can be.

A quick Hello from the Author

Welcome to reading my thesis. I'm glad to see that you made it this far. Some of you read it because you have to, some of you read it because you want to and some of you might do it because of both previous reasons. Before getting started I would like to point out some personality traits, that if you haven't already seen them, might be interesting to know. If you look at my publication list, you might discover three things. 1. There are rather many publications. This is because I love to talk to people and get involved in different projects and when I see that I can be a helping hand I do take the opportunity to say so, and that's how the numbers go up. 2. If you look at the author list there is actually quite few papers that I have written completely by myself, and "completely by myself" it should also be noted that nothing is by myself, someone always corrects, gives inputs and so on. This is because I believe that 1+1=5, thus meaning that the sum of the combined individual parts is MUCH greater than the sum of the individual parts alone. 3. I can't keep my hands out of the cookie jar, and I want many different cookies! As you also might see is that the topic of the publications both included in this thesis but also the papers not included are from different areas: Fundamental science, applied science, computational science, biology and technology. I really enjoy testing different areas and I do to be honest easily get bored if I have to do the same task for too long, variety is the key to my wellbeing.

It has been such a fantastic journey and I'm eternally grateful for being allowed to do all of these things.

1 A TALE OF THE CENTRAL DOGMA

The earth shook, and a loud rumble was heard. The earth shook again. A volcano in the distance just had an eruption, spewing out its ashes into the atmosphere. Ionized particles ignited the sky and thunder and lightning was all around, turning the night sky fiery red. Zaap! Lightning struck a puddle nearby. It was not the first time that lightning had struck this puddle, but this time something was different. Earth was unrest and in great pain, mother earth was in labor, about to give birth to something spectacular. In that very puddle a new construct was being created, unlike anything the universe had seen before. A molecule capable of self-replicating had seen the dawn of day, and life was formed.

The central dogma is the story of how the single most important way of storing information, RNA, started replicating itself and thus life was formed. This occurred some 3.5-4.2 billion years ago, probably even earlier. It did of course not just appear at once. It was rather a buildup of all the components, RNA, fat and protein, that at some point reached a critical concentration that in combination with an energy burst kickstarted life itself. Miller and Urey found in 1952 (Miller 1953) that most of the amino acids, lipids, sugars and some nucleotides are rather "easy" to form if the environmental conditions that earth was exhibiting in its youth are used. It has also been shown that fatty acids could help in building protein-like structures (Murillo-Sanchez et al. 2016) as well as catalyze the formation of RNA (Black and Blosser 2016). A simple type of RNA called proto-RNA can also be self-assembled from nucleotides if the right molecules are in close proximity, and these molecules might have been present in the early days of the earth (Cafferty et al. 2018). But why did life arise at all? As Erwin Schrödinger stated, "How can the events in space and time which take places within the spatial boundary of a living organism be accounted for by physics and chemistry?" We turn to the laws of physics, to be precise the second law of thermodynamics, stating that a system goes from order to disorder. Life exists because it can cause disorder better than spontaneous disorder. By taking a molecule from the surrounding and incorporating it, life actually increases order, but it gives energy in the form of heat to the surroundings, thus increasing the disorder in the system as a whole.

Proteins were formed based on the sequence that RNA was carrying, this allowed replication of RNA. RNA could convert between two structures, one that carries information and the other, ribosomes, that could read RNA. This system is simple and efficient, but mutations occur easily and so RNA evolved. Evolution generated the storing facility, DNA which was more stable to be able to keep the information intact. The central dogma, as we now are referring to occur in the following steps: i) replication: DNA is replicated, ii) transcription: DNA is transcribed to RNA iii) Reverse transcription: RNA is transcribed to DNA iv) replication: RNA is replicated and v) translation: RNA is translated into proteins (Figure 1). This is a much-simplified version of the process, but the central dogma holds true as a concept. In this thesis we will cover mostly one part of the central dogma and that is transcription, but to do so we need to dig deeper into the tale.

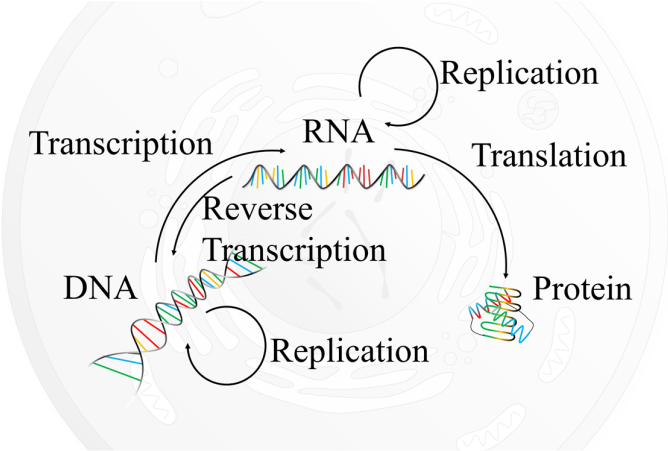


Figure 1 The central dogma of biology. DNA undergoes several stages of transformation: transcription to form mRNA and translation to form proteins. The DNA also needs to replicate itself to be able to be part of the dividing cells.

As RNA became the prominent way on Earth to increase disorder, evolution allowed it to be encased into cells. Probably this occurred to increase the probability of the stochastic events that allowed RNA to replicate and generate proteins to become more frequent as the encapsulation increased the concentration of molecules. The cells started to proliferate and became specialized into different tasks. To be able to tackle a continuously changing environment, a method to control the level of production of each protein was beneficial. Control of the transcription allows the cells to do just that, and maybe the RNA was the first transcription factors, controlling the gene expression in the form of riboswitches (Breaker 2012). By activating different genes in different conditions or at different levels the cell could both cope with an external changing environment and the internal environment. This way of protecting and adapting became very useful over the eons of time and at some point, even the cells became specialized in different tasks and soon multi-cellular organisms saw the light of day. We, humans, are one of evolutions finest creations, at least according to me. We have strength, endurance, flexibility, fine motoric skills, advanced hearing, tasting and seeing, and as we all know, we have the most powerful brain (that we yet know of) in the entire universe. This is thanks to the many different cell types that come together to form one entity.
Unfortunately, it is difficult to study the transcription of such enormously complex and slowly replicating system that is us humans. To scale it down and study transcription in a more efficient way we turn to our favorite model organism: *Saccharomyces cerevisiae.*

Saccharomyces cerevisiae, or the sugar loving (saccharo) fungus (myces), which makes beer (cerevisiae), has been used by humans since the Neolithic period for its great capability of turning carbohydrates into ethanol and carbon dioxide (Mortimer 2000) and the term enzyme meaning “in yeast” was coined by Kühne in 1877. The ethanol production is used for making beer and wine, and the yeast additionally provides some nice flavors in terms of esters to the beverage. In baking, the carbon dioxide helps to make the bread “fluffy” and the yeast also

helps to generate flavors and texture to the bread (Querol and Fleet 2006). *S. cerevisiae* has not only been used by humans for its great food and beverage production, it is also well studied in all omics fields (gen-, transcript-, prote-, metabol-,flux-, phen-) and was one of the first organisms to get its genome sequenced (Goffeau et al. 1996). *S. cerevisiae* is used in industrial settings as a cell factory due to its advantageous qualities of short generation time, high osmotic tolerance, broad range of pH tolerance, growth on complex and minimal media and as it is generally recognized as safe, holding the GRAS status (Hampsey 1997). The success of using yeast as a model organism is also due to the high degree of conservation of many key cellular processes between yeast and human cells, such as autophagy, protein translocation and secretion, heat shock and regulation hierarchies (Nielsen 2019). There is also a high degree of conservation between genes, as 47% of the 414 essential yeast genes can be replaced by their human orthologs (Kachroo et al. 2015). When it comes to engineering, *S. cerevisiae* is a good workhorse as it has a very efficient homologous recombination, which allows for integration of genetic fragments directly into the genomic DNA, which generates more robust engineered strains (Gietz and Woods 2001; Scherer and Davis 1979). These features also allow us to study proteins, and in my case transcription factors, in detail through various techniques.

The *S. cerevisiae* genome contains around 6300 genes and the genome size is around 12 million base pairs. However, only 9 million of these are protein encoding, while the remaining 3 million base pairs, or 25%, of the whole genome are used for other processes (Goffeau et al. 1996; Mackiewicz et al. 2002). In humans this number is a baffling 98%! In yeast most of the 25% are regulatory elements, promoters. This is where we most likely will find our usual suspects and the focus of this thesis: Transcription factors.

1.1 AIMS

In this thesis, I hope to provide some answers and progress into the following broad questions:

- Can we understand the regulation of genes by studying the transcription factors in a holistic, systems biology way?
- Can we build transcriptional regulatory networks (TRNs) that implicates the role of a transcription factor in different metabolic states?
- Can we utilize this information to understand the underlying function that constitutes transcriptional activation, and by doing so increase our understanding to construct better cell factories?

1.2 PROMOTER ARCHITECTURE

1.2.1 PATTERNS, THEY ARE EVERYWHERE

We humans love to find patterns. As Carl Sagan said: “Humans are good at discerning subtle patterns that are really there, but equally so at imagining them when they are altogether absent”. Since our entire genome is made up of patterns, perhaps it is therefore understandable that we try to find them everywhere. We will now look closer at some reoccurring genomic patterns in *S. cerevisiae*.

The promoter is a DNA sequence located upstream of a gene that regulates the gene expression. The typical architecture of *S. cerevisiae* promoters includes the following core elements: the TATA/-like box, the transcription start site (TSS) and upstream activating/repressing sequences (UAS/URS) (**Figure 2**). The TATA/-like box is a sequence found in many promoters that contains a repeat of the nucleotides T and A. This sequence allows for binding of the TATA-binding protein (TBP) that is part of the preinitiation complex (PIC) involved in gene transcription, which is covered in more detail in section 1.3.1. The TSS defines the start of mRNA transcription, where a gene can have multiple TSSs, and is directly upstream of the start codon: 5’-ATG-3’ (Zhang and Dietrich 2005) and also covered in **Paper XIII** (not included in this thesis). Upstream UAS/URS contains sequences that attracts the transcription factors (motifs). Most promoters have a nucleosome depleted region (NPR) of 400 bp where UAS/URS is located (Ozonov and van Nimwegen 2013).

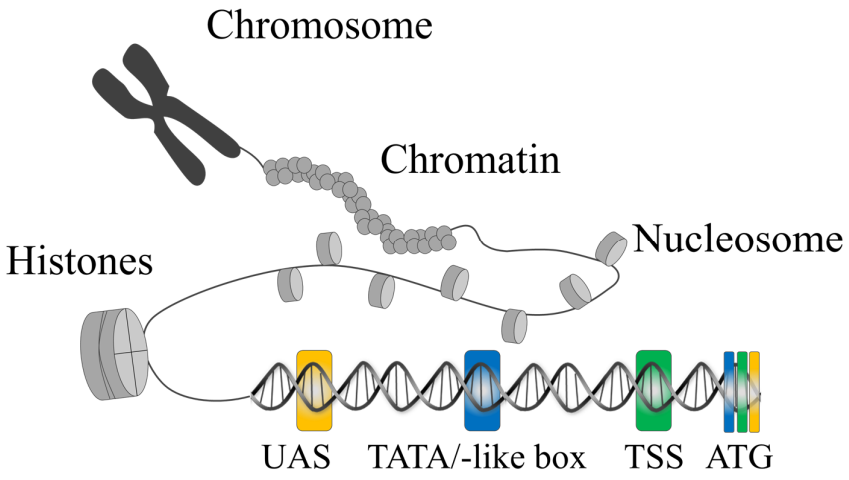


Figure 2 The packaging of DNA into chromosomes. The chromosome is a condensed state of the chromatin which is composed of DNA and nucleosomes. Unwinding the chromosome reveals individual nucleosomes composed of histones and DNA. The promoter then is composed of short sequences that are required for binding of transcription factors (UAS/URS), or the pre-initiation complex (TATA/-like box). This leads to the formation of the transcript starting from the TSS and then reaching the coding sequence starting from the ATG.

1.3 TRANSCRIPTION FACTORS

In *S. cerevisiae*, there are roughly 200-260 transcription factors (TFs) (Hughes and de Boer 2013). The concept of transcriptional control was first coined by Jacob and Monod (Jacob and Monod 1961), and it was later established that this control was due to DNA binding proteins: transcription factors. These transcription factors belong to different families depending on their DNA binding domain (DBD). The major classes of transcription factors in *S. cerevisiae* are displayed in **Figure 3**. The first and most abundant class is the one containing a Zn^{2+} stabilized DBD consisting of ~120 proteins. This class includes the two major subclasses C2H2 and Zn2Cys6 and minor subclasses such as C4. We have studied several Zn^{2+} stabilized DBD transcription factors, including Cat8, Sip4, Ert1, Rds2, Rgt1, Hap1, Stb5, Oaf1, Pip2, Sut1 and Leu3. The C2H2 TF subclass forms an array, or tandem repeats, of zinc-stabilized alpha helixes that can interact with the DNA (Bohm et al. 1997). The Zn2Cys6 TF subclass are homodimers or heterodimers that together form the DBD. This class of zinc fingers is unique to fungi. Due to variations in the overall proteins, the dimerization mechanism can be different, but the principle of having two zinc fingers forming the DBD remains the same (MacPherson et al. 2006). The second class is one containing a zipper DBD. We have studied several transcription factors from this class, including Cbf1, Tye7, Ino2, Ino4, Cst6, Gcn4 and Rtg1. This class is also divided into two subclasses: basic leucine zipper (bZIP) TFs (Fernandes et al. 1997) and basic-helix-loop-helix (bHLH) TFs (Robinson and Lopes 2000). This class of TFs can both form homodimers and heterodimers. In addition, smaller classes of transcription factors include the helix-turn-helix (HTH) and the forkhead (Fkh) TFs.

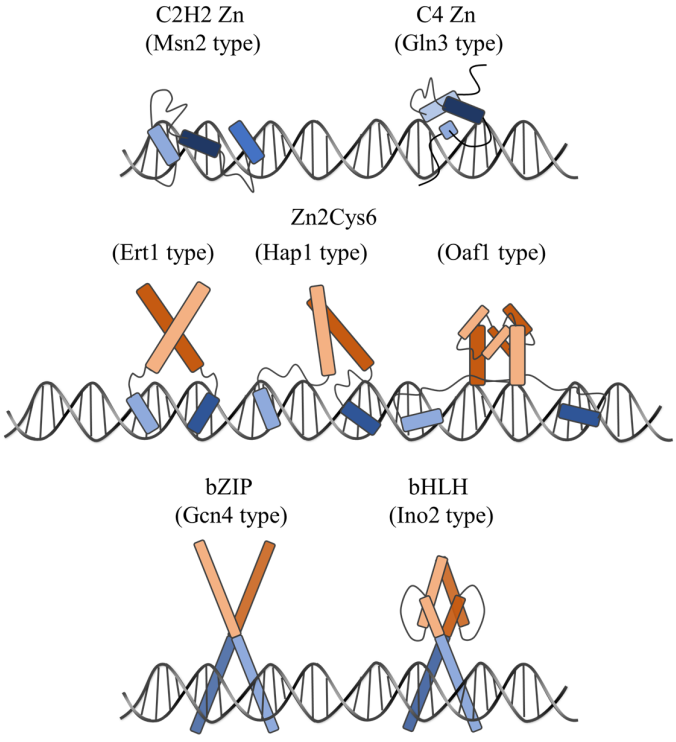


Figure 3 The major classes of transcription factors in *S. cerevisiae*. The zink fingers C2H2, C4 and Zn2Cys6, and the zippers bZIP and bHLH.

Most transcription factors are dimers, where both proteins are required for DNA binding. However, there are also examples of heterodimers where one peptide contains the binding domain while the other contains the activation domain. Such an example is Gcr1 and Gcr2, where Gcr1 binds to DNA and Gcr2 contains the activating domain (Uemura and Jigami 1992). The different domains of the transcription factor also constitute its role in the regulatory machinery. While activating or repressing domains act by recruiting coactivator or corepressor complexes to the naked DNA, chromatin remodeler domains act upon recruiting other transcription factors to the DNA structure while assembled into chromatin (Workman and Kingston 1998). To understand how this is achieved, we need to return to the chromatin structure.

1.3.1 WHERE'S THAT ON SWITCH?

In its most common state, the DNA is covered with nucleosomes that cover most of the naked DNA. Nucleosomes consist of four histone pairs around which DNA is tightly folded and are used for packing the DNA into chromatin and then to chromosomes. Chromosomes are extremely compact and allow DNA to take up less space in the nucleus. Each nucleosome occupies a ~147 bp stretch on the DNA, which allows it to also act as repressors of transcription as it physically blocks the TATA/-like box, TSS or UAS from interaction with transcription factors or other proteins involved in transcription initiation (Juan et al. 1993). Transcription factors can however overcome this physical blockage through different mechanisms. **Figure 4** explains this initial setup that is required for gene expression to occur.

The SWI/SNF complex, that was first discovered in yeast (Winston and Carlson 1992), is a nucleosome remodeler that can either act on its own or through interactions with transcription factors that guide the remodeling complex to the right location (Neely et al. 2002). These remodeling complexes work by modifying the histone tails that are susceptible for modifications. The most common modifications are acetylation and methylation, but also phosphorylation, ubiquitination and sumoylation occur (Kouzarides 2007). Another example are pioneering transcription factors, which have higher affinity to the DNA than the nucleosome (Zaret and Carroll 2011). And the last group are the cooperative transcription factors that have multiple binding sites adjacent to each other, or multiple transcription factors that have binding sites next to each other. This increase the probability of DNA binding if one or more transcription factors are already bound, thereby outcompeting the nucleosome(s) (Adams and Workman 1995).

When the nucleosome has been removed, other transcription factors can interact with the DNA to attract the proteins necessary for transcription. However, there is still an additional nucleosome blocking the TSS. Other transcription factors attract other chromatin remodelers: the SAGA complex and the TFIID. The SAGA and the TFIID complexes, contain subunits of histone acetyltransferase (HAT). These two complexes remodel the histone tail to remove the downstream of TSS (+1) nucleosome making the TATA and TSS available for binding. The

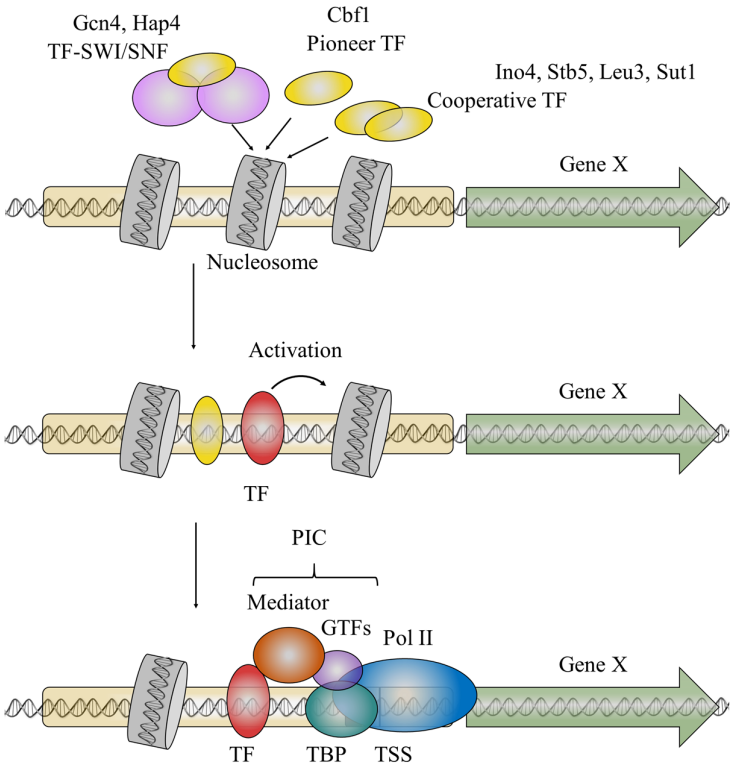


Figure 4 Transcription factor interaction with DNA for gene expression. Removal of nucleosomes can occur through different mechanisms such as the remodelers, pioneer TF or the cooperative TFs. The underlying DNA is revealed and allows for other TFs to bind. The TF attracts the TFIID or SAGA which leads to activation, gene expression, through first removal of the +1 nucleosome and second attracting the PIC.

TATA-binding protein (TBP) is then recruited by the SAGA or TFIID to the TATA/like-box (Huisinga and Pugh 2004), which attracts and assembles with the general transcription factor complexes (GTFs) TFIIA and TFIIB into a stable complex. This recruits the RNA Polymerase II and TFIIF, followed by binding of TFIIE and TFIIH. Together all these parts form the preinitiation complex (PIC) (Rhee and Pugh 2012) that initiates the transcription of said gene.

1.3.2 Regulation of the regulators

To complicate gene regulation further, transcription factors are also regulated themselves. This regulation occurs primarily through two processes: change in concentration and activation (Calkhoven and Ab 1996). The simplest regulation of a transcription factor is through other transcription factors that bind to the promotor of said transcription factor gene, thus changing the concentration of the transcription factor (**Figure 5A** 1). This can also occur in an autoregulatory manner, where the transcription factor is involved in the transcriptional activation of its own gene. This can occur in a simple, direct manner through binding on its own promoter (**Figure 5A** 2), or indirectly through binding to the promoter of other transcription factors that then bind to the promoter of said transcription factor (**Figure 5A** 3).

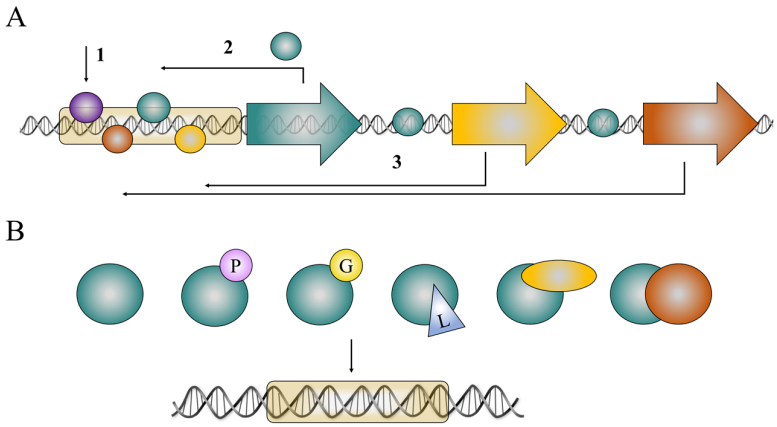


Figure 5 Transcription factor regulation through abundance or activation. A) The abundance of transcription factors is regulated either through 1. Other transcription factors 2. Direct autoregulation or 3. Indirect autoregulation. B) Activation of transcription factors can occur through phosphorylation, glycosylation, ligand binding, cofactor binding or TF-TF dimerization.

Transcription factors can be active in their natural state; however, many transcription factors require activation through external stimuli (**Figure 5B**). This activation, or, for that matter, inactivation, occurs through direct interaction. Phosphorylation and glycosylation are two common posttranslational modifications that can activate/inactivate a transcription factor. These are useful modifications as they can be reversed, thus allowing the transcription factor to switch between active or inactive states. Transcription factors are the largest protein group to be subject to phosphorylation (Ptacek et al. 2005) and around 10 transcription factors are subject to glycosylation (Comer and Hart 1999) where for instance Cat8 is one of them (Cullen et al. 2006). Transcription factors can also interact with ligands, e.g. Oaf1, which contains a ligand binding domain (LBD) for oleate, leading to activation of Oaf1 (Phelps et al. 2006). Rgt1 is a fascinating transcription factor. Rgt1 acts as a repressor in low levels of glucose and as a de-repressor, or activator, in high levels of glucose (**Figure 6**). This regulation of Rgt1 is mediated through two mechanisms: phosphorylation and ligand binding. Rgt1, in low glucose, is bound to co-repressors Ssn6-Tup1, as well as Mth1 and Std1, which inhibits phosphorylation. Ssn6-Tup1 forms a repressive structure together with histones, to assemble nucleosomes, thus repressing transcription through physical blockage (Davie et al. 2002). In high glucose media, Mth1 and Std1 are released from Rgt1, Rgt1 then becomes phosphorylated, which changes its protein structure and results in blocking of its DNA binding domain, thus releasing the repression (Polish et al. 2005). The recruitment of nucleosomes through Rgt1 binding can clearly be seen from the overlay of Rgt1 binding data from T-rEx

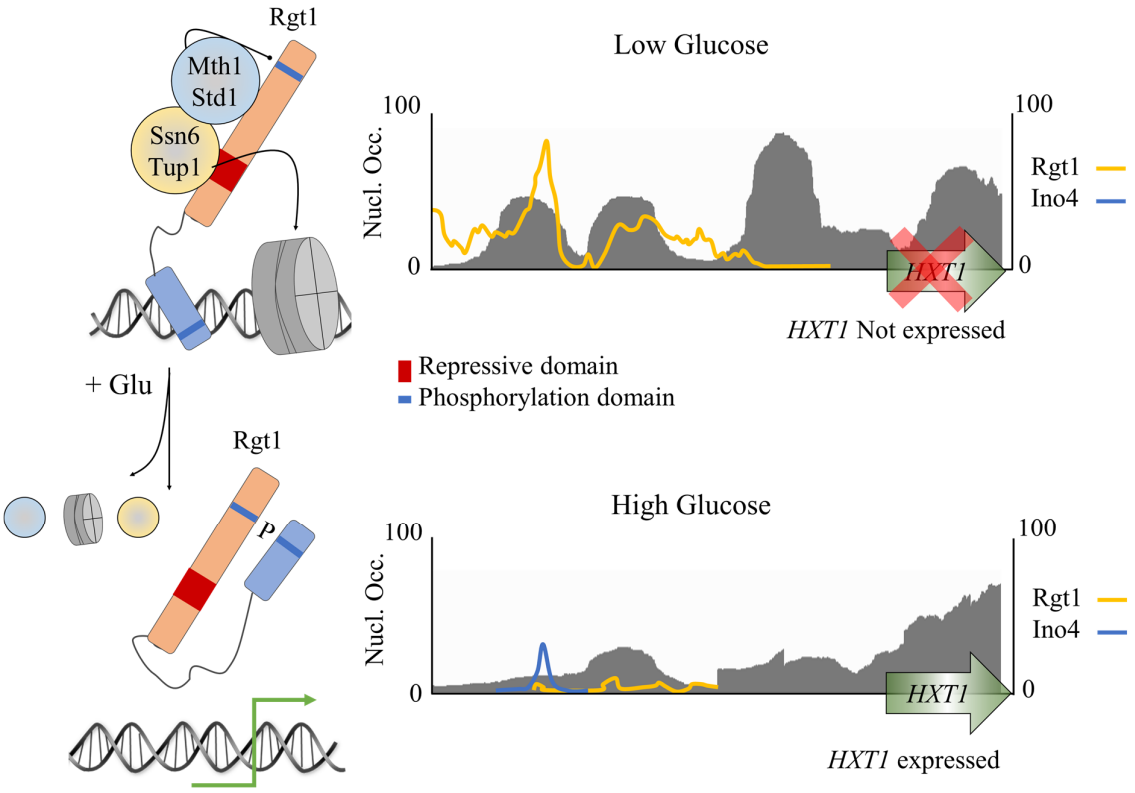


Figure 6 Rgt1 repression and de-repression and its influence on expression of the hexose transporter gene *HXT1*. Left panel: In low glucose media, Rgt1 binds to Mth1/Std1, which blocks phosphorylation of Rgt1. This allows for binding of Ssn6-Tup1, which attracts histones and the assembly of nucleosomes (grey shade) blocking expression of *HXT1*. In high glucose media, Mth1/Std1 is degraded and released from Rgt1. Rgt1 can then become phosphorylated which causes a change in the protein structure of Rgt1, blocking Ssn6-Tup1 from binding, therefore releasing the repression. Right panel: Binding profiles of nucleosomes, Rgt1 and Ino4 on the *HXT1* promoter. Top: Rgt1 is present and attracts the nucleosomes. Bottom: in high glucose Rgt1 is phosphorylated and the nucleosome is removed from the promoter revealing for instance an Ino4 binding site, and *HXT1* can be expressed.

(**Paper VI**) in Glu-lim (low glucose condition) and N-lim (high glucose condition) with nucleosome data from 0.05% Glucose and 2% glucose media (Dang et al. 2014) (**Figure 6**). Furthermore, transcription factors can also bind to other co-factors such as SWI/SNF mentioned earlier, and lastly the transcription factors can interact with other transcription factors. This occurs at a very large extent, where transcription factors can form both homodimers and heterodimers.

2 THE PROMISCUOUS TRANSCRIPTION FACTOR

The transcription factor moves stochastically in the cell, “searching” in three dimensions for DNA to bind to. When DNA is found, the transcription factor executes a linear “sliding” search along the DNA strand to find a motif (Hu et al. 2008). Motifs, I find them very fascinating, are a stretch of DNA containing a sequence of nucleotides that the transcription factor binds to. As mentioned before, transcription factors belong to different families depending on the DBD. Each family has a similar sequence motif that they bind to, but with some variations that allows varying degrees of precision in the binding. The consensus motif of a transcription factor is variable, where some positions in the motif allow several nucleotides, whereas other positions have a fixed nucleotide. **Figure 7** A) illustrates the motif of the bHLH transcription factor Cbf1 and the DNA binding sequences map. While some positions are fixed, others are variable. This promiscuity of the transcription factor allows extraordinary flexibility and ability to adapt to a changing environment, as each transcription factor binds with varying degree of affinity to many motifs, and each motif can in turn be controlled by many transcription factors. The sequence map shows how each binding (each row) has a core set of nucleotides that taken together (each column) form the consensus motif of the transcription factor. The transcription factor can also change its binding preferences depending on numerous factors, such as TF-TF interactions, TF-cofactor interactions, DNA shape (such as major or minor groove), genomic context such as GC rich regions surrounding the motif and the fact that some transcription factors have multiple binding motifs altogether (Inukai et al. 2017).

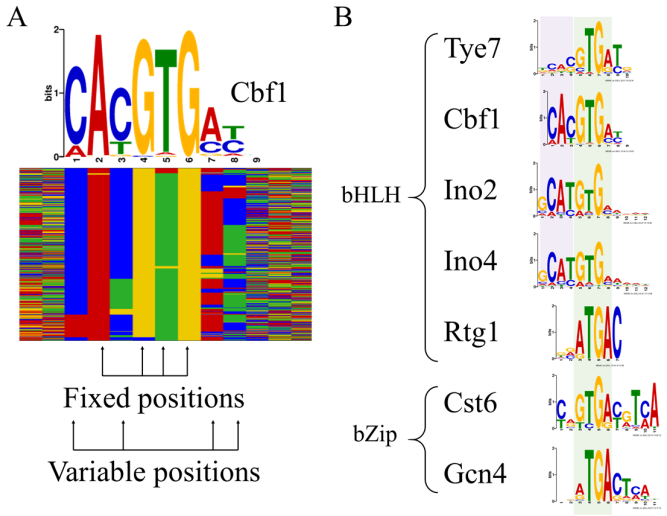


Figure 7 Transcription factor binding motifs. A) A motif of a TF has fixed positions where the nucleotides do not change while other positions are variable and can be exchanged for other nucleotides, usually with a preference of two nucleotides. B) Transcription factors from the same DBD family e.g. leucin zippers exhibit similar binding motifs (green shade) while sub-families have almost identical binding (purple shade).

2.1 WHY ARE THEY ALL THERE?

2.1.1 A RECURRING PATTERN

A common motif for the leucin zipper family is the E-box motif CAnnTG. The individual transcription factors have different nucleotides in the nn part, and there are many examples where multiple transcription factors bind on the same position. For instance, the transcription factors Ino2, Ino4, Cbf1 and Tye7 all belong to the leucin zipper family (bHLH), with the motifs CATGTGA (Ino2 and Ino4) and CACGTGA (Cbf1 and Tye7), where the blue nucleotide indicates the major difference in their motifs. An example of binding for these transcription factors is the *ACS1* promotor which contains an E-box motif 329 bp upstream of the TSS, TCACGTGTGACT, with the E-box motif marked in red. All four transcription factors bind at the same position, despite a mismatch in comparison to the Ino2/Ino4 consensus motif. Interestingly, also Gcn4, Rtg1 and Rtg3, which also belong to the leucin zipper family, bind to the *ACS1* promotor. This is likely due to the motif (G)TGAC, marked in blue, that follows the E-box motif. Worth mentioning is that 5 nucleotides downstream of the E-box is the motif of Sip4/Cat8 which are also bound at the same location as the leucin zippers. Another example is the *ADH3* promoter. At 326 bp upstream of the TSS there is an E-box motif of TCACGTGT. The 8-mer (including the T's at the 5´- and 3´-end) is identical to that of the *ACS1* promoter and also here all mentioned transcription factors bind, including Gcn4 and Rtg1. A third example is the *ADP1* promoter. Here, there are two E-box motifs, one at 202 bp upstream of the TSS, CCACGTGC, and one at 410 bp from TSS, CCACATGC: There is only one nucleotide different in between these two motifs. Interestingly, the motif at 202 bp shows a strong binding of Cbf1, a weak binding of Ino4 and no binding of the other two transcription factors, while the motif at 410 bp has a strong binding of Tye7, moderate binding of Ino2 and Ino4 and very weak binding of Cbf1. This illustrates how the surrounding nucleotides and

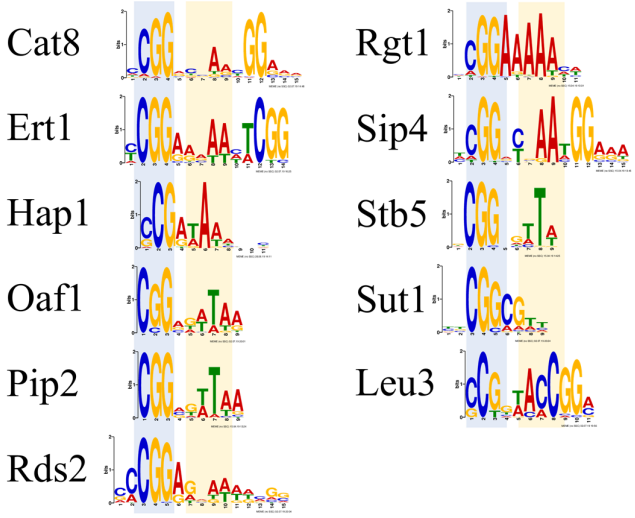


Figure 8 The zinc-finger motifs. The zinc-fingers bind to CSG (blue shade) with a region of A's or T's (yellow shade) in between the two binding sites.

possibly the DNA shape are important for the binding. **Figure 7** B) shows that the motifs of the leucin zipper family have a core set of nucleotides, GTGA, marked in green but that the sub-families (bHLH and bZip) differ in their preferences for the surrounding nucleotides.
We, and others have identified numerous additional examples of such overlaps (Brindle et al. 1990; Chen and Lopes 2007), not only for the leucin zipper TFs but also for the zinc fingers. The two zinc-fingers bind to a CCG/CGG motif on each side of a spacer in our case the spacer contains A's or T's (**Figure 8**). One of the CCG/CGG motifs cannot be identified for all transcription factors in a simple consensus motif as the length can vary between the two fingers. The A-T rich region gives the DNA an electronegative charge in the minor groove, allowing a positively charged linker of the Zn2Cys6 protein to interact (Rohs et al. 2010), see **Figure 3** Leu3 and Oaf1 types for a visual representation of this interaction. Interestingly, all transcription factors we have studied share a common motif of CSGnnWW (S=C/G, W=A/T), although the total length of the motif varies.

2.1.2 NO ONE CAN ESCAPE THE LAW

Why does the motif of a specific transcription factor vary at different locations and how can so many different transcription factors bind to the same location? This boils down to thermodynamics, as small variations in the motif of the transcription factor will change the affinity to each potential target. Briefly, a transcription factor has to be precise in its binding to ensure specificity, but still, the binding affinity cannot be too strong, as it may interact with the DNA permanently. Transcription factors have a transient binding behavior, were these DNA-interactions occur for milliseconds to seconds (Swift and Coruzzi 2017). The disassociation and dynamics of transcription factors are thus very fast, and precision is the price for this fast dynamic. Concerning the sliding mechanism along the DNA, transcription factors from the same family with a similar DBD will have a certain probability of binding to any site that has a similar motif and that they encounter during this sliding process. These low affinity bindings are not only stochastic events, but may also be important for gene regulation (Crocker et al.

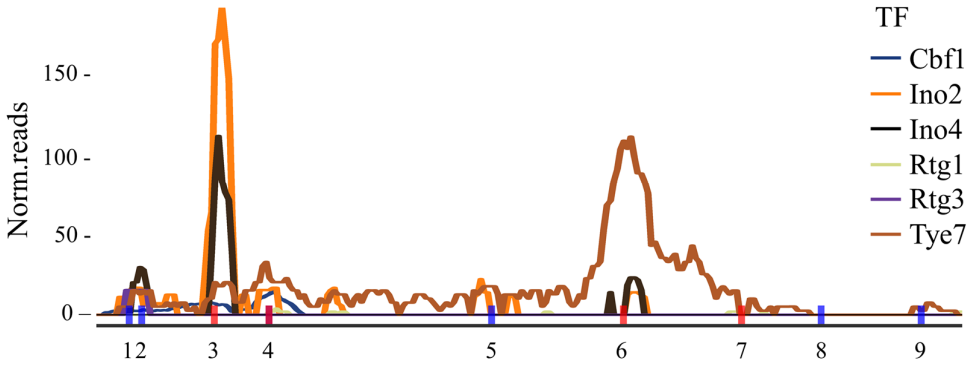


Figure 9 The low and high affinity TF binding on the *ENO1* promoter. Six TFs, all belonging to the bHLH family are bound at 8 of 9 CWCnTG motif sites (blue forward, red reverse). Three motifs (nr 1,2 and 4) are covered by five TFs

2016). The *ENO1* promoter is a prime example of where the bHLH sub-family is showing this behavior (Chen and Lopes 2007). Using the motif CWCnTG (W=A/T) we can find 9 sites within the promoter, 8 of these sites have at least one of the six transcription factors (Ino2, Ino4, Cbf1, Tye7, Rtg1 and Rtg3) bound. At three positions, five of the six transcription factors are bound. It has also been shown that even though many transcription factors are said to work in pairs as homodimers, many of them, especially in the bHLH sub-family, can also interact with each other as heterodimers. Ino4 is for instance recorded to work as a heterodimer not only with Ino2 but also Rtg1, Rtg3, Pho4 and Tye7 (Robinson et al. 2000). These different TF-TF dimerizations are probably what causes some of the differences in binding motifs as the formed heterodimer then may have a higher affinity for a third motif compared to what the two individual homodimers would have (Rodriguez-Martinez et al. 2017).

In summary, there is rarely one transcription factor that controls one gene in eukaryotes. Transcription is dynamic and responsive to the environment, and the system is highly complex with many transcription factors working together. To illustrate and store our understanding of these relationships, we explain the interactions of transcription factors and their targets through transcriptional regulatory networks (TRNs).

3 METABOLISM

In our group, one aim is to improve cell factories for biofuels or other high value chemicals. At the center of all cellular metabolic networks, and therefore of value to this aim, is a set of twelve chemicals. These are called precursor metabolites from which all cellular building blocks and chemical products can be derived (Nielsen 2003). Three categories exist that all metabolic reactions can be divided into. Catabolic reactions comprise pathways that convert feedstock (e.g. carbon source) into precursor metabolites, reducing power and energy in the form of ATP. Anabolic reactions comprise pathways that consume reducing power and energy to produce cellular components (e.g. lipids, nucleic acids, cell wall) or desired chemical products. Central metabolic reactions are those that enable the cell to interconvert between the twelve precursor metabolites, and thereby permitting production of all cellular components from a single catabolic pathway (**Figure 10**). How these reactions and their products can be used in industrial processes was one of the first things I worked on when I started my project, and this is covered in a review (**Paper IX**).

After performing the literature research for this review, my interest in using metabolic engineering and synthetic biology in the lab increased. Fortunately, a new project had just started, looking into the possibility of producing cocoa-butter as a food additive in yeast. Many engineered strains were created utilizing either the endogenous yeast enzymes or heterologous cocoa enzymes with the synthetic biology concept in mind, specifically, to use promoters that

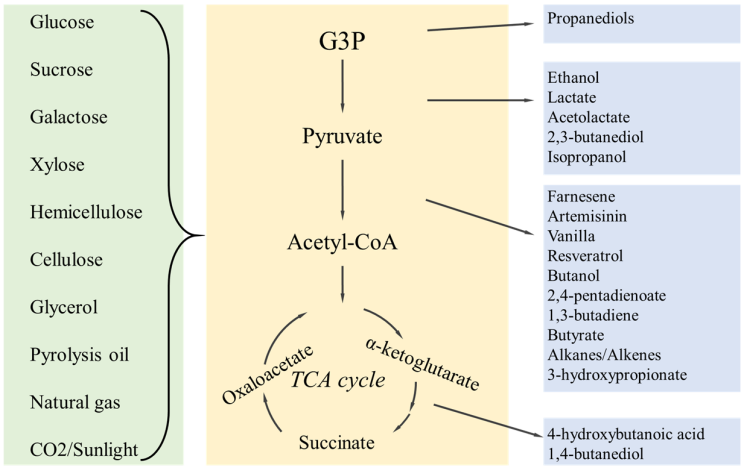


Figure 10 The bowtie structure of metabolism, adapted from Paper IX. Metabolism is shaped like a bowtie, with many pathways funneling into a small number of central metabolites that then branch out into a wide range of anabolic pathways. The 12 precursor metabolites are: glucose-6-phosphate, fructose-6-phosphate, ribose-5-phosphate, erythrose-4-phosphate, glyceraldehyde-3-phosphate (G3P), glycerate-3P, phosphoenolpyruvate, pyruvate, acetyl-CoA, α-ketoglutarate, succinate and oxaloacetate.

acts like switches, turning genes on or off, at certain growth phases. This work is presented in papers **Paper X-XII** but is not included in this thesis.

Central carbon and lipid metabolism are core processes that generate many molecules needed for the production of biofuels, food additives, commodity chemicals, fine chemicals or proteins. To study central carbon and lipid metabolism at a regulatory level helps to understand how to engineer better cell factories and possibly understand human regulation better, as many of the enzymes and pathways are similar.

3.1 CENTRAL CARBON METABOLISM

3.1.1 GLYCOLYSIS

A sugar molecule such as glucose, fructose, mannose or other hexose molecules is transported in to the cell via the hexose transporters (HXTs). The promoter regions of many of the HXT genes have been shown to be bound by the transcriptional regulator Rgt1 (Ozcan and Johnston 1999). (See section 1.3.2 for more info.) The catabolic reactions start by converting the sugar molecule, in this case glucose as it is *S. cerevisiae*'s favorite food, to precursor metabolites. The first part is the glycolysis (**Figure 11**). Here, the sugar molecule is phosphorylated by hexokinases Hxk1 and Hxk2, generating glucose-6-phosphate (G6P). G6P is then converted into fructose-6-phosphate (F6P) by G6P isomerase Pgi1. F6P is converted to fructose-1,6-bisphosphate (F1,6P2) via Pfk1 and Pfk2. F1,6P2 is split into two three-carbon compounds, glyceraldehyde-3-phosphate (GA-3P) and dihydroxyacetone phosphate (DHAP) by aldolase Fba1. DHAP can then be converted to GA-3P via triose phosphate isomerase Tpi1. Glycolysis has so far yielded 2 GA-3P molecules and consumed 2 ATP. The two GA-3P molecules are further converted to 1,3- bisphosphoglycerate (1,3P2G) via glyceraldehyde 3-phosphate dehydrogenase Tdh1, Tdh2 or Tdh3. 1,3P2G is converted into 3- phosphoglycerate (3PG) via 3-phosphoglycerate kinase Pgk1. 3PG is converted to 2-phosphoglycerate (2PG) via phosphoglycerate mutase Gpm1. Phosphopyruvate hydratase, Eno1 or Eno2, converts 2PG into phosphoenolpyruvate (PEP). Finally, PEP is converted to pyruvate via the pyruvate kinases Pyk1 (Cdc19) and Pyk2. Gcr1 and Gcr2 are two key player TFs in the regulation of glycolytic genes (Baker 1986; Uemura and Fraenkel 1990) where Gcr1 contains the DBD and Gcr2 contains the activating domain. Tye7, or Sgc1, is another transcription factor that has shown to be bound to many genes in the glycolysis (Nishi et al. 1995). Abf1 and Rap1 have also been shown to bind to several genes in the glycolytic pathway (Brindle et al. 1990).

The glycolysis has now in total generated two pyruvate molecules, 2 NADH and 2 ATP. The pyruvate molecules can further be converted into a central precursor: Acetyl-CoA.

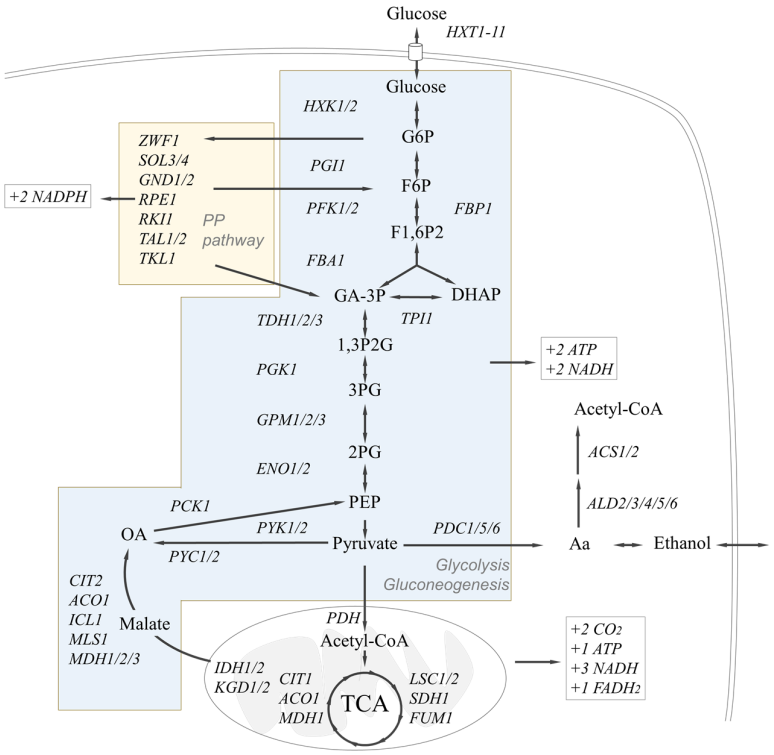


Figure 11 Glycolysis, gluconeogenesis, the pentose phosphate pathway and tricarboxylic acid cycle. The carbon source (glucose) is transferred to the cell where it undergoes many enzymatic reactions to form precursor metabolites.

3.1.2 PENTOSE PHOSPHATE PATHWAY

The pentose phosphate pathway (PPP) generates NADPH and precursors for nucleotide and amino acid synthesis. The first step of the PPP is to convert G6P into 6-phosphogluconolactone (6PGL) by G6P dehydrogenase Zwf1. 6PGL is converted to produce 6-phosphogluconate (6PGC) by 6-phosphogluconolactonase Sol3 or Sol4 and finally 6PGC is oxidized to ribulose-5-phoshate (Ru5P) by 6PGC dehydrogenases Gnd1 and Gnd2. This first part is called the oxidative PPP and generates 2 NADPH and CO_2. The Ru5P formed from the oxidative PPP is converted via Rki1, Rpe1, Tkl1, Tkl2, Tal1 and Nqm1 to form the glycolytic intermediates GA-3P and F6P or ribose-5- phosphate (R5P) that can be used in amino acid metabolism as well as nucleotide and nucleic acid metabolism. Upon oxidative stress Stb5 is the main transcription factor identified to act on the PPP genes (Larochelle et al. 2006)

The 2 NADPH generated in the pentose phosphate pathway and the acetyl-CoA can e.g. be further used in anabolic reaction in the lipid metabolism.

3.1.3 GLUCONEOGENESIS

Gluconeogenesis is basically the reversal of the glycolysis, with some additional steps and enzymes. It is highly important for the utilization of nonfermentable carbon sources to generate energy in the form of ATP and precursor metabolites. Pyruvate cannot directly be converted back to PEP but is so through conversion into the intermediate oxaloacetate (OA) by Pyc1 and Pyc2 and then from OA to PEP via Pck1. Oxaloacetate can also be generated through the tricarboxylic acid cycle (TCA) (see below). A common feature in the gluconeogenesis promoters is the UAS_{CSRE} (CSRE: carbon source responsive element) CGGnnnAAnGG, which is the motif of Cat8-Sip4 (Hedges et al. 1995; Rahner et al. 1999; Roth and Schuller 2001). Gluconeogenesis has a strong connection to β-oxidation and so the UAS_{ORE} (ORE: oleate responsive element) bound by Oaf1-Pip2 can also be found in many of the gluconeogenic promoters. Just as Oaf1-Pip2 and Cat8-Sip4, Hap4 also activates the gluconeogenesis pathway (Zampar et al. 2013). Rds2 and Ert1 are two other transcription factors involved in the gluconeogenesis to utilize nonfermentable carbon sources (i.e. ethanol) (Turcotte et al. 2010).

3.1.4 TRICARBOXYLIC ACID CYCLE

Glycolysis is the primary source of energy (ATP) for yeast cells under fermentative conditions. However, when yeast is grown on alternative carbon sources or when glucose is depleted, the metabolism shifts from fermentative to respiratory and carbon is shunted to the mitochondrial tricarboxylic acid (TCA) cycle thus increasing electron transport and respiration. The TCA cycle occurs in the matrix of the mitochondria, where pyruvate is converted through oxidization to form energy and precursor metabolites. It starts with pyruvate being converted to acetyl-CoA via the pyruvate dehydrogenase complex (PDH), consisting of Pda1, Pdb1, Pdx1, Lat1 and Lpd1. Acetyl-CoA combines with a four-carbon acceptor molecule, oxaloacetate (OA), to form a six-carbon molecule, citrate, by Cit1. Isocitrate is formed from citrate by Aco1. A carbon is released as CO_2, and NADH is generated in the next step generating α-ketoglutarate by Idh1 and Idh2. Kgd1 and Kgd2 catalyze the reaction to form succinyl-CoA, again generating CO_2 and NADH. Succinyl-CoA undergoes a series of additional reactions, first producing an ATP molecule by Lsc1 and Lsc2, then reducing the electron carrier FAD to $FADH_2$ by SDH1. Fumarate is converted to malate through introduction of a water molecule by Fum1 and finally generating another NADH by Mdh1. This set of reactions regenerates the starting molecule, oxaloacetate, and so the cycle can repeat. From pyruvate, two CO_2, three NADH, one $FADH_2$ and one ATP molecule are generated. TFs Rtg1 and Rtg3 have shown to be both involved in the regulation of genes involved in the TCA cycle and in peroxisomal assembly (Chelstowska and Butow 1995).

3.1.5 AMINO ACID METABOLISM

The pathways for the biosynthesis of amino acids (AA) are diverse. However, they have an important common feature as their carbon skeletons come from intermediates of glycolysis, the pentose phosphate pathway, or the tricarboxylic acid cycle. Yeast cells provided with an

appropriate source of carbon and nitrogen can synthesize all amino acids used in protein synthesis. Glutamate and glutamine are key components in AA metabolism as they are used in the transamination reactions required in the synthesis of each AA. There are five families of amino acids. These are the **glutamate family** (glutamate, glutamine, arginine, proline, and lysine), the **aspartate family** (aspartate, asparagine, threonine, and the sulfur-containing amino acids cysteine and methionine generated from the TCA cycle via α-ketoglutarate or OA), the **aromatic family (**phenylalanine, tyrosine, and tryptophan and histidine generated from the PPP), the **serine family** (serine, glycine, cysteine and methionine) and finally the **pyruvate family** (alanine and the branched amino acids valine, leucine, and isoleucine generated from glycolysis) (Ljungdahl and Daignan-Fornier 2012).

The transcriptional activator Gcn4 is a key activator of amino acid metabolism. Gcn4 binds to promoters of genes possessing the consensus UAS_{GCRE} sequence motif GAGTCA (Hinnebusch 1988). Leu3 is another transcription factor involved in amino acid metabolism and as the name suggests, it is mostly involved in leucine metabolism (Zhou et al. 1990).

3.2 LIPID METABOLISM

The lipid group is vast and contains many different molecules. The major groups are fatty acids, sphingolipids, phospholipids, triacylglycerol, sterol esters and sterols (**Figure 12**). Fatty acids are the major component of most of the lipid classes, where the only exception are the sterols.

3.2.1 FATTY ACIDS

Acetyl-CoA is the building block of fatty acid synthesis (**Figure 12**), where it is converted to Malonyl-CoA via the enzyme Acc1. Malonyl-CoA and acetyl-CoA is merged, via the fatty acid synthase complex (Fas1 and Fas2), to form the base of fatty acids, where a new Malonyl-CoA is added in each cycle. The reaction is typically terminated when the acyl chain reaches 16-18 carbons. Elongation to 18 carbons is mediated through Elo1, and further elongation is mediated via Elo2, or Elo3, plus the accessory enzymes Ifa38, Phs1 and Tsc13 in the ER membrane. This reaction uses 2 NADPH. C16 and C18 fatty acids are the desaturated via Ole1 that introduces a double-bond in the Δ9-position and is oxygen requiring (Oh et al. 1997; Page et al. 1994; Stukey et al. 1990; Toke and Martin 1996).

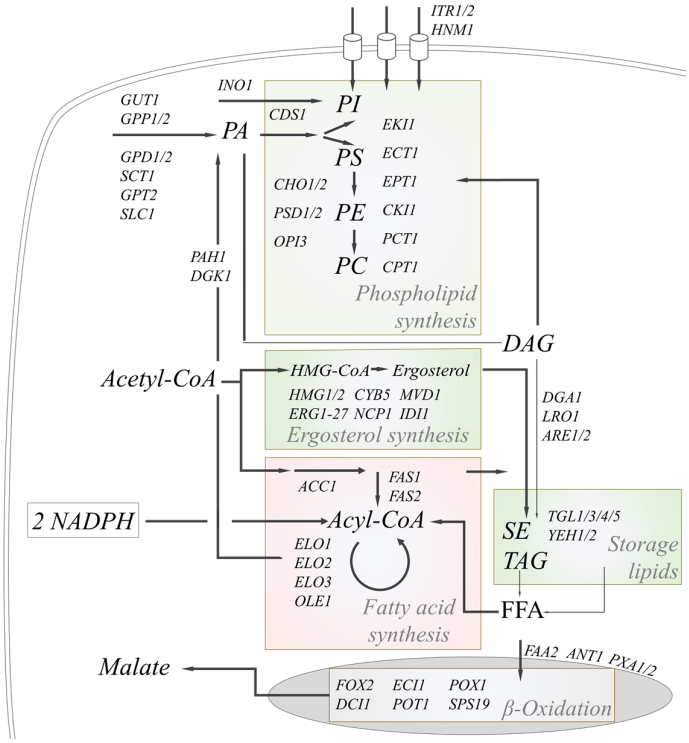


Figure 12 Genes involved in lipid metabolism. Fatty acid synthesis generates the acyl-CoA chain used in phospholipid, sterol ester and triacylglycerol synthesis. Free fatty acids can be used as a carbon source in the b-oxidation. PA: Phosphatidic acid, PI: phosphatidyl inositol, PS phosphatidyl serine, PE: phosphatidyl ethanolamine, PC: phosphatidyl choline, DAG: Diacyl glycerol, SE: Sterol ester, TAG: Triacyl glycerol, FFA: Free fatty acid

3.2.2 PHOSPHOLIPIDS

Phospholipids are the main constituents of the membrane together with sterols, where the phospholipids are formed from the fatty-acyl-CoA chains that are merged with and glycerol-3-phosphate and then inositol, ethanolamine or choline, which are formed through the CDP-DAG and the Kennedy pathway (**Figure 12**). A common feature is a short regulatory sequence named the UAS_{INO} (GCATGTGAA) found in the promotor region of genes involved in the fatty acid and phospholipid synthesis (Chen et al. 2007; Chirala et al. 1994; Lopes and Henry 1991). This sequence is related to the two transcription factors Ino2 and Ino4. The regulation of Ino2 and Ino4 has a third component, Opi1, which binds to Ino2 and represses it. Opi1 is bound to the ER when levels of phosphatidic acid (PA), which is an important intermediate in phospholipid synthesis, are high, allowing Ino2 and Ino4 to activate their gene targets, but when PA levels drop, Opi1 is released from the ER and can interact with and repress Ino2 in the nucleus.

3.2.3 ERGOSTEROL

Synthesis of the sterols uses acetyl-CoA as precursor, which is converted through the many Erg enzymes in the ergosterol (sterol) pathway (**Figure 12**). Ergosterol and DAG can then be

converted into storage lipids such as triacylglycerols (TAG) and sterol esters (SE), which form the reservoir of cellular energy and building blocks for membrane lipids. The TAG is made from fatty acyl-CoA (or an acyl-chain derived from a phospholipid) and DAG, while the sterol esters are made from sterols and fatty acyl-CoA. The ergosterol pathway is oxygen consuming and are thus regulated by the heme and oxygen responsive transcription factor Hap1 (Hickman and Winston 2007). Sut1 is another transcription factor that is regulating the sterol biosynthesis (Bourot and Karst 1995; Ness et al. 2001). Upc2 and Ecm22 are other transcription factors involved in sterol biosynthesis (Vik and Rine 2001).

3.2.4 B-OXIDATION

-oxidation is the process where fatty acids are broken down to generate energy. First, storage lipids such as TAGs and SEs are broken down to free fatty acids (FFA) by enzymes in the triacylglycerol lipase (TGL) family. The FFAs are then imported to the peroxisomes where the β -oxidation occurs. FFAs are metabolized in a multistep reaction cascade from acyl-CoA to

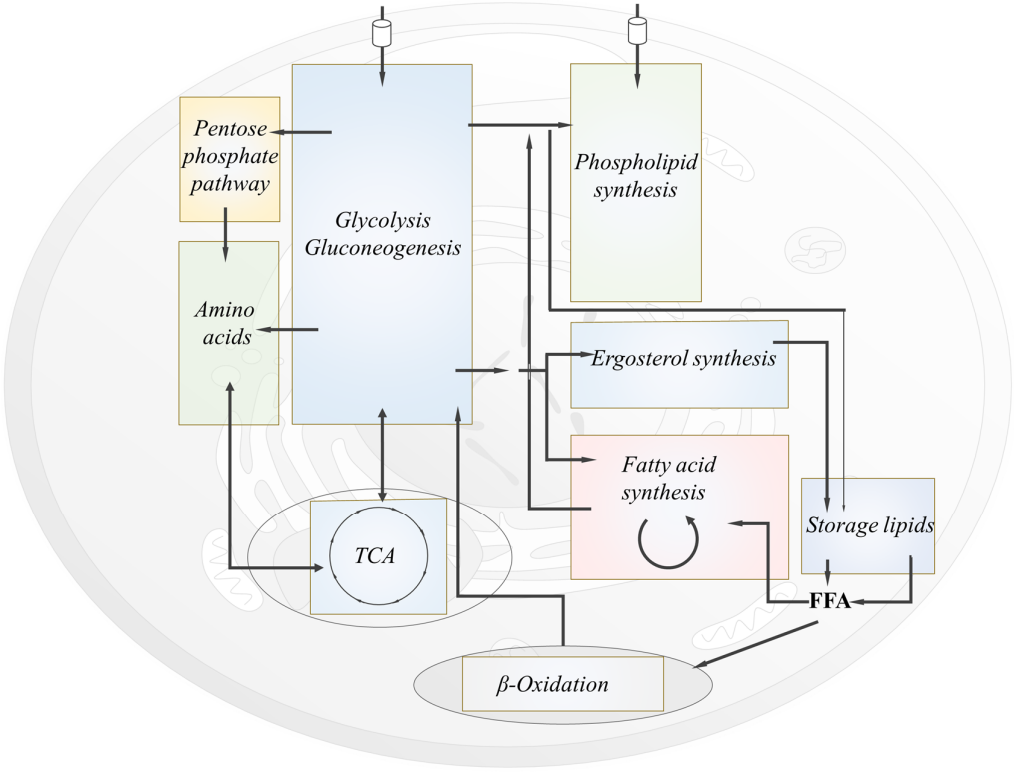


Figure 13 Metabolic pathways included in this thesis. Overview of the major metabolic pathways and their interrelationships.

trans-2enoyl-CoA to 3-ketoacyl-CoA, and finally to acetyl-CoA. This is done by the enzymes Fox1(Pox1), Fox2(Mfe2) and Fox3(Pot1) (**Figure 12**). The transcription factors Oaf1 and Pip2 were shown to be the most prominent regulators of the β-oxidation together with Adr1

(Hiltunen et al. 2003; Karpichev et al. 2008). Acetyl-CoA is transported out of the peroxisomes as malate, which can be used to generate OA, and further be used in gluconeogenesis.

In the overview **Figure 13**, we can see how all the mentioned pathways are connected. Pathways funneling into a small number of central metabolites in the glycolysis and TCA cycle then branch out into a wide range of anabolic pathways. Glycolysis, PPP and TCA generate energy and amino acids. Glycolysis and fatty acid synthesis generate the membrane lipids; sterols and phospholipids. Excessive energy can be stored as storage lipids which are broken down in the event of carbons source limitation through β-oxidation and gluconeogenesis to generate all the central metabolites and thus completing the circle of metabolism.

4 SYSTEMS BIOLOGY

4.1 A HOLISTIC VIEW ON BIOLOGY

To study complex systems such as living organisms in a holistic manner we need a toolbox able to store and connect vast amounts of information of different types. The field of systems biology aims to build and understand the networks that form the whole of a living organism. This is done through the use of mathematical models. This is a cross-functional field where biology, engineering, mathematics and computational modelling are required to advance our understanding of very complex systems such as humans and organs all the way down to protein and molecule levels. There are in principal two viewpoints of systems biology. Bottom-up approaches encompass manual reconstruction of the networks through mathematical methods where reactions and relationships are built based on our current understanding of the system. These models may have varying complexity and detail and are often validated using literature and/or own data used to fit the models. Top-down approaches encompasses metabolic network reconstructions using 'omics' data (e.g., transcriptomics, proteomics) generated through DNA microarrays, RNA-Seq or other modern high-throughput genomic techniques using appropriate statistical and bioinformatics methodologies (Shahzad and Loor 2012). Models developed using top-down approaches are thus data-driven rather than knowledge-driven. These models are unbiased by previous knowledge, and therefore useful to confirm hypothesized or identify previously unknown relationships and handle big data sets and systems where bottom-up approaches simply become too complex. This is the strength of systems biology as the two approaches are complementary. On one hand we can map cellular functions at the genome scale, and on the other hand we can get in detailed timescale resolution of the impact of individual components on overall system properties.

I used a top-down approach to study the transcriptional regulatory networks at a genome-scale level through mainly two high-throughput techniques: transcriptomics and what we sometimes refer to as regulomics. Transcriptome analysis is commonly used to identify genes that are involved in the response to different perturbations (i.e deletions or environmental conditions) and to find mechanisms that are likely to occur in the cell. To characterize biologically meaningful groups of genes with similar changes in expression, i.e. co-regulation, one can use clustering techniques (Eisen et al. 1998). Regulomics, or regulatory genomics, is the study of un-transcribed noncoding regions that contain genomic features, for example that attract transcription factors, and how these features regulate gene expression. Both transcriptomics and regulomics rely on genomics that reveals the full genetic material of the cell. Without the prior knowledge about the genetic material and their function we would not be able to integrate our findings.

4.2 NETWORKS ARE ALL AROUND US

Figure 14 A subsection of the yeast transcriptional regulatory network analyzed in this study. Each dark blue node is a TF and each light blue node is a gene. Many TFs have individual gene targets (the genes at the edge) but many genes are also shared between TFs (genes at the center). The layout of the network is not static but rather highly dynamic and changes as a response to environmental changes.

Atomic, chemical, biological, physical, social, cosmic networks; networks are truly all around us and they all share a common feature: interactions. Interactions occur at all scales, from cosmic scale to sub-atomic. Metabolism in yeast is a complicated network of chemical reactions catalyzed by enzymes. This network can be analyzed through computational models called genome scale metabolic models (GEMs), which can be used to calculate experimentally verifiable phenotypic predictions (Duarte et al. 2004). One step deeper into the network is the transcriptional regulatory networks (TRNs). Transcriptional regulatory networks are maps of the network of regulator-gene interactions that describe potential pathways the yeast cells can use to regulate global gene expression, much like how maps of metabolic networks describe the potential pathways that may be used by a cell to accomplish metabolic processes (Lee et al. 2002). Pioneering work in this field was done by the Young lab, where nearly all transcription factors were mapped in rich media and some in other media using ChIP-chip (Harbison et al. 2004; Lee et al. 2002), and the transcription factor resources developed since: YeTFaSCo (de Boer and Hughes 2011), Yeastract (Teixeira et al. 2017) and SGD (Cherry et al. 1998). Thanks to this, the underlying mechanisms started to be revealed. However, it also became apparent that a more complete picture of the yeast TRNs can be generated by studying the transcription factors in multiple conditions. Figure 14 shows the network of the transcription factor-gene interactions identified and used in our studies. Clearly, the network exhibits so many interconnections that we require computational modelling to analyze such systems. In fact, computational models are an essential component of TRN research (He and Tan 2016).

5 EXPERIMENTAL SETUP

In this chapter I will briefly describe the methodology I have used for transcription factor analysis in the work described in this thesis. These are cultivations in a chemostat, RNA-seq and ChIP-exo and bioinformatics methods.

During batch fermentation on glucose, *S. cerevisiae* typically undergoes a predictable series of growth stages (**Figure 15** A). Initially, the cells are in a lag phase during which they are adapting to the new environment, e.g. rewiring the metabolism to current conditions, such as glucose and oxygen levels in the medium. Once the cells have adapted, the exponential respire-fermentative phase begins. In this phase the cells grow at maximum growth rate, consuming glucose and oxygen and producing ethanol in the process. When glucose is depleted, the cells must adapt again to their new environment and rewire the metabolism to be able to consume ethanol. This is called the diauxic shift and can be seen as a small peak in oxygen levels and as a boulder in optical density (OD), representing growth. When the cells have adapted to the ethanol, they consume large amounts of oxygen to be able to ferment ethanol. This phase is therefore called the exponential respiration phase. Once the last carbon source, ethanol, is depleted, the cells stop growing and oxygen is no longer consumed. This is the stationary phase. As demonstrated, the cell undergoes many different transformations with varying growth rates during batch fermentation. This is not ideal for studying transcription factors as these are integral parts of the machinery that rewires the cells. Thus, we need a more robust system to study the transcription factors where the cells are in a controlled steady-state during the whole experiment. For this reason, we turn our focus to the chemostat.

The chemostat is a bioreactor that uses pumps to control the growth rate of the yeast cells (Novick and Szilard 1950). After all carbon sources are consumed, and the cells have reached the stationary phase, the pumps are started feeding controlled amounts of the selected carbon source to allow the cells to continue to grow at a fixed rate. In a chemostat, there are two important parameters: the growth limiting factor and the media outlet. Without these, the biomass would increase, resulting in a fed batch instead of a chemostat. The limiting factor is commonly the carbon or nitrogen source and is quickly consumed by the cells. Media must be removed at the same rate as media is fed in to ensure a constant volume. From this, we get the important equation $\mu = D = \frac{f}{V}$ where the growth rate, μ, is equal to the dilution rate, D, which is the same as the media inflow, f, over the volume, V, of the reactor. As the volume remains constant, the growth rate is directly proportional to the inflow. Thanks to this fine control of the growth rate through adjusting the rate of inflow, we can control the environment to ensure that it remains constant throughout the cultivation (**Figure 15** B).

We have mainly used four different metabolic conditions in our chemostat experiments: nitrogen-, glucose-, ethanol- and glucose (anaerobic)-limitations (N-lim, Glu-lim, Eth-lim and Ox,Glu-lim). A nitrogen-limited condition keeps the production of amino acids and therefore proteins at a limited level. The medium is rich in glucose and the cells mostly ferment but some degree of respiration does occur. This state allows the cells to store excessive amounts of energy as lipids in lipid bodies. In glucose-limited condition, respiration is fully active as an

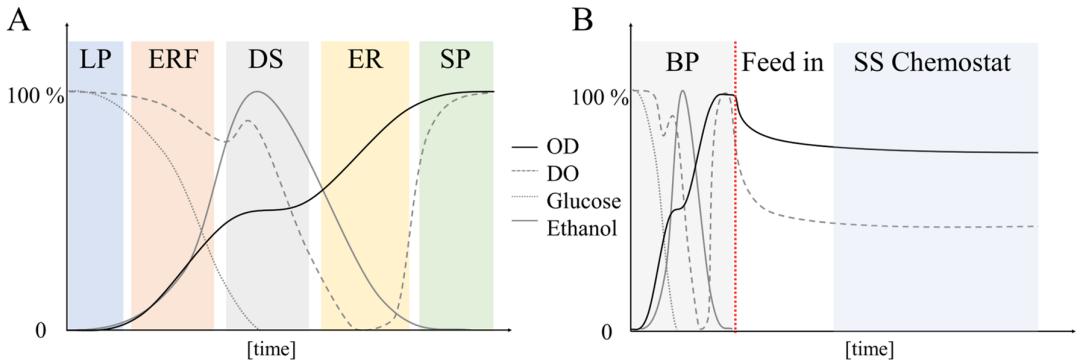


Figure 15 Batch and chemostat cultivation. During a batch cultivation the yeast cells undergo many transformations to adapt to their new environment. In a chemostat cultivation, the growth rate and environment are to remained constant allowing for robust measurements of a highly dynamic system that is the transcriptional regulatory network. LP: Lag Phase, ERF: Exponential Respiro-Fermentation, DS: Diauxic Shift, ER: Exponential Respiration, SP: Stationary Phase, BP: Batch Phase, SS: Steady State, OD: Optical density, DO: dissolved oxygen, glucose and ethanol.

---

additional energy source to glycolysis. In ethanol-limited conditions, gluconeogenesis and respiration takes place. This is for the same reason as for the glucose-limited condition, but here the cells need to use gluconeogenesis as ethanol is a non-fermentable carbon source. Glucose anaerobic limited is a true fermentative state as there is no oxygen supplied to the cells. The conditions have been studied more extensively in (Jewett et al. 2013). These conditions were selected to ensure a wide range of metabolic states of the cells to capture much of the span a transcription factor can have in its regulation.

Transcriptomics, or RNA-seq, is used to measure the transcripts of all genes, the mRNA expression level, in a single cell or in a population of cells. This is an indirect measurement of the activity of the transcription factors. Transcriptomics only captures a snapshot of the dynamic regulation. Therefore, the chemostat is of high importance to ensure a stable and controlled environment and reproducibility of our results.

Chromatin immunoprecipitation using exonuclease treatment, **ChIP-exo.** This method lies at the center of my studies and is the method for studying the interaction between a transcription factor and DNA. It is built upon the method of chromatin immunoprecipitation, which covalently locks any bound protein to the DNA (Solomon and Varshavsky 1985), thus enabling a genome-scale view of DNA-protein interactions. The bound DNA is then enriched and sequenced. Enrichment is enabled by adding a tag (TAP or Myc) at the C- or N- terminus of the protein of interest. An antibody that binds to the tag allows to selectively enrich for proteins containing the tag. The protein is removed from the DNA, followed by DNA sequencing. To ensure that the protein of interest is accurately tagged and not perturbed, we use quantitative

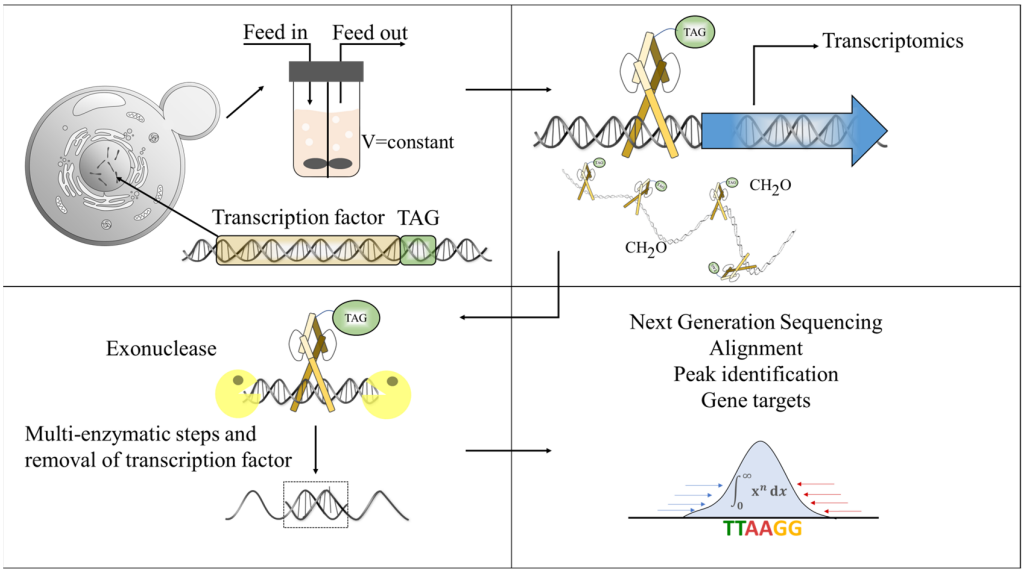


Figure 16 Experimental setup. The yeast cells express a transcription factor of interest with a tag and are cultivated in a chemostat. When steady-state is achieved the cells are harvested and the mRNA levels are measured (transcriptomics). The genomic DNA is crosslinked (CH_2O) with the transcription factor. The crosslinked DNA fragments are extracted and treated with exonuclease and other enzymes leaving only the protected DNA intact. The DNA is then sequenced, aligned and then we can identify the binding (peaks) location of the transcription factor.

real-time PCR (qPCR). We use a cutoff between a reference gene and a gene reported to be a target of the transcription factor to validate that the tag can be used to accurately enrich for DNA bound to the transcription factor.

Predecessors to ChIP-exo are ChIP-chip and ChIP-seq, where ChIP-chip was first version and used microarrays to detect binding events. The next generation was ChIP-seq, which utilizes deep sequencing techniques instead of microarrays. ChIP-seq is the dominant tool for studying gene regulation and epigenetic mechanisms. ChIP-exo uses an extra step of lambda exonuclease treatment (Rhee and Pugh 2011). The exonuclease digests the DNA into single stranded DNA. However, at locations where proteins are bound the DNA is protected. This treatment increases the resolution of the binding down to single bp and removes contaminating DNA, thus reducing the background noise. Importantly, the covalent binding (formaldehyde treatment) occurs during 10 min for all ChIP-techniques. This means that any interaction that occurs between the DNA and the protein during these 10 min will be captured. For chemostat experiments, the cells are in steady state, therefore these 10 min are a representation of the general condition. The -exo protocol is also very labor intensive and requires many enzymatic steps which reduces the DNA concentration to low levels. Binding eventsthat occur at low frequency may therefore not be captured with this method. A third problem with all ChIP methods is encountered if the protein of interest does not contain any DBD but instead a protein binding domain. To be able to capture the binding of such protein requires that the binding

between the protein of interest and its associated protein is intact and that the associated protein is still bound to the DNA after all method treatments.
Fortunately, a new version of the ChIP-exo protocol is now available (Rossi et al. 2018b) where the number of enzymatic reactions has been greatly reduced. This method is therefore improved both in terms of low affinity binding proteins and protein-protein interactions.

Bioinformatics

To identify the genomic locations, we use next generation sequencing (NGS) to sequence the extracted DNA. The reads (sequenced DNA fragments) are aligned to a reference genome using analytical tools such as Bowtie (Langmead et al. 2009), and then a peak identification software is used to map all the binding events. These binding events are then assigned to genes which we group to identify underlying mechanisms. One way to group genes is to use Gene Ontology (GO) terms. This vocabulary of gene products is applicable across organisms and is therefore widely used for analysis of omics data (Ashburner et al. 2000). These can also be linked to metabolic pathway reactions together with gene set enrichment analysis (GSA) that focuses on gene sets, groups of genes that share common biological function, chromosomal location, or regulation (Subramanian et al. 2005). Another way of grouping identified genes is based on their expression patterns, where either linear models or clustering methods could be used. The advantage of using such methods is the ability to identify genes that are co-regulated but do not belong to the same GO-term (Wu 2008).

6 DEVELOPMENT OF A FRAMWORK FOR TRN ANALYSIS

I applied for a PhD project focusing on transcription factors and their regulation networks in yeast with little understanding of the magnitude of the task. As the realization came, the shear amount of experiments required made my head explode: 200 TFs*2 duplicates*4 conditions=1600 experiments! I also realized that to use the chemostats at the time available in the lab I would need to duplicate also myself. The current chemostat setup included 8 1L reactors, meaning 1600/8= 200 experiments, where each experiment runs for about 5 days. With 1000 days of experiments I would need very long time to complete my PhD. The first thing that struck me was the need for more reactors, and the second was the volume as the experiments only require a fraction of the volume, 40 ml of culture. Based on these two criteria we set out to build our own system. This resulted in **Paper I: the mini-chemostat.**

The ChIP-exo protocol was still in its infancy and few labs were using it, and there are still very few who are using it today. Through our affiliation with the Center for Biosustainability (CFB) at DTU we have connections with many groups. One group is the Bernhard Palsson lab, where they had recently started using ChIP-exo in *E. coli*. From them we got some tips and tricks on how to set up the method and soon we were on our way on testing it out. This resulted in **Paper II: Cst6**, where we analyzed this largely unknown transcription factor, previously indicated to be involved in gene regulation in response to growth on nonfermentable carbon sources and stress.

Since the ChIP-exo protocol is (or at least was) rather new, there was no proper unified way of treating the data, to address this we started working on a pipeline. This resulted in **Paper III: Bioinformatics pipeline for analyzing ChIP-exo data**, where we take advantage of existing programs together with our own scripts.

6.1 THE MINI-CHEMOSTAT

As mentioned in the introduction of this chapter, we wanted to establish a large-scale, high-throughput work flow for systems biology research of microorganisms. To do this we needed reliable data from robust cultivation systems. Chemostats have become our favorite cultivation system for this purpose, as it ensures reproducibility and high quality by providing a stable, well-controlled environment for the cells. However, many of the available chemostat systems require large amounts of media and are complex to set up and expensive to purchase and maintain. To address these concerns, we developed a mini-chemostat (MC) system with 16 reactors, each at a working volume of 40 ml.

6.1.1 PHYSIOLOGICAL PARAMETERS

This chapter describes in more detail the different parameters that are important to study the physiology of the yeast cell. For our studies, where a stable metabolic state is required in order to ensure reproducible results, the parameters that need to be kept constant are the dilution rate, the pH, the dissolved oxygen level and the temperature (Furukawa et al. 1983; Lahtvee et al. 2016; Larsson et al. 1993; Regenberg et al. 2006; Verduyn et al. 1990). All chemostats allow control of the dilution rate and thus the growth rate. Another important parameter to observe is the pH as yeast cells produce compounds that lower the pH. This can have as much effect on the state of the cell as the carbon source. Dissolved oxygen is another parameter that is of high importance. Without sufficient oxygen the cells are unable to respire, *S. cerevisiae* can use the (fermentable) carbon source in a fermentative state. However, this would also lead to a complete rewiring of the metabolism. The forth parameter is the temperature, as changing the temperature also causes the cell to change its metabolism.

Moreover, chemostats can be used to study the effect of changing the above described parameters as well as other conditions. By testing multiple conditions, the effects of changing a specific parameter can be studied in detail on a system-wide scale. By systematic changes of the parameters, keeping all parameters but one constant, the importance of each parameter can also be evaluated. As an example, it is possible to identify the transcription factors that controls growth rate dependent genes by altering the dilution rate (Fazio et al. 2008). Chemostats can also be used for selecting clones that have adapted to certain conditions using the concept of adaptive laboratory evolution. This gives new insight into how the cell can evolve to regulate the growth under selective pressure. Omics analyses such as transcriptomics and proteomics benefits especially from this tight control as the whole system is at steady state, and thus the gene expression, regulatory network, and the metabolism remain the same throughout the experiment. This is also important for generating the transcriptional regulatory networks. In addition to the above described parameters, CO_2 can be measured, which provides an indirect way of observing cell growth and an indication of the metabolic state. The CO_2 levels can also be used to make calculations for flux balance analysis (Nissen et al. 1997) that are being used in genome scale metabolic models (GEMs) (Duarte et al. 2004).

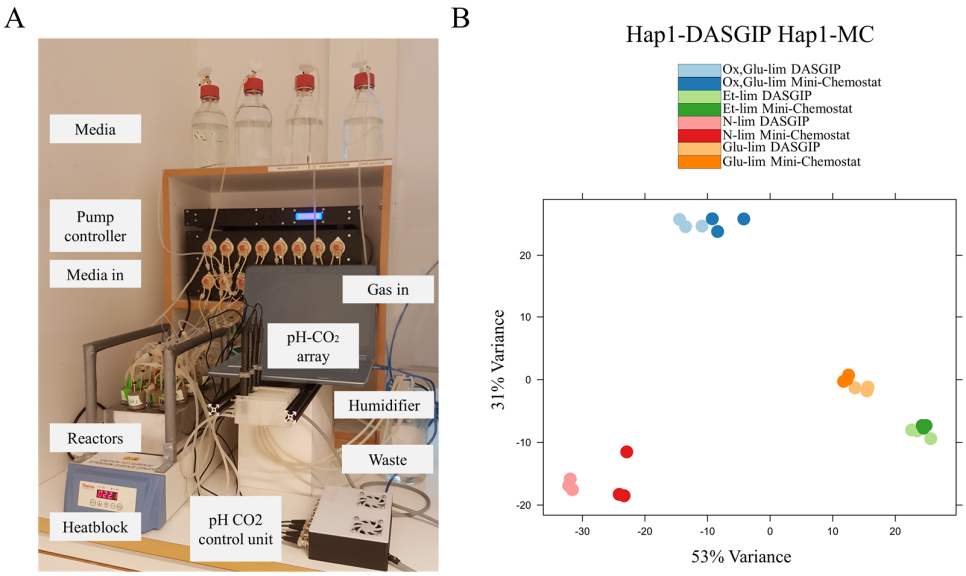


Figure 17 The mini-chemostat setup and its performance. A) The setup including the 50 ml reactors, heatblock, pumps, gas in, humidifier ph-CO_2 sensor array and controller. The DO sensor cannot be seen as it is inside the reactors. B) The system was tested against the standard in our lab, a 1L Dasgip bioreactor. The PCA analysis shows the clustering of the replicates in either system and how the two systems cluster compared to each other.

6.1.2 THE DESIGN

The initial setup was simple: 16 glass vessel of 50 ml each, a single pump head (but with individual tubing) for each reactor, gas and media inlets and gas and media outlet. The temperature was controlled through a heat block while OD and pH were measured in the outflow. However, this setup turned out to be too simple. The selected pumps did not allow sufficient precision at the low flow rates that were required resulting in unexpected differences in flow rates depending on the content of the media. The gas resulted in evaporation from the media resulting in loss of water and decreased volume. We also chose not to include a DO sensor as this seemed troublesome to fit into our small system. However, this proved completely vital when running the chemostat in respiratory conditions such as the glucose and ethanol limited.

So, we redesigned the system. The pumps were swapped for individual pumps of higher quality, a humidifier was included (gas was led through water before entering the vessel) and a DO sensor. We also built a pH-CO_2 sensor array capable of continuous measurements to allow real-time data to be collected while experiments are running (**Figure 17** A).

6.1.3 A SYSTEM COMPARABLE WITH COMMERCIAL SYSTEMS

The system was validated against a high-quality commercial system, evaluating both the stability of the physiological parameters and the ability to replicate expression data using four different conditions. The four conditions were N-lim, Glu-lim, Eth-lim and Ox,Glu-lim (see section 5). Transcriptomics data as well as DO, OD, pH and CO2 measurements were taken from the 1L chemostats and from the mini-chemostats at the four different conditions and compared. Differential expression of genes was assessed using Deseq 2 (Anders and Huber 2010) and visualized with PCA analysis. The triplicates from both the Dasgip and mini-chemostat clustered together indicating that the systems are comparable (**Figure 17** B). Most genes showed similar expression pattern, although we identified some genes that were differently expressed. To analyze the data further we used gene set enrichment analysis (GSA) on GO-terms to assess groups of genes in known pathways simultaneously (Varemo et al. 2013). At a p-value < 0.01 there were no GO-terms enriched for any of the conditions, showing that the systems were comparable. We increased the p-value threshold to 0.05 and found significant GO-terms. Interestingly, there were no GO-terms that overlapped between the conditions which showed that there was no systematic difference between the two systems.

At this stage, potential improvements to the system were identified, such as automatic measurement of DO and OD as manual measurements are quite labor intensive. Also, the gas is not run through any mass flow control making the oxygen uptake rate (OUR) difficult to estimate. To address this, the development of this device was transferred to a company, D2Biotech. The system has since then been redesigned and now includes automatic reading of OD and DO, magnetic stirring and a mass flow controller (data not published but available at d2biotech.com).

6.2 CST6: A STRESS-INDUCED TRANSCRIPTION FACTOR

While I was working on developing the mini-chemostat system, Guodong Liu was setting up the ChIP-exo protocol. We selected Cst6 as a first target to evaluate the protocol and start the journey into yeast transcription factor regulation. Cst6 had previously been mapped in a large-scale ChIP-chip study (Harbison et al. 2004). However, in this study, no consensus motif could be identified nor could the later identified target *NCE103* be found (Cottier et al. 2012). It was known that deletion of Cst6 leads to poor growth on respiratory carbon sources like ethanol (Garcia-Gimeno and Struhl 2000). Another study indicated that Cst6 expression is dependent on the type of carbon source (Osterlund et al. 2015). We used the ChIP-exo protocol to identify binding sites for the transcription factor Cst6 in *S. cerevisiae*. Our aims were both to confirm the previous findings and also identify new targets for Cst6. This experiment was conducted using batch cultivation as the mini-chemostats were still under development. We evaluated Cst6 using both glucose and ethanol as the carbon source.

6.2.1 BINDING TARGETS

Interestingly, the identified binding targets were very different between the two media. In total, we identified 40 binding sites when the cells were grown in media with ethanol as the carbon source. We could identify the binding site motif GTGACGT from the bound region sequences. This motif was slightly different from the previously *in vitro* identified motif TGACGT. In contrast, only 6 binding events were detected when the medium containing glucose was used, of which 4 were also found during growth on ethanol. This indicated that Cst6 hardly binds to its targets in a glucose-rich environment. 16 targets were identified to belong to genes encoding mitochondrial proteins. Together with the previous knock-out studies this indicated that the transcription factor is involved in respiratory functionality. Another interesting finding was that Cst6 binds to the promoter of ten different DNA-binding or transcriptional regulatory proteins, of which many are stress regulators.

6.2.2 NCE103 AND THE BICARBONATE PATHWAY

The next step was to link the transcription factor binding events to the effect this has on gene expression levels. To do this, we generated a *CST6* deletion strain (*cst6*Δ) strain. One of the major targets of Cst6 is *NCE103*, which encodes a carbonic anhydrase converting CO_2 to HCO_3^-. In the conversion of sugar to energy, CO_2 is produced. Nce103 converts this CO_2 to bicarbonate HCO3-. There are several processes that then use this bicarbonate. Pyruvate carboxylase, Pyc1, converts pyruvate to oxaloacetate while using HCO_3^-. Acetyl-CoA carboxylase, Acc1, catalyzes carboxylation of cytosolic acetyl-CoA to form malonyl-CoA using HCO_3^-. When glutamine is converted to glutamate by carbamoyl-phosphate synthase subunit Cpa1, HCO_3^- is used. In the TCA, a carbon is released as CO_2, and NADH is generated when isocitrate is converted to α-ketoglutarate by Idh1 and Idh2. All of the above-mentioned enzymes share a common feature, their promoter is bound by Cst6. The pathways are illustrated in **Figure 18** A). Expression of *NCE103* was severely impaired in the *cst6*Δ strain. Several other identified targets also had an impaired expression, indicating that Cst6 does in fact regulate the expression of its targets mainly by activation.

6.2.3 CST6 IMPACTS CELL GROWTH

The cell growth, evaluated by OD, was reduced for *cst6*Δ compared to the control using ethanol-rich medium (**Figure 18** C) but similar between the two strains when using glucose-rich medium (**Figure 18** B). On ethanol, an extended lag phase was seen, and the final biomass concentration was slightly lower compared to control. *NCE103* is likely the key player here. On glucose, the difference in expression of *NCE103* between the two strains was negligible and the rapid CO_2 production during fermentation may ensure sufficient supply of HCO_3^-. When ethanol is used as the sole carbon source, the decreased *NCE103* and the slower CO_2 production by respiration may not be able to provide enough HCO_3^- for the key biosynthetic reactions required for cell growth.

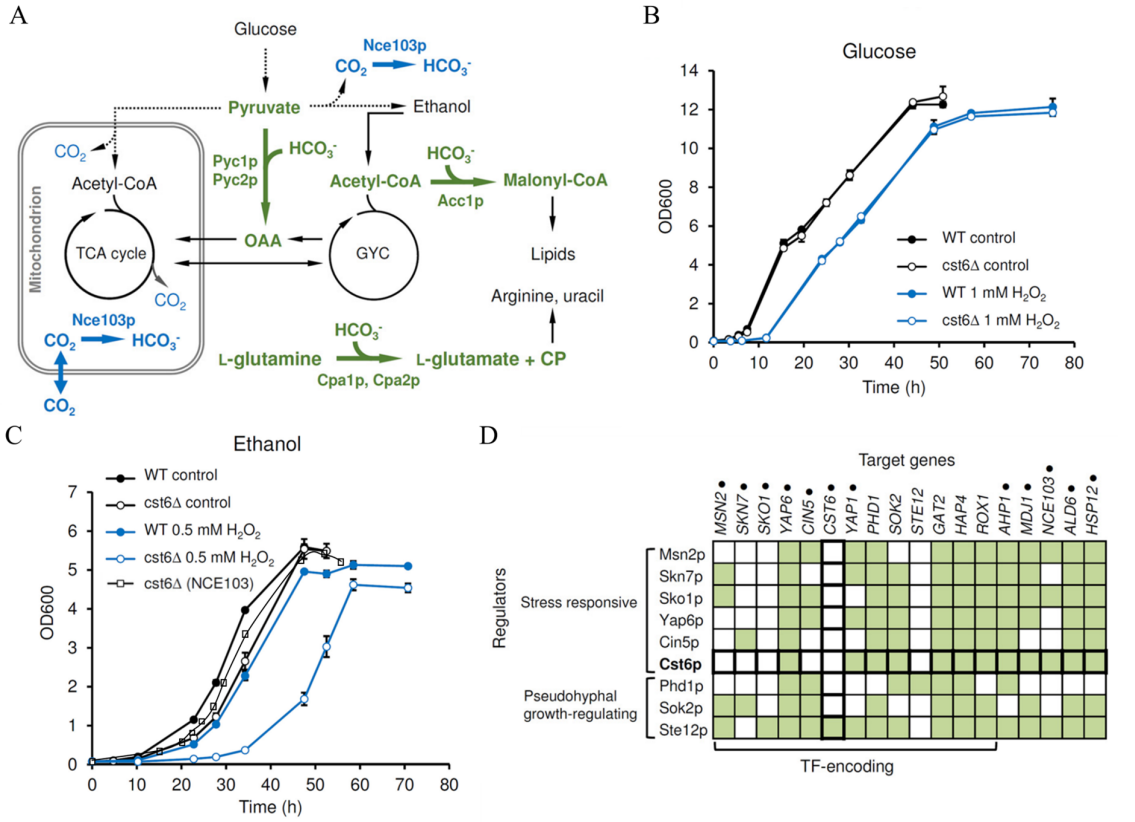


Figure 18 Cst6 role in stress response. A) Major pathways for CO_2 and HCO_3^- metabolism. The reactions producing or consuming HCO^{3-} are shown by thick arrows. Pathways active on glucose are sown by dotted arrows. OAA, oxaloacetate; TCA cycle, tricarboxylic acid cycle; GYC, glyoxylate cycle; CoA, coenzyme A; CP, carbamoyl phosphate. B) Growth on 2% (wt/vol) glucose with or without supplementation of H_2O_2. C) Growth on 1% (vol/vol) ethanol (EtOH) with or without supplementation of 0.5 mM H_2O_2. A strain containing *NCE103* under control of the *TEF1* promoter could partially restore growth on ethanol. D) Heat map showing coregulation of Cst6 targets. Targets with known functions in the stress response are marked by dots. Filled squares indicate TF-target binding relationships.

Since the expression of *NCE103* was decreased to a greater extent than the expression of the other target genes during growth on ethanol, we evaluated whether the low expression of this gene was contributing to the slower growth of the *cst6*Δ strain. We expressed *NCE103* under control of the constitutive *TEF1* promoter in the *cst6*Δ strain. This could partially restore the growth on ethanol at the lag phase. This implies that the carbonic anhydrase activity or the resulting HCO_3^- concentration in the *cst6*Δ strain is the limiting factor for the initial growth of the *cst6*Δ strain on ethanol (**Figure 18** C black lines).

6.2.4 STRESS RESPONSE

In *S. cerevisiae*, Cst6 is an ATF/CREB family transcription factor with a basic leucine zipper (bZIP) domain. All three members in this family, namely Sko1, Aca1, and Cst6 (alias Aca2),

bind to the TGACGTCA sequence *in vitro* (Garcia-Gimeno and Struhl 2000). The function of Sko1 in osmotic and oxidative stress responses and its genome-wide regulatory network have been well documented (Proft et al. 2005; Proft and Struhl 2002). That also Cst6 is active in stress response might therefore not seem too farfetched. Contradictory to our results, previous studies based on phenotypic and gene expression analysis have indicated that Cst6 is not involved in the stress response (Garcia-Gimeno and Struhl 2000). The fact that Cst6 in our study was bound to many stress response genes indicated otherwise. We evaluated the WT and *cst6*Δ response to the stress factor H_2O_2 with glucose or ethanol as carbon source. Growth tests on ethanol showed that the *cst6*Δ strain was more sensitive than the wild type to H_2O_2. On glucose, the *cst6*Δ strain showed sensitivity to H_2O_2 similar to that of the wild type (**Figure 18** B) and C) blue lines).

With the stress test and the binding to stress-related genes, we could establish that Cst6 acts as a stress-responsive transcription factor. Integrating our findings with other published data on stress-related transcription factors showed the complexity of the stress response regulatory network in yeast. Interestingly, we found a hierarchical role of Cst6 in the stress response. Cst6 binds to the promoter of *YAP6*, *YAP1*, *PHD1*, *SOK2*, *GAT2*, *HAP4*, and *ROX1*, which are all transcription factors. Yap1 and Yap6 are involved in stress response while Phd1 and Sok2 are involved in pseudohyphal growth (**Figure 18** D). Overlay of five transcription factors that had high similarities in their set of target genes, all involved in stress response, and Cst6 shows how complex this regulation is. This analysis reveals extensive combinatorial regulation and a transcription factor cascade that is active only in certain conditions.

6.3 PIPELINE FOR ANALYZING CHIP-EXO DATA

When we started analyzing the ChIP-exo data, we soon realized that in order to ensure correct and reproducible interpretation of our data, we needed to develop a pipeline for the data handling. For example, we needed to define what is a peak and what is noise, and exactly how to treat the raw data. In section 5, **Figure 16** I gave an overview of the whole process from tagging the protein to viewing the binding profile. Here, I will describe how this consensus pipeline was constructed.

6.3.1 CHIP-TECHNIQUES

The earliest protocol of ChIP methodology was the ChIP-chip protocol, which uses a microarray with predefined sequences to identify binding events. Typically, this results in a binding spectrum with broader peaks with low resolution to separate binding sites of close

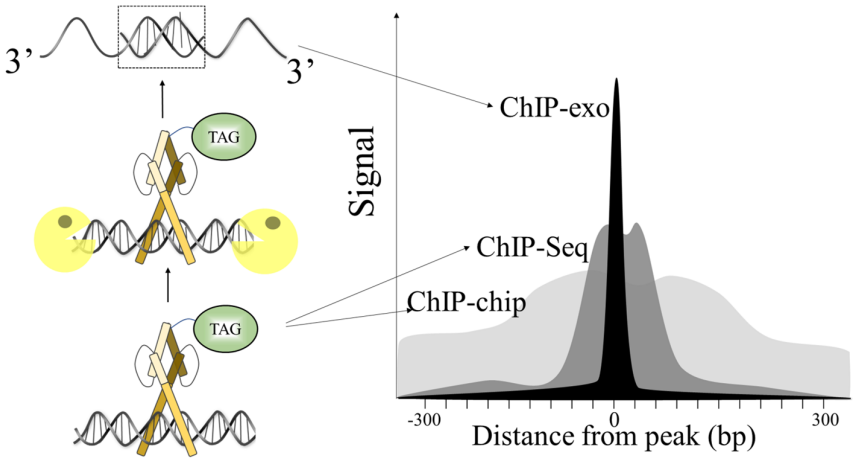


Figure 19 ChIP techniques. ChIP-chip and ChIP-seq are derived from the binding of TFs that are then analyzed using microarrays or NGS. ChIP-exo uses an exonuclease (pacman) that digests contaminating DNA and the flanking regions surrounding the TF binding. This results in superior resolution and binding detection.

proximity. This protocol also results in high noise, which makes it difficult to identify weak binding events (small peaks) and also requires a good control. The next generation was ChIP-seq, where the main difference is the analyzing technology. ChIP-chip does not use sequencing technologies while ChIP-seq does, which allows for higher resolution. However, the noise level is high also for this protocol due to contaminating DNA sequences and controls are therefore needed. In ChIP-exo, the sequencing technology is the same as in ChIP-seq, but the pretreatment of the DNA is different. This pretreatment removes the contaminating DNA, reducing noise and making the control redundant. The resolution is increased as the flanking regions of the DNA surrounding the transcription factor are degraded (**Figure 19**). An advantage of this method is the potential to identify multiple binding sites at close proximity. In ChIP-seq, this could only be inferred if several motifs were identified within the region of interest but does not necessarily mean that the transcription factor binds there. The only option to identify multiple binding in ChIP-seq is by using mutated promoters and assess changes to the signal. In ChIP-exo, these multiple bindings can instead be observed directly thanks to the nucleotide resolution. With ChIP-exo, there is a flanking region at the 3' end of up to 300 bp that the exonuclease does not degrade. This is due to shearing of the DNA. However, it always occurs at the 3' end while the 5' end always point to the transcription factor-DNA crosslinking. Also, the reduced amount of DNA in the sample requires the use of PCR to amplify and increase the DNA concentration.

6.3.2 Data treatment

The sequenced DNA data are in the form of reads with a predefined length based on the sequencing specifications, where we usually use 75 bp. These reads need to be aligned to the

genome for identification, and for this we use Bowtie and map it to the CEN.PK113-7D genome sequence (Salazar et al. 2017). CEN.PK113-7D (or its derivatives) is the laboratory strain of *S. cerevisiae* that we use in our experiments. The PCR amplification can result in duplicates that need to be filtered out. The 5' end of the read may be identical to another 5'end of another read, which identifies them as reads for the same transcription factor-DNA crosslink. However, the shearing causes the 3'end to be at different locations. Therefore, any two reads that have identical 3' and 5' ends are considered duplicates and are removed using the software samtools (Li et al. 2009).

As mentioned, the length of the reads is 75 bp due to the sequencing technique. To increase resolution, the length of the reads must be adjusted, trimmed, to be able to detect binding sites. Transcription factors vary in size and so will the stretch of DNA that transcription factors bind to. We developed a mathematical formula to address this, which we refer to as the TFfootprint formula. This formula assumes that the transcription factor has a spherical shape (adapted from (Erickson 2009)) and also that the transcription factors bind as dimers and thereby overlap with half their size. Trimming is done using the software bamUtils (Wing 2010) applying the TFfootprint formula according to

$$TFweight = sequence\ length\ [nuc] * \frac{1\ AA}{3\ nuc} * 110 \frac{daltons}{AA}$$

$$TFradius = 0.066 \frac{nm}{daltons} * TFweight[daltons]^{1/3}$$

$$TFfootprint\ [bp] = 3 * TFradius * 3.03 \frac{bp}{nm}$$

When the trim length is determined, we identify regions that have overlapping reads from both forward and reverse 5' ends. Overlapping reads indicate that we have identified the two borders of transcription factor-DNA crosslinking. Then the two strands are combined and only regions present in both directions (e.g. complimentary regions in both strands) are stored while the remaining DNA is removed. The stored DNA is then reported as the transcription factor binding profile using the software Bedtools (Quinlan and Hall 2010).

For peak identification we use the software GEM (Guo et al. 2012), which uses an iterative learning process to identify peaks. Peaks with a normalized binding strength below 2 times the noise level (signal to noise ration SNR≤ 2) will be filtered out. Peaks are linked to a specific gene if they are closer than 1000 bp to the TSS.

The whole workflow is visualized in **Figure 20**.

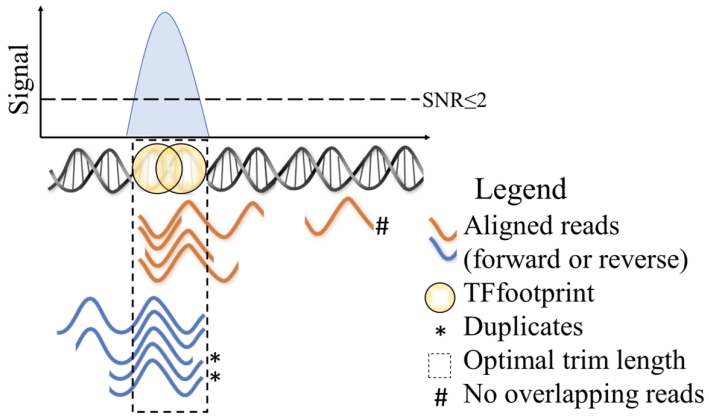


Figure 20 Visualization of the data treatment. Reads are aligned (forward or reverse) to the genome and duplicates are removed. A TFfootprint is calculated that trims the reads to match the size of the TF. Only overlapping reads are kept. Peaks with a SNR≤2 will be filtered out. Peaks are assigned to genes if they are closer than 1000 bp to the TSS.

6.3.3 PIPELINE OUTPUTS

The pipeline generates seven different outputs (**Figure 21**): Strand overlapping reads, Sample correlation, Read profiles, Average Read distribution, List of TF peaks, List of gene targets and Peak centered read distribution.

"Strand overlapping reads" are the resulting data used to analyze and visualize our findings (**Figure 21** A). For this we use browsers like IGV. The graphical output "Sample correlation" shows how well our data correlate e.g. between replicates, which can be used as a data quality control (**Figure 21** B). "Read profiles" mapped to genes are the next output that can be used for visualizing and analysis of the reads, i.e. in our software T-rEx. From the read profile, average "Read distributions" are created including all reads ±1000 bp of the TSS for all genes in the genome (**Figure 21** C).This plot can also be used as a quality control as most binding events occur upstream of the TSS in yeast, and therefore we should see read enrichments in this region if the experiment was successful. "List of TF peaks" is a file containing all peaks identified throughout the genome. This list can be filtered by assigning a "List of gene targets". Then each peak is assigned to a gene and peaks with SNR≤ 2 are filtered out. "Peak centered read distributions" uses the list of peaks and a similar file to the "Strand overlapping reads", but where the strands are still separated. The plots generated show how the reads are distributed around the identified peaks (**Figure 21** D & E).

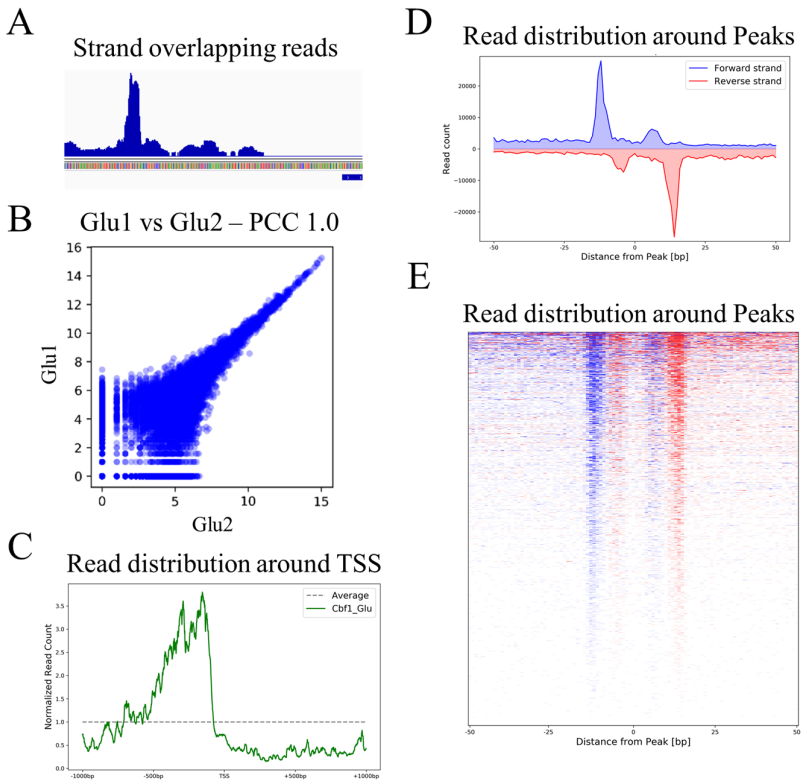


Figure 21 The graphical outputs of the pipeline for Cbf1. A) The strand overlapped reads which can be viewed in IGV browser. B) Sample correlation plot can be used as a quality control that the samples has similar binding. C) Read distribution around TSS shows if and where the TF has more binding than the average noise. D-E) Read distribution around Peaks shows the TF-DNA crosslinking borders.

This concludes this chapter, where I have demonstrated our framework for analyzing transcription factors. In summary, this framework consists of three parts: First, we developed a miniaturized high-throughput fermentation system capable at running several different conditions and with high reproducibility. Second, we demonstrated that the ChIP-exo technique is useful for studying transcription factor-DNA interactions and that the use of different conditions is in many cases vital for the identification of transcription factor targets. Finally, we generated a pipeline that unifies how to treat ChIP-exo data and reports back quality controls as well as data that can be used for further analysis.

7 IMPLICATIONS OF TRNS

This chapter includes the major work of this thesis: transcriptional regulatory networks. It covers **Papers IV-VI**, focusing on the generation and analysis of high throughput data that can be used for reconstructing TRNs. **Paper IV** covers five transcription factors, namely Ino2, Ino4, Oaf1, Pip2 and Hap1. These transcription factors are primarily involved in lipid metabolism, although we show that their TRNs also span beyond this process. **Paper V** comprises a more in-depth investigation of the transcription factor Stb5 regulating the pentose phosphate pathway. In this paper, we reconstruct the TRNs, investigate the biological implications (NADPH levels) and use GEMs to fit the flux distribution of the proposed TRN. Finally, **Paper VI** explores the use of models to predict expression levels of genes based on transcription factor binding and the biological role of transcription factors.

7.1 REGULATORY NETWORK OF LIPID METABOLISM

Lipid metabolism is an important process for the cell, both for generating structural components of the cell and for energy storage (see section 3.2, Figure 11). The five transcription factors Ino2, Ino4, Hap1, Oaf1 and Pip2 have all been implicated in the lipid metabolic processes and were therefore selected for our analysis. Our primary aim was to identify binding sites for each transcription factor with the goal to link them to potential target genes. The experiments were carried out in four different conditions in chemostats, either in a Dasgip bioreactor or in the mini-chemostat system.

7.1.1 HIGH RESOLUTION, NEW TARGETS AND MULTIPLE BINDING

The high resolution of ChIP-exo, where the binding site of a transcription factor can be determined at nucleotide resolution, provided high precision in our data. We could for example without trouble determine multiple bindings on a given promoter (**Figure 22** A), something that would not have been possible with the predecessors ChIP-seq or ChIP-chip. Overall, we found that Ino2 and Ino4 had a high degree of overlap in their targets, but they were not entirely identical. Interestingly, both had a very high number of targets (804 and 652, respectively) and were found to be bound at multiple positions on many promoters (**Figure 22** B). Ino4 had previously been hypothesized to be a global regulator of gene expression (Santiago and Mamoun 2003) and Ino4 had been shown to have an overall regulatory function in DNA

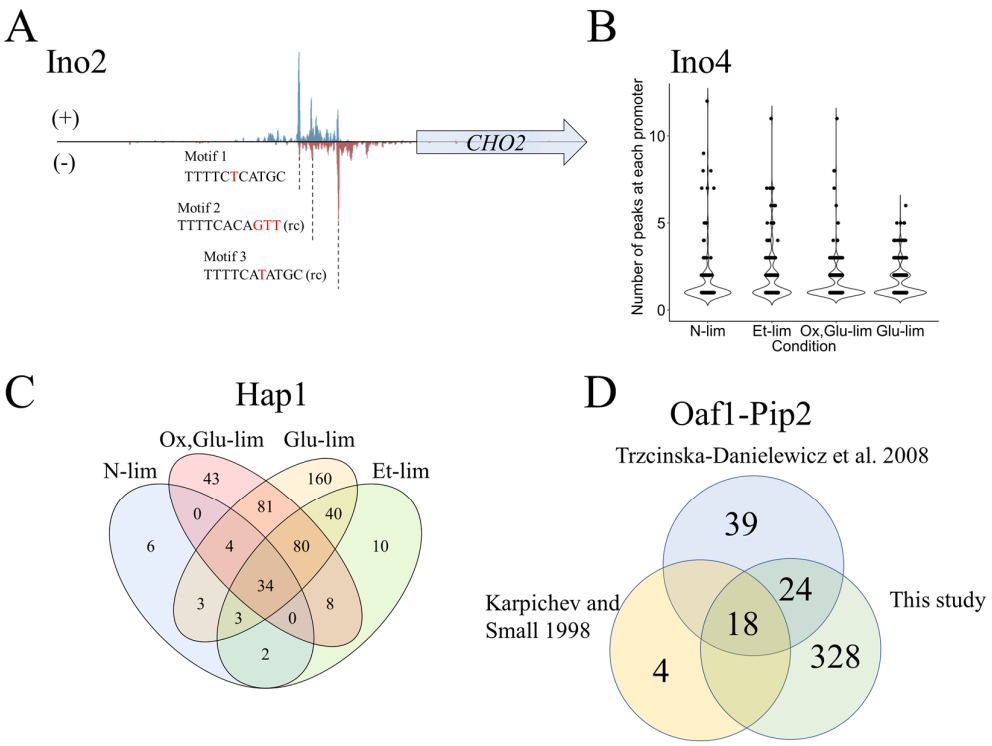


Figure 22 High-resolution, multiple and condition dependent binding. A) Ino2 interaction with the *CHO2* promotor showing three bindings with three different, but highly similar motifs. B) The distribution of bindings of Ino4 in four different conditions indicating that most promotors have one binding site but that are plenty of promoters with multiple binding sites. C) Venn diagram of Hap1 targets showing the four different conditions. 34 genes are shared between the conditions but overall many targets are condition specific. D) Comparison of detected UAS$_{RE}$ from two previous studies and the binding of Oaf1-Pip2 in this study.

damage, targeting 1078 genes when methyl-methanesulfonate was added to the media (Workman et al. 2006). Ino4 belongs to the pioneering transcription factors (see section 1.3.1) that uses multiple bindings to outcompete the nucleosome (Yan et al. 2018), which is well in agreement with the high number of binding sites identified in this study. But are these multiple bindings true, or are they artifacts of the method or even noise? Although some may be noise, much speaks for most peaks reflecting true binding events. As an example, in our data, the *ENO1* promoter shows what could be 5 (possibly even 6) Ino2 peaks or, alternatively, 1 peak and the rest is noise (see **Figure 9**). Looking at the binding sites and including a motif CWCnTG (closely resembling the E-box motif mentioned in section 2.1), all 5 (6) sites contain this sequence indicating that the peaks are true. For Hap1, we found a good overlap with previously reported targets, but we also expanded the list with 320 new potential targets where we see a core set of genes for all conditions but that many targets are condition dependent (**Figure 22** C). For Oaf1 and Pip2, a computational approach to find oleate responsive elements (OREs) identified 85 sites in total (Trzcinska-Danielewicz et al. 2008). It had previously been shown that 22 of these were bound by Oaf1-Pip2 (Karpichev and Small 1998), while we

showed that 42 of these sites were bound in any of the 4 conditions (**Figure 22** D). This not only raises the importance of using different conditions, but it also speaks for that binding sites identified from computational approaches can be correctly predicted given the right conditions. Identifying what causes the changes in chromatin structure allowing the transcription factor to bind could lead to phenotypic predictive models.

7.1.2 CONDITION-DEPENDENT BINDING

All studied transcription factors had a core set of target genes, but they also had plenty of target genes that are condition-dependent thus indicating their response to environmental changes. Hap1 for instance binds to both *ERG2* and *HMG1* in Glu-, Nit-, and Ox,Glu-lim but these binding events were not present in Eth-lim. We also found that the binding is related to the chromatin state (**Figure 23**). For the low affinity hexose transporter gene *HXT3*, nucleosomes are depleted in N-lim (high glucose condition), allowing Ino4 to bind. In Glu-lim condition the nucleosomes are present and Ino4 cannot bind. This is in line with the Rgt1 binding seen for the *HXT1* gene (see section 1.3.2) where Rgt1 binds in Glu-lim but not in N-lim conditions. However, the opposite was observed for the high affinity glucose transporter genes *HXT6* and *HXT7*, where nucleosomes were present in N-lim but not Glu-lim conditions. For *HXT6*, Rgt1 is almost depleted, allowing Ino4 to bind. For *HXT7*, Rgt1 binding is more abundant and lower binding of Ino4 is observed (**Figure 23**). However, the fact that more nucleosomes are present in the N-lim condition suggests that there is a third component that can block the expression of the high affinity transporters. This third component is mediated through the glucose sensor

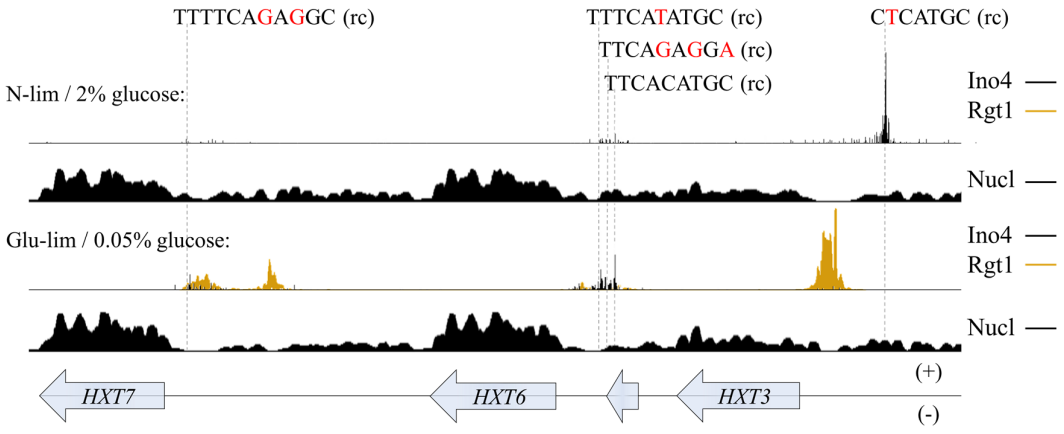


Figure 23 Binding of Ino4 and Rgt1 on the promoters of three hexose transporter genes. For *HXT3* a strong binding of Ino4 can be seen in N-lim while no binding occurs in Glu-lim. Overlaying nucleosome data shows that there are nucleosomes present when Ino4 cannot bind. However, Rgt1 is present, which attracts nucleosomes in Glu-lim condition. For *HXT6* and *HXT7* the opposite behavior can be seen suggesting a third component, Mig1 and Mig2, to the hexose transport machinery.

Snf3, which inhibits the repressive effect of the protein Mth1 resulting in Rgt1 phosphorylation and inactivation (Liang and Gaber 1996). This releases the repression on the promoters of *MIG1* and *MIG2*, which encode two transcriptional repressors that also work by binding to Ssn6-Tup1, which in turn assembles nucleosomes thus repressing *HXT6* (Westholm et al. 2008). This example shows the usefulness of this data set, where previous data and hypotheses can be strengthened and further built on.

7.1.3 REGULATORY NETWORK

To be honest the data are rather beautiful, but one should take care not to get lost in its complexity. I have stared myself blind looking at the different nucleotide compositions in and around the binding sites of each transcription factor. But it is worth it as elegant patterns do emerge. Fortunately, we can apply computational tools to do the hard work of network identification instead of ruining our sight. We used GSA from the Piano R package (Varemo

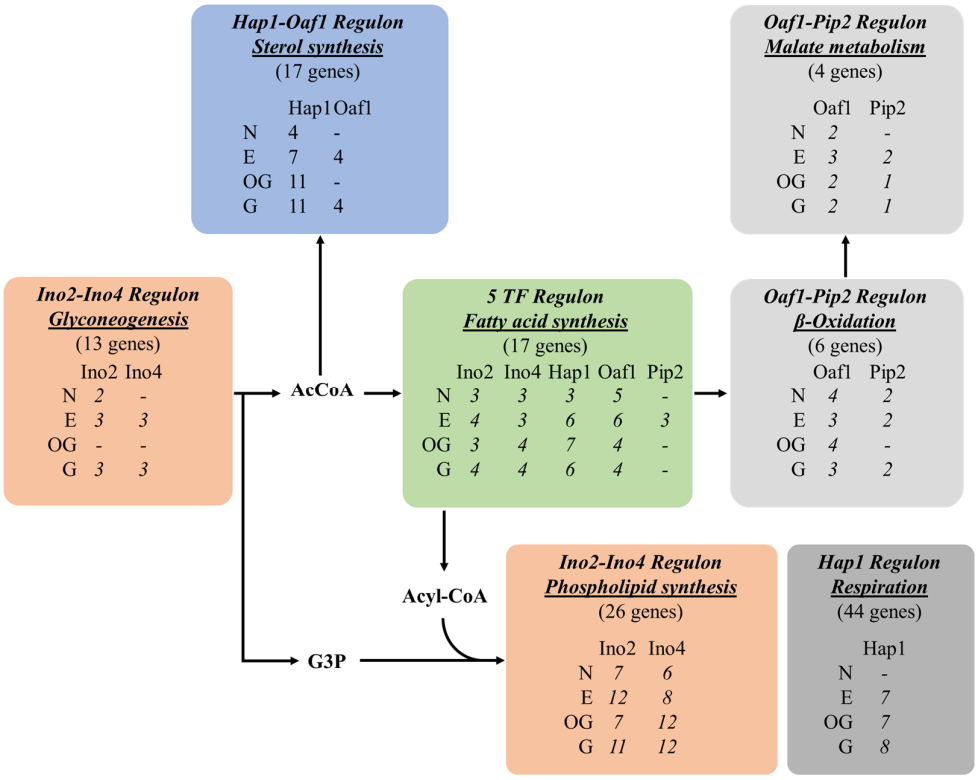
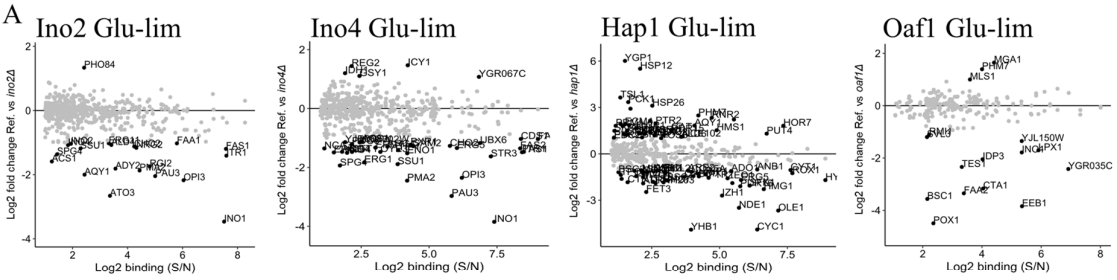


Figure 24 The lipid TRN. The biological processes reported to have statistical significance (*P* value < 0.01) in gene set analysis of target genes with binding ratio log2(S/N) of >1 are shown. For a TF significantly associated with a certain process, the number of bound genes is shown. A minus symbol indicates that the process is not significantly reported for the TF under the corresponding condition. Abbreviations of growth conditions: N, N-lim; E, Et-lim; OG, Ox,Glu-lim; G, Glu-lim.

et al. 2013), applying a cutoff of p value < 0.01 for enriched GO-terms. We found that for Ino2 and Ino4, lipid metabolic processes were enriched as well as amino acid synthesis, cell wall biogenesis, gluconeogenesis and respiration. For Oaf1-Pip2, lipid metabolic processes were enriched as well as β-oxidation and malate metabolism. Oaf1 also had independent enrichments in sterol synthesis and respiration. Hap1 showed enrichments in lipid metabolic processes, sterol synthesis and respiration. Returning to the overall picture in section 3, **Figure 13**, it all makes sense. Ino2 and Ino4 have the majority of their targets in phospholipid synthesis, and fatty acids are required to synthesize phospholipids. The precursors for fatty acids are products of the glycolysis and gluconeogenesis, and so we have traced the network for Ino2 and Ino4. Oaf1-Pip2 is bound to the promoters of genes that are involved in the break-down of fatty acids and sterols. To further generate new precursors in gluconeogenesis, an important key intermediate metabolite is malate, Oaf1-Pip2 targets are enriched for malate metabolism genes. This connection is illustrated in **Figure 24**, where also the different conditions are represented.

7.1.4 GENE DELETIONS AND CHIP-EXO

We used available data on deletions for the five transcription factors to see how well the binding of a transcription factor corresponds to the transcriptional output. For all deletions, we found direct targets, indirect targets and nonresponsive targets. Direct targets have a significant binding (log2(S/N) and a significant down- (or up-) regulation |log2(FC)| >1 in a deletion strain. Indirect targets have a significant FC but no binding, and the nonresponsive targets have binding but no significant FC (**Figure 25** A). We used t-tests including all genes, both one-



B

TF	Condition	Two-sided	One-sided Down	One-sided Up
Ino2	Glu-lim	0	0	0.00081
Ino2	N-lim	0.00423	0.00033	0.24424
Ino4	Glu-lim	0	0.00022	0.02047
Ino4	N-lim	0	0.00002	0.64716
Hap1	Aerobic	0.331	0.00025	0.00225
Hap1	Anaerobic	0.0333	0.05372	0.00937
Oaf1	Glu-lim	0.09453	0.00076	0.58461
Pip2	Glu-lim	0.00079	0.00592	0.98412

Figure 25 Data integration for elucidating the regulatory effect of TFs on their targets. A) Gene expression changes in deletion strains of *INO2*, *INO4*, *HAP1*, *OAF1* and *PIP*2 were correlated with the binding strength (S/N) of the respective TF. Direct significant targets are named. B) P values from the t test of the five TFs under different conditions showed that in a one-sided t test, all TFs had significant (P<0.01) down- or upregulation of genes that were bound compared to non-bound genes.

sided and two-sided, to identify significant up- and/or down-regulations caused by the TFs. The one-sided tests showed most interesting results (**Figure 25** B), as almost all transcription factors showed a significant down-regulation of bound genes in the deletion strains indicating direct targets. Ino2 and Hap1 also showed upregulation of target genes in the deletion strains in a one-sided t-test. For Hap1, this is in line with previous observations that this TF is a repressor in anaerobic conditions (Ox,Glu-lim) (Hickman and Winston 2007). However, this analysis suggests that it also has an activating effect on many other targets. Interestingly, we found that some of the genes that showed upregulation by Ino2 binding had a slightly different motif compared to the consensus. Therefore, we postulate that the upregulation of genes due to deletion of *INO2*is not as much of a repressive effect of Ino2 as it is an enhancing effect of another transcription factor that can take its place. This infers that there is competition between activating transcription factors, and that the transcription factors have different levels of activation (described in section 2.1).

7.2 STB5 A MODULAR NADPH-REGULATOR

For lipid synthesis to function, NADPH is needed. NADPH is mostly generated in the pentose phosphate pathway (PPP, section 3.1.2, **Figure 11**). Stb5 has previously been reported to be a regulator of the PPP and found through ChIP experiments to be bound to many of the genes (i.e. *ZWF1*, *GND1*, *GND2*, *TAL1*,*TKL1*) of this pathway (Larochelle et al. 2006). The PPP is the most prominent route to generate NADPH when cells are grown on glucose. Alternative routes for generating NADPH include Ald6 encoding aldehyde dehydrogenase for conversion of acetaldehyde to acetate using $NADP^+$ as cofactor, and Idp2 encoding a cytosolic NADP-specific isocitrate dehydrogenase. NADPH is also important for reductive biosynthesis such as fatty acid synthesis and oxidative defense mechanisms (Juhnke et al. 1996).

7.2.1 STB5 TARGETS

The experimental procedure to investigate Stb5 was the same as for the lipid metabolic transcription factors and included four different limited chemostat conditions. We found 50 targets that were shared between all conditions, but the two respiratory conditions had an increased number of targets. GO-term analysis showed that PPP, NADPH regeneration and oxidation reduction processes were enriched. Interestingly, for Glu-lim conditions, lipid biosynthesis and ergosterol biosynthesis was also enriched, highlighting the strong connection between PPP, NADPH and lipid metabolism (**Figure 26** A). We conclude from the changes in binding between the different metabolic states of the cell that Stb5 consistently binds a set of NADPH-associated genes (*ZWF1*, *SOL3*, *GND1*, *GND2*, *TKL1* and *ALD6*) (**Figure 26** B), independent of the metabolic condition. Stb5 also shows increased recruitment to distinct groups of additional genes in ethanol and aerobic glucose limited conditions.

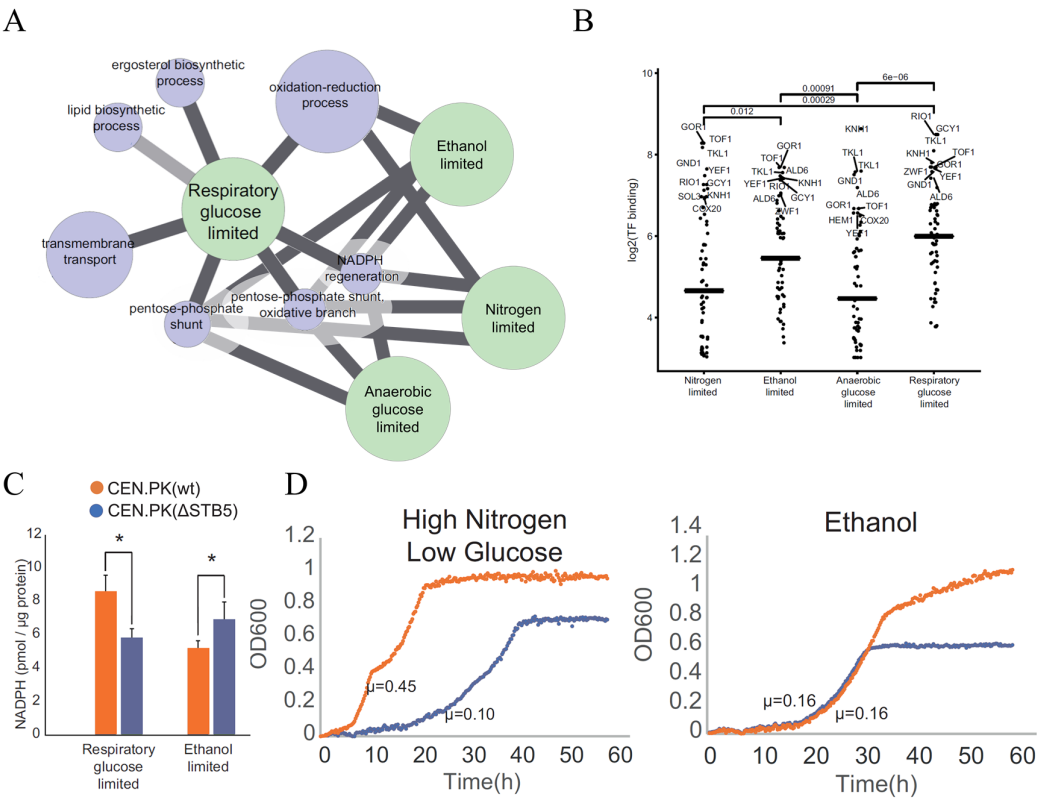


Figure 26 Stb5 targets and the effect of *stb5Δ*. A) The GO-terms connected to the Stb5 targets in the four different conditions. B) Highlighting the strongest bound targets of Stb5 where main targets are in involved in the PPP and/or NADPH generation. C) NADPH levels are reduced in the respiratory glucose limited chemostat condition in *stb5Δ*, whereas in ethanol condition NADPH levels are increased. D) The reduced NADPH levels have a severe effect on the growth rate in Glu-lim media while growth rate was not affected in ethanol media.

7.2.2 NADPH AND GENE EXPRESSION LEVELS IN WT AND *STB5Δ* STRAINS

Measuring NADPH levels in all conditions showed that NADPH levels were decreased in N-lim and Glu-lim conditions but increased in both Eth-lim and Ox,Glu-lim (not significant for Ox,Glu-lim). The reduction of NADPH in the *stb5Δ* strain for N-lim and Glu-lim indicates a functional role of Stb5 in NADPH generation. In Ox,Glu-lim, there is a lower demand of NADPH as oxidative stress is reduced due to oxygen limitation and thereby reduced respiration. Growth on ethanol is an interesting condition as glucose is not the carbon source and gluconeogenesis are the main route for generating precursor metabolites. The higher NADPH levels indicate that an alternative route of NADPH generation is active, which might be inhibited by Stb5 (**Figure 26** C). Research on overexpression of *STB5* supports this theory. In the glucose phase of a batch culture (similar to Glu-lim as tested by transcriptome comparison) when Stb5 can use glucose to generate NADPH, free fatty acids were increased, while after 72 h (similar to Eth-lim as tested by transcriptome comparison) fatty acid levels

were decreased (Bergman et al. 2019). The decrease of fatty acid production during growth on ethanol indicates that the alternative route(s) for generating NADPH from ethanol cannot function properly when *STB5* is overexpressed.

When the deletion strains were grown in batch cultures, it was clear that the low nitrogen and low glucose grown strains where under stress. We measured the growth rate using medium similar to the chemostats. The growth rate in a batch culture at High Nitrogen/Low Glucose was reduced from μ=0.45 to μ=0.1 and in batch culture Low Nitrogen/High glucose from μ=0.22 to μ=0.09. It was fortunate that the growth rates were not considerably lower than μ=0.1, as this was the chosen dilution rate for the chemostats. In batch ethanol media the deletion did not affect growth rate (**Figure 26** D). Transcriptomic analysis of the deletion strain in chemostat showed that only four genes were consistently downregulated in all four conditions: *TAL1*, *GND1*, *KNH1* and a gene of unknown function, namely *YBR085C-A*. In Glu-lim, N-lim, Ox,Glu-lim and Eth-lim there were 230, 500, 480 and 27 genes with significantly changed expression levels, respectively. We found that these genes were enriched for transcription factor genes and GO-terms connected to rRNA biogenesis and processing genes. rRNA biogenesis and processing is a strong GO-term connected to growth rate (Regenberg et al. 2006). Other GO-terms were also connected to stress and chemical resistance. However, only few of the differentially expressed genes were direct targets of Stb5. Connecting this to what we observed before indicates that the cells undergo high stress due to NADPH limitations. The NADPH limitation has an impact on growth rate and when the cells are grown in chemostat close to their μmax, this stress has profound impact on the transcriptome. The overweight of transcription factor genes in the up-regulated targets (not target by Stb5) indicates that this is a stress adaptation by the strain to cope with the deletion of *STB5*. PPP and NADPH regeneration were also enriched GO-terms.

7.2.3 GEM SIMULATIONS

We used genome scale metabolic model (GEM) simulations to obtain information about how fluxes were changed in the deletion strain. We focused on the respiratory glucose limited condition (Glu-lim), as most of the Stb5 targets were identified in this condition. We used the metabolite concentrations from HPLC analysis to constrain the exchange fluxes. The model was optimized for biomass formation and we found that the predicted growth rate was nearly identical to the measured value. Since NADPH is mostly affected in the *stb5Δ* strain, we focused on changes in fluxes of reactions involving NADPH. Interestingly, only five enzymes were predicted to be generating NADPH: Zwf1, Gnd1, Gnd2, Idp1 and Ade3. All these five enzymes have a reduced flux in the model. 63 gene reactions were identified to consume NADPH, which could be grouped to 13 pathways, where the fluxes of these reactions were also reduced (**Figure 27** A). We conclude that the loss of Stb5 function leads to reduced NADPH generation and use in biosynthesis pathways. Which changes in reaction fluxes are then most likely to be caused by transcriptional changes in the *stb5Δ* strain? We used random

A

GEM simulations

NADPH generating reactions		NADPH consuming reactions	
Gene	Flux stb5Δ-WT	Pathway	Flux stb5Δ - WT
ZWF1	-2.72e-2*	Riboflavine	-6.98e-6*
GND1	-2.73e-2*	Alanin and aspartate	-3.07e-3*
GND2	-2.73e-2*	Folate	-3.35e-4*
IDP1	-7.61e-3	Sterol	-2.66e-5*
ADE3	-1.86e-3	Fatty acid	-3.17e-4*
		Threonine and lysine	-2.02e-3
		Arginine and proline	-1.15e-3
		Glutamate	-3.47e-2
		Glycine and serine	-3.07e-3
		Valin, leucine and isoleucine	-2.66e-3
		Tyrosine, tryptophan and phenylalanine	-1.86e-3
		Cysteine	-4.04e-4
		Purine and pyrimidine	-4.89e-4

B

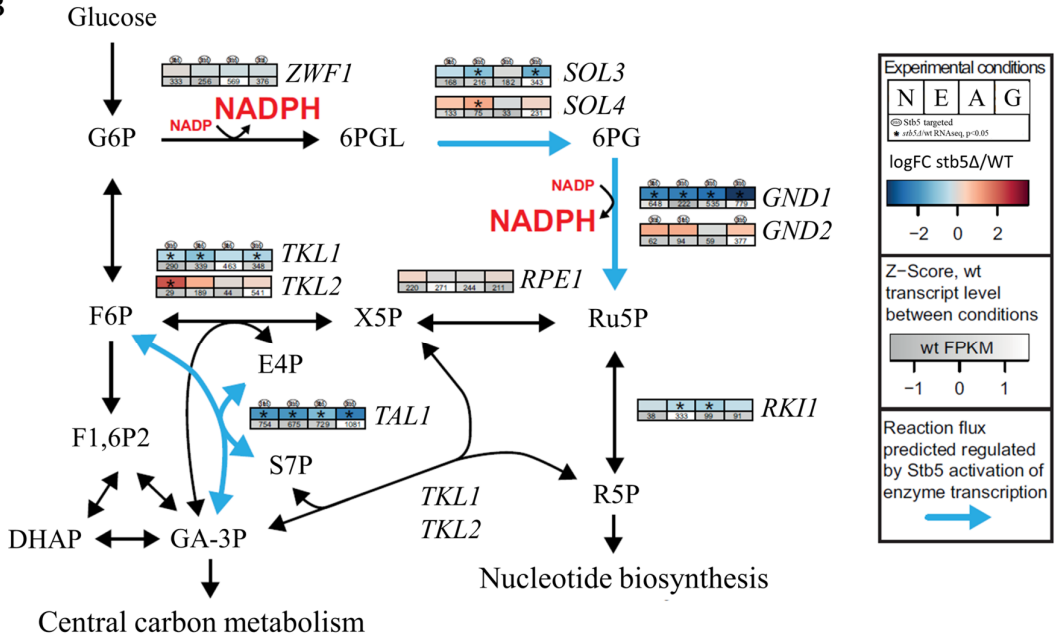


Figure 27 GEM simulations and Stb5 regulation. A) NADPH generating and consuming reactions and pathways where all fluxes are negative indicating reduction of fluxes in all of these reactions due to the *STB5* deletion. B) Stb5 regulation of the main PPP genes and their reactions. N: N-lim, E: Eth-lim, A: Ox,Glu-lim, G: Glu-lim

---

sampling to assess this, where reactions with consistent changes in flux values and gene expression values are most likely to be our targets. We found that of the top nine reactions, four were connected to the PPP. Of these four reactions, three are reactions involving Tal1,

Sol3, Sol4, Gnd1 and Gnd2, for which we previously had identified strong evidence of control by Stb5 (**Figure 27** B).

7.2.4 ADDITIONAL FINDINGS

An interesting observation was that *ZWF1*, one of the major drivers of the PPP, is bound by Stb5, but does not show any signs of being regulated by Stb5 as its gene expression does not change upon *STB5* deletion. However, our previous study on lipid metabolism transcription factors shows that Oaf1-Pip2 are also strongly bound to *ZWF1*, indicating that these may have a stronger regulatory role over *ZWF1* compared to Stb5. In addition, the two alternative routes for producing cytosolic NADPH encoded by Idp2 and Ald6, were slightly down-regulated in the *stb5Δ* strain (*ALD6* only significantly in Glu-lim). Both the *IDP2* and the *ALD6* promoter are bound by Stb5 and by Oaf1. This is in line with that NADPH is needed in the cytosol for the regeneration of thioredoxin/glutathione that are in turn needed for the detoxification of H_2O_2 generated in the peroxisomal β-oxidation (Minard and McAlister-Henn 1999) which is controlled by Oaf1-Pip2 (unpublished results).

7.3 PREDICTIVE MODELS OF TRANSCRIPTIONAL REGULATION

Next, we set out to develop predictive computational models of transcriptional regulation using data from a large-scale experiment including the previous 6 and 15 new transcription factors, requiring roughly 70 days wet lab experimental time. We were interested in how well we can predict the functionality of binding of a transcription factor, and by doing so identify the TRN. It is clear from deletion studies by both us and others that the resulting gene expression change upon TF gene deletion correlates poorly with the targets of the studied transcription factor, with on average 3% of targets overlapping with differentially expressed genes (Gitter et al. 2009). To deal with this we used several different models, including linear regression models and adaptive regression splines. We aimed to keep it as simple as possible to reduce risk of overfitting.

Linear regression models are defined by $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$, where i is the gene, Y is the gene expression (FPKM), β_0 is the intercept, β_1 is the coefficient that is selected to fit the level of transcription factor binding, X_{1i}, to the intercept and ε_i is the predicted error. The direction (positive or negative) value of β has a biological implication. A negative value indicates that the transcription factor has a negative effect on transcription and thus works as a repressor and a positive value indicates that the transcription factor acts as an activator. As it is highly unlikely that only one transcription factor is the sole contributor to the gene expression, we use multiple linear regression models where several predictors are added together. This model is defined by $Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$, where k indicates the index of a transcription factor. For this analysis, we included data from several previous publications (**Paper I, IV and V**) as

the data were generated in the same conditions. Multivariate adaptive regression splines, MARS, is advantageous for prediction performance for example when the transcription factor effect on gene expression is only present at certain transcription factor binding levels. This algorithm allows a type of peak definition where the peak threshold (spline) is introduced to create a linear relationship between transcription factor binding and transcript levels only for a certain range of transcription factor binding strength. Variable selection in MARS will select only the best combination of transcription factors to predict as much as possible while penalizing increasing the complexity of the model. The targets of the selected transcription factors were all enriched in genes involved in central carbon metabolism in the large scale ChIP-chip experiment (Harbison et al. 2004) and the 849 selected target genes are present in the metabolic GEM model v7.6 (Sanchez 2016).
As expected from our previous analyses, the different chemostat conditions, and therefore also the metabolic state of the cells, had profound effects on the binding targets of each of the 21 transcription factors according to our computational analyses. We could both identify new roles and functions and confirm most of the literature-reported GO-terms for all investigated transcription factors, e.g. amino acids for Gcn4, branched chain amino acids for Leu3, glucose transport for Rgt1 and glycolysis for Gcr1.

7.3.1 PREDICTING GENE EXPRESSION WITH MARS

We used MARS to predict the gene expression levels of the 849 genes using 21 transcription factors in all 4 conditions. The MARS model managed to predict 34-43% of the variation of the expression levels between the conditions (**Figure 28** A-B). It also identified transcription factors that benefit from splines over linear models (**Figure 28** A-B). Some of these splines have a threshold before which there is a linear correlation between binding and transcript levels, while others have a saturation effect where more binding does not lead to higher expression. Transcription factors often work in pairs or in bigger complexes. Sometimes these pairs do not give any extra predictive power to the model. MARS can filter these pairs out and only keep the transcription factor that has the highest predictive power (**Figure 28** C-D). To find cases of collinearity in the MARS models where a transcription factor is not included because there is a slightly better predictor selected, we tested if all included transcription factors could be substituted by other transcription factors with significantly correlated binding. Such cases are shown with black borders. However, sometimes these pairs are multiplicative (synergistic) and have a far higher impact on the predictive power if they are defined as a pair in the model (**Figure 28** E-F). All the collinear transcription factor pairs were tested for synergy in this way, and we found that Ino2-Ino4 was multiplicative in all conditions. Ino4 also had more combinations where it could increase the predicative power, in line with what was mentioned in section 2.1. Other known synergistic interactions such as Cat8-Sip4, Gcr1-Gcr2, Rtg1-Rtg3 were also identified.

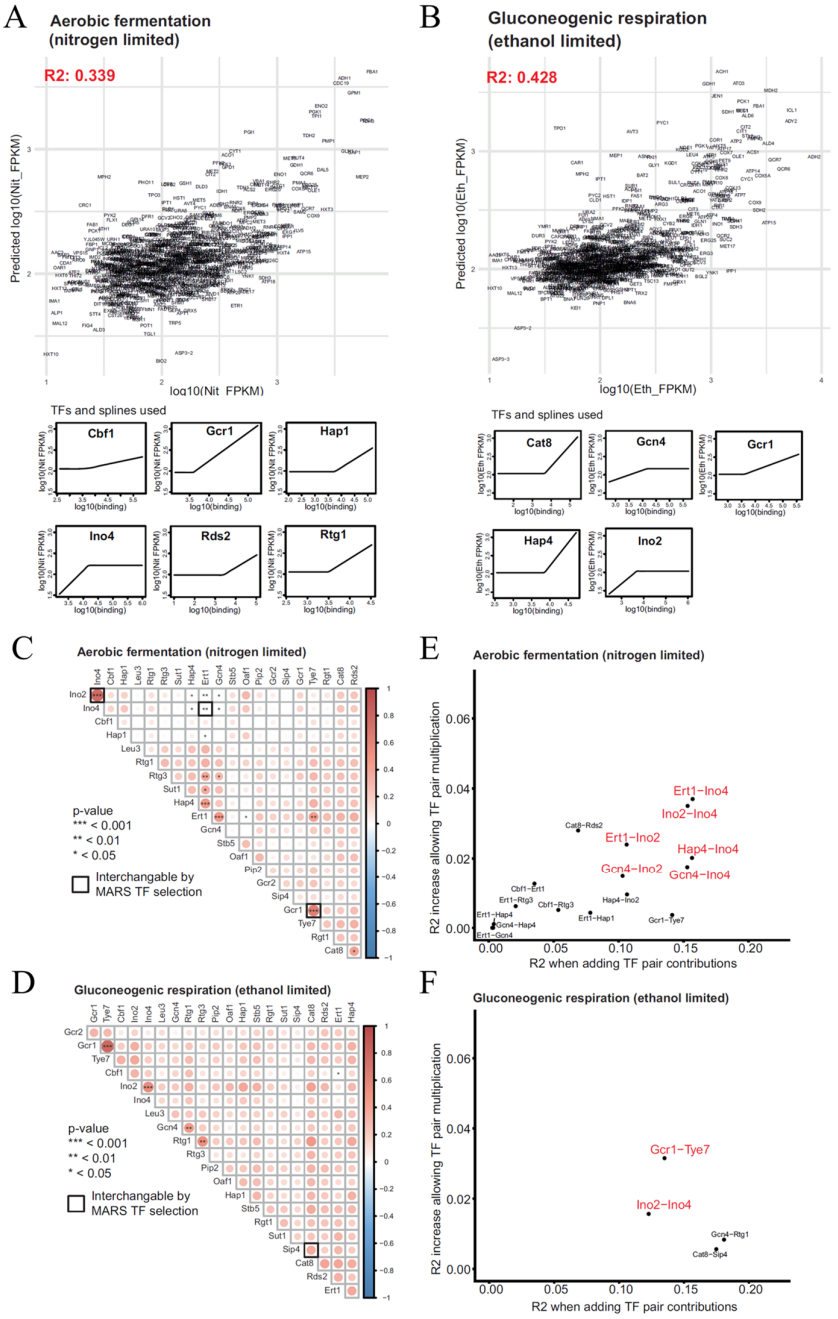


Figure 28 Predicting metabolic gene transcript levels of the different conditions using MARS. (A-B) Predicted vs observed transcript levels, where the boxes below the prediction plots represent the significant TFs and their respective splines. (C-D) Correlation plots between TF binding in promoters of metabolic genes. Significance for TF pairs indicated by asterisk(s). Pairs of significantly collinear TFs that are interchangeable in the MARS TF selection are indicated by a stronger border. (E-F) Linear regressions of collinear TF pairs were tested with and without allowing a multiplication of TF signals of the two TFs. TF pairs indicated in red and with larger fonts have an R_2 of the additive regression >0.1 and increased performance.

7.3.2 IMPROVING PREDICTIVE POWER THROUGH METABOLIC CLUSTERING

From the MARS analysis we could identify transcription factors that give an overall predictive power to the expression levels of all the metabolic genes. However, if a transcription factor only regulates a small set of genes, it could potentially be filtered out in this process. This method instead focuses on genes with similar trends that can be clustered together between conditions. Using k-means clustering, we found that 16 clusters were optimal (defined by Bayesian information criterion) for the metabolic genes. These clusters are represented by several GO-terms, but the strongest GO-term is indicated with a p.adj-value. From this analysis, we selected 4 clusters that were all strongly coupled to central carbon metabolic processes and had a large transcriptional change between the conditions (**Figure 29**). A good indication that this analysis is accurate is that cluster 9, "cellular response to nitrogen starvation", has higher transcript levels in the nitrogen-limited condition. For analyzing the identified clusters, we used the MARS algorithm again, but this time we decided to use it without the splines to increase stability and reduce the risk of overfitting. The predictive power of the model increased significantly when analyzing the clusters compared to analyzing all the genes. To assess the relative influence and predictivity of each transcription factor in a cluster we calculated a "TF

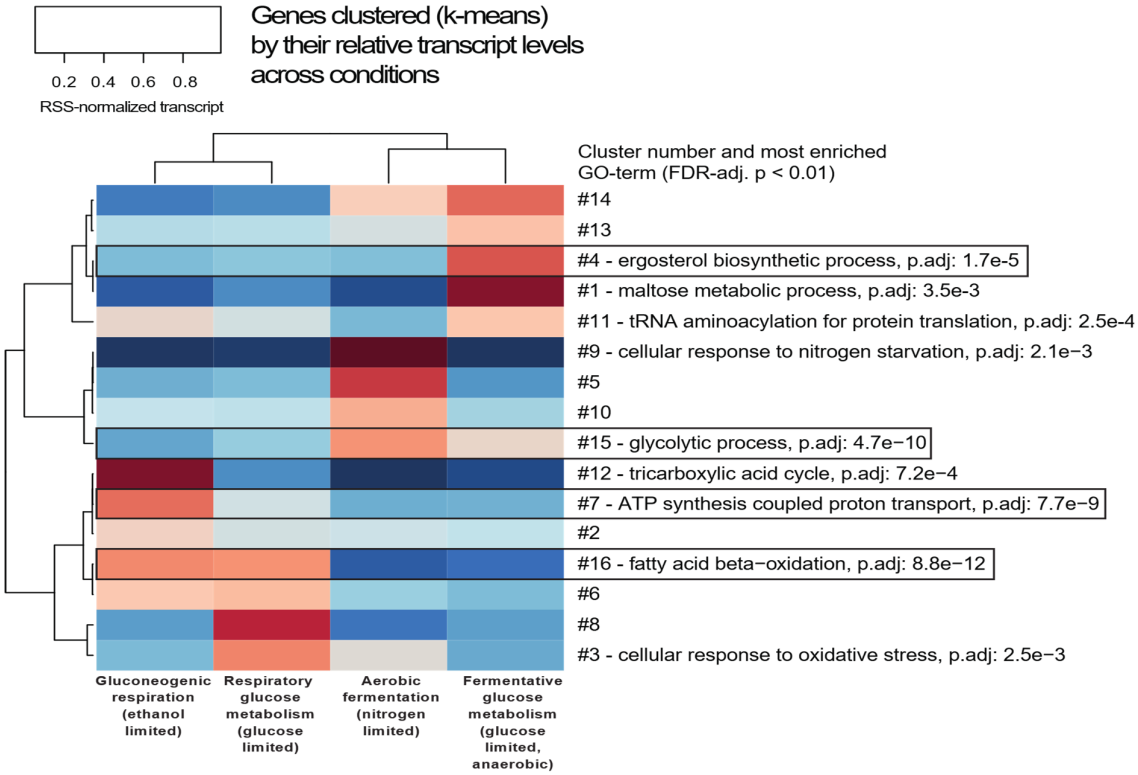


Figure 29 Clustering genes by their relative change in transcription level in between conditions. For clusters that have one or several significantly (FDR-adj $P < 0.01$) enriched GO terms, the top GO term is indicated with p.adj-value. Clusters containing central metabolic processes selected for further analysis with linear regressions are indicated by a black frame

importance" variable, defined as the product of the linear regression R2 value, the relative

contribution of the transcription factor (0-1) and the coefficient for activation or repression (-1,1).

Cluster 4, including ergosterol, steroid, sterol and lipid metabolic processes, showed a relatively high number of changes to the calculated "TF importance" values when comparing the different conditions. Hap1 was the only transcription factor that was identified in all conditions, while Gcn4 was identified in three and Oaf1 in two conditions (**Figure 30** A). All transcription factors showed positive importance, indicating an activating effect. The predictive score of this cluster in fermentative glucose metabolism (Ox,Glu-lim) using the model selected important transcription factors Ert1, Cat8, Gcn4 and Hap1 (Oaf1 is not selected by MARS as it is interchangeable with Hap1) showing an R2 = 0.68 (**Figure 30** C). This strongly indicates that the model explains much of the variability of the gene expression for these processes in his cluster. Contribution from each transcription factor can be displayed in a heatmap, showing the measured transcript levels as well as the binding signal of each transcription factor normalized column-wise (Z-score) (**Figure 30** E). Cluster 15 is enriched for glycolytic processes, and as expected, the glycolytic regulators Gcr1, Gcr2 and Tye7 are major contributors to the model's predictive power (**Figure 30** B). Also, Ert1 is an interesting predictor with negative TF importance value, indicating a repressor role in this cluster. In the linear regression, Gcr2 and Tye7 can be replaced by only Gcr1, resulting in Gcr1, Hap1, Ert1 and Rtg1 being identified as predictors in Glu-lim conditions. The model has an R2 = 0.66, again a strong indicator of high explanatory power (**Figure 30** E, F). In Cluster 7, we found an enrichment of genes involved in mitochondrial ATP biosynthesis, and these genes were upregulated in the two respiratory conditions (Glu-, and Eth-lim). Hap4 and Oaf1 were identified in both these conditions. Interestingly, Rgt1, previously identified as a repressor in glucose-limited conditions (reported in section 1.3, **Figure 6** as well as 7.1.2, **Figure 23**), was reported to have a repressive role also in both cluster 7 and cluster 16, providing some validation to the models.

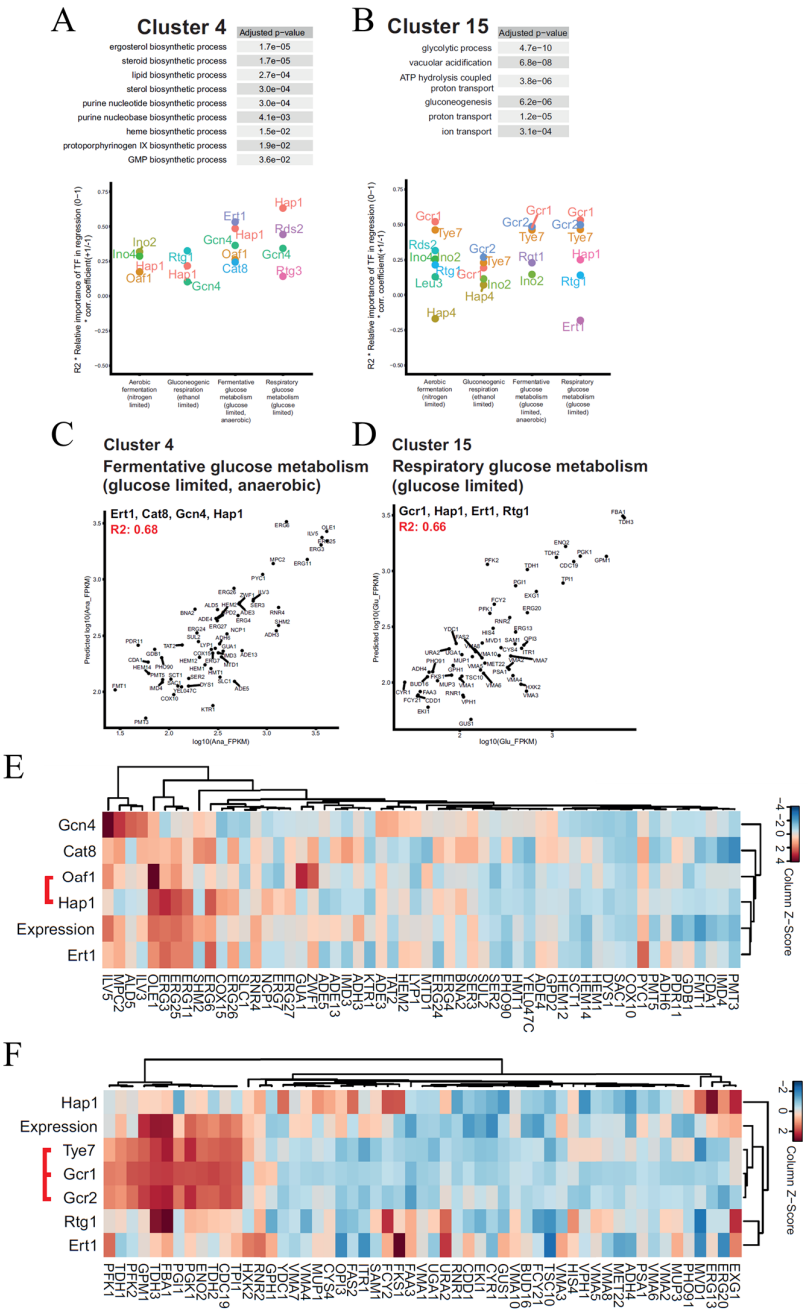


Figure 30 Clustering genes by relative expression. (A-B) All significant (*P*.adj < 0.05) GO terms for the clustered genes and the relative importance of the TFs selected. (C-D) Prediction plots showing the predicted transcript levels compared to the measured transcript levels from using the selected TFs. (E-F) Heatmaps demonstrate the measured transcript levels as well as binding signal of each TF normalized column-wise (*Z*-score). TFs linked by a red line under the heatmap have significant collinearity over the cluster genes and were demonstrated to be able replace the other.

In summary, I have demonstrated that we can build TRNs that not only confirm previously reported functions of the transcription factors, but also expand the target lists and in turn the TRNs. Varying the conditions is vital in the discovery of novel functions, and in combination with perturbation systems (e.g. deletions), this method allows studying the overall effect of a transcription factor and its direct targets. The functions of the transcription factor are often hidden in the perturbation systems, and what we see is the cells' adaptation to the environment. Computational models can help in our endeavor to understand how the transcription factors affect the transcript levels. These models can accurately identify transcription factors that interact with other transcription factors and what their regulatory role is (activator or repressor).

8 UTILIZATION OF TRNS

This chapter describes the incorporation of all our generated data into a combined database and toolbox, T-rEx. T-rEx is a user-friendly web application that can be used for finding regulatory modules or to perform in depth promotor studies. It can be used to investigate any of the already included transcription factors or to upload own data (generated according to our pipeline, described in section 6.3). It includes statistical modelling functionality, and I will exemplify how it may be used to identify and improve the engineering of cell factories. **Paper VII** describes the development and utility of T-rEx. **Paper VIII** is about the utility of our high-resolution database (included in T-rEx). This paper exemplifies how the toolbox and database can be used to fine-tune the expression of genes and how it can help improve the design of gRNAs in CRISPRi/a systems.

8.1 T-REX: A TOOLBOX FOR ANALYZING TRANSCRIPTION FACTORS

Along the way it became clear that we needed a toolbox to simplify and speed up visualizations of and interactions with the beautiful data we were generating. As most researchers in fundamental research are neither computational experts nor statisticians, we constructed the transcription factor explorer T-rEx to bridge this gap. T-rEx is an R Shiny web-application capable of visualization, summarizing statistics and analysis of transcription factor data.

T-rEx consist of four separate web pages with different functionality, where the data are visualized, analyzed or new data can be integrated. All presented data are generated from the ChIP-exo pipeline, although the derived data may have been treated differently between different analyses.

The first page contains a summary of the selected transcription factor, including a table of targets, the consensus motif and the identified sequences that generate the consensus motif. The read distribution around peaks (Peak distribution profile) and the read distribution around the TSS (read distribution profile) are also presented. The user can download the data for further analysis.

The second page contains a visualization tool for the Read profiles, mapped to genes in the ChIP-exo pipeline, and some additional features. Some of these features include data overlay of for example multiple transcription factors, the TATA/-like box, motif search, transcript levels and binding sites.

The third page contains the analysis page, including standard statistical tools that can be used for data interpretation. The analysis starts with the selection of a growth condition, GO-terms of interest and which dataset to use. The data are automatically treated correctly for use in any of the different statistical tools:

- **Fisher's exact test** uses a hypergeometric distribution to assess if two transcription factors are co-localized.
- **Heatmap** displays all the selected genes for each condition and the bound transcription factors.
- **Network plot** is generated as a visual overview of the selected GO-terms. The number of edges (gene connections) for each TF is used as weight for the node size of the TF.
- **Cluster test** uses the Partitioning Around Medoids (PAM) method, which is a form of k-means clustering. The two most prominent dimensions are displayed, and their individual separation is presented. The medoid coefficient indicates the contribution of the TF to the cluster, where a positive value indicates presence and negative value indicates absence.
- **Linear model** fits a model to describe gene expression using the TF binding data. This model does not take any TF-TF interactions into account. The linear model uses the number of bindings of each TF on each gene to allow for dynamics in the system.
- **Shared Targets** compares the targets of a single selected transcription factor to either one or several transcription factors.

The fourth page allows the user to include new data. The new data are integrated into the current data set and can be analyzed as previously described. The following chapters demonstrate the utility of T-rEx.

8.1.1 UTILITY OF T-REX: NETWORK IDENTIFICATION

T-rEx can be used to identify interesting TF networks. As an example, let's say that we are interested in ergosterol biosynthesis, phospholipid synthesis, pentose-phosphate shunt and fatty acid β-oxidation. To look closer at these pathways we select these GO-terms as search terms. We also select the Glu-lim condition and limit our gene set to Yeast8 genes (Lu et al. 2019). 64 genes are selected based on these criteria. First, we test the co-localization of transcription factors using Fisher's exact test. We find that Cbf1-Tye7, Gcr1-Tye7, Oaf1-Pip2 and Ino2-Ino4 are co-localized, and that the transcription factors Stb5, Sut1, Hap1, Gcr2, Gcn4, Ert1 and Cat8 are also enriched in these GO-terms. We visualize the GO-terms in a network plot to get an overview of the connections between the genes and the transcription factors (**Figure 31**). In the network, we observe similar trends as identified in the Fisher's test, e.g. many overlapping edges (genes) are observed between Ino2-Ino4, Oaf1-Pip2 and Cbf1-Tye7. However, in this plot also Hap1 and Gcn4 have overlapping targets. In the center of the plot, there is a group of genes with high numbers of bound transcription factors. Using the heatmap, we identify these transcription factors as Cbf1, Gcn4, Hap1, Oaf1, Pip2, Ino2, Ino4, Tye7 and Stb5. Using the Shared Targets function and the previously mentioned transcription factors, we find 7 genes bound by all TFs. Of these, we choose *ROX1* to take a closer look at the binding profile.

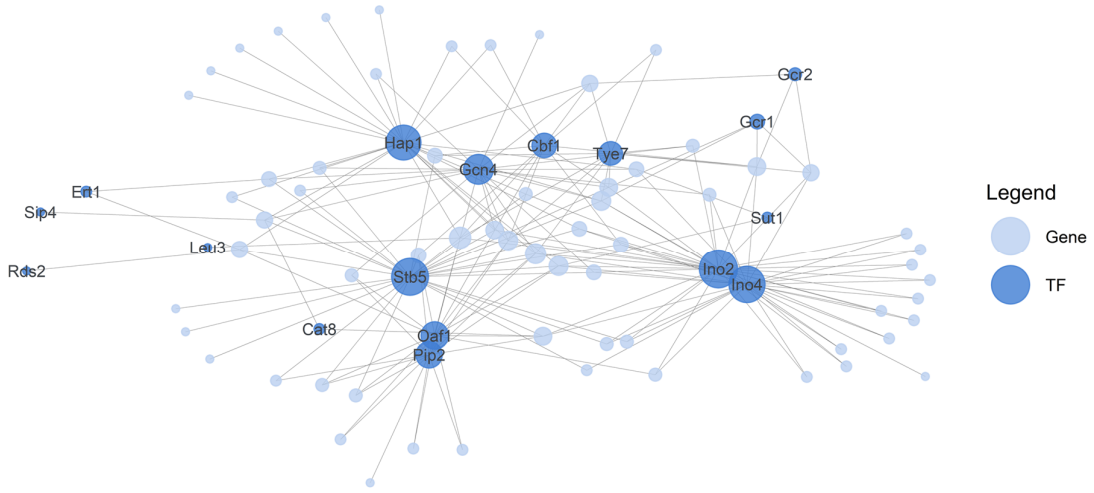


Figure 31 A network plot of the selected GO-terms in Glu-lim. The nodes are weighted by the number of gene connections. Several TFs group closely together, such as Ino2-Ino4, Oaf1-Pip2 and Tye7-Cbf1. Genes in the center are connected to many TFs. One of these genes is *ROX1*.

8.1.2 Utility of T-rEx: Promoter study

To look closer at the *ROX1* promoter and its bound transcription factors, we use the second page "Transcription Factor Binding Data". On the *ROX1* promoter, 9 different transcription factors are bound. Hap1 and Oaf1 show the strongest binding on the promoter followed by Stb5 (**Figure 32**). These three zinc-fingers share the motif CSGnnWW (S=G/C, N=any, W=A/T), which we observe at multiple locations within the strongest binding region. Hap1 prefers CCGnnAW, observed at three locations at the center of each Hap1 peak. Oaf1, Pip2 and Stb5 have almost identical motifs, and the binding of Oaf1-Pip2 but not Stb5 overlaps with the Hap1 binding. Stb5 has 4 peaks in the region, where we also observe the motif CGGnnWT. Tye7, Cbf1, Ino2 and Ino4 are all bHLH transcription factors that bind to the E-box motif CAnnTG. All four are bound at the same location, and we can find two E-box motifs there. Cbf1 and Tye7 have one additional peak, and we can find two E-box motifs within this peak. Gcn4 shows one binding site in the region and at the same location we can find the motif ATGACT, which is in agreement with the consensus motif of Gcn4.

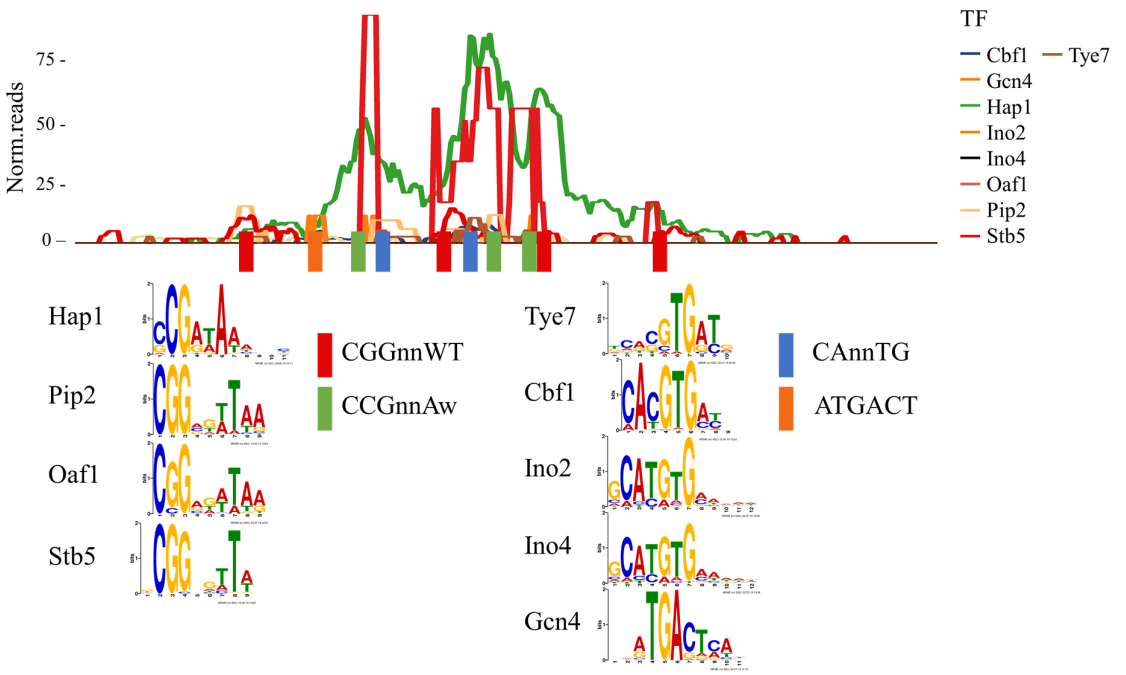


Figure 32 Binding of 9 TFs on the *ROX1* promoter. Hap1 and Oaf1-Pip2 show the strongest binding on this promoter and three distinct peaks can be seen, the motif of Hap1 can be found at all these locations. Stb5 has a lower binding strength but with at least 4 distinct peaks and we can find the motif at all these sites as well. The 4 bHLH Ino2, Ino4, Cbf1 and Tye7 are bound at the same location where we can find the E-box motif. Gcn4 is also bound to the promoter and a motif is identified at the center of the peak.

8.1.3 Utility of T-rEx: Identification of regulatory models

Now we are interested how well we can explain the transcript levels of this set of genes applying the linear model. In this case, the model explains the transcript levels well with an R2 of 0.65. We download the output from the model and see that the significant predictors are Ino2, Gcn4, Gcr2 and Hap1 (**Figure 33** A). Using the cluster function, we first test 5 clusters. However, one cluster is removed as it only contains one gene. For the 4 remaining clusters, we use a cutoff of 0.5 for the medoid coefficient to remove transcription factors with less predictive power. Cluster 1 is primarily regulated by Ino2 and Ino4 and includes genes involved in phospholipid and ergosterol metabolism. Cluster 2 contains 4 genes implicated in carbohydrate metabolic processes, and is primarily regulated by Ino2, Ino4, Cbf1, Gcr1, Gcr2 and Tye7. Cluster 3 comprises ergosterol and a small number of PPP genes, and is primarily regulated

only by Hap1. Cluster 4 contains β-oxidation genes and some PPP genes, and Oaf1-Pip2 are the major transcription factors for this cluster (**Figure 33** B).

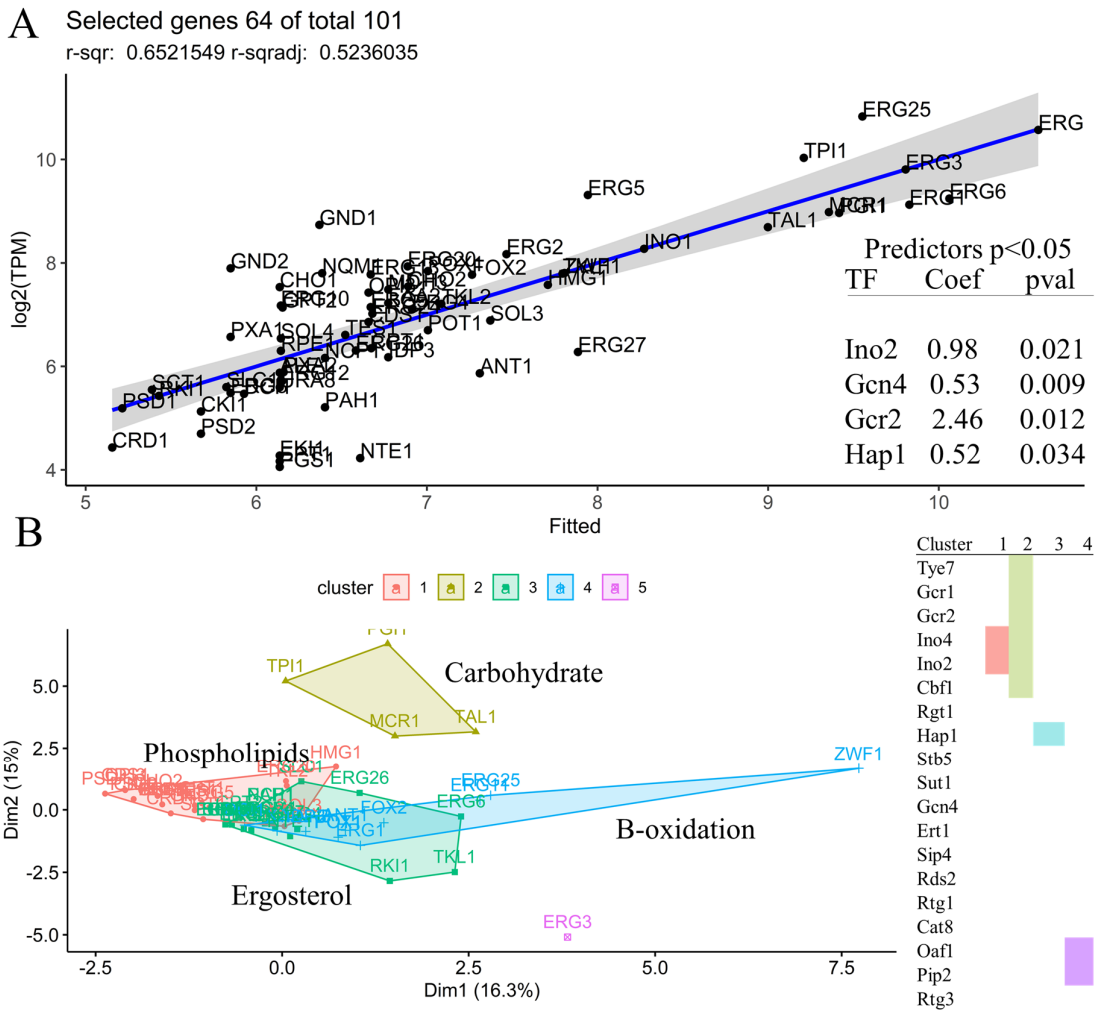


Figure 33 Transcript levels prediction and regulatory modules. A) Using a linear model, we can predict the transcript levels where the major contributors to the prediction are Ino2, Gcn4, Gcr2 and Hap1. B) Using clustering we can identify regulatory modules of genes clustered based on their identified peaks. 4 clusters are identified and based on the medoid coefficient we can determine which TFs have most important regulatory role over which cluster.

8.2 DESIGNING GRNAS BASED ON TRANSCRIPTION FACTOR BINDING

This study focuses on using CRISPRi/a to understand the interplay between transcription factor binding and binding of dCas9. CRISPRi/a uses a catalytically inactive Cas9, often referred to as dCas9 (endonuclease-deficient Cas9) (Perez-Pinera et al. 2013b; Qi et al. 2013). The gRNA consists of two regions a scaffold and a spacer. The scaffold interacts with the Cas9 while the spacer is complementary to the target DNA sequence. CRISPR interference or activation (CRISPRi/a) is a programmable tool for gene regulation. Such regulation enables both repression of the target gene when fused to a repressor domain, such as the mammalian transcriptional repressor domain Mxi (Bernards 1995), or activation when fused to an activator domain, such as the tripartite activator VPR (Chavez et al. 2015). This results in a CRISPR-based transcription factor (crisprTF) system. Achieving predictive and precise gene regulation is, however, challenging, and this is mainly due to the complexity of the regulatory processes and our limited understanding of it (Deaner et al. 2017; Jensen 2018). Can our large-scale studies of transcription factor binding sites aid in this predictive and precise gene regulation?

We used T-rEx to identify interesting clusters of genes and observed that many genes involved in central carbon metabolism were connected to the transcription factors Gcr1-Gcr2 and Tye7. There are 18 genes that are direct targets of these three transcription factors, and 10 were selected for further analysis (**Figure 34** A). On each of the 10 promoters, there are between 3 and 5 binding sites for the three transcription factors (**Figure 34** B). The promoters of each of these genes were cloned together with a GFP gene, which acted as reporter for analyzing the transcription factor-dCas9 interaction. dCas9-VPR is often used for gene activation and was used in this study. gRNAs were designed to bind either on or next to the identified TF motif, and expression of dCas9-VPR without a gRNA was used as control (**Figure 34** C). An example of how the gRNA targets sites are located is shown for the *ENO1* promoter (**Figure 34** B) where one gRNA target (ENO1-11) is located on top of a Gcr1-Gcr2 binding site, two gRNAs (ENO1-02 and ENO1-03) are targeted outside of any TF binding site, one gRNA target (ENO1-04) is located on top of a Tye7 (and Ino2-Ino4) binding site and one gRNA (ENO1-01) is targeted close to the TATA box. Cells were grown in batch cultivations for 35 h and their fluorescence was analyzed over time and normalized to the OD. A typical GFP profile of 6 different strains, 5 of which were expressing gRNAs, can be seen (**Figure 34** D). While some gRNAs had no effect (ENO1-21), others resulted in upregulation (ENO1-31) or in downregulation (ENO1-11 and ENO1-41) of the reporter.

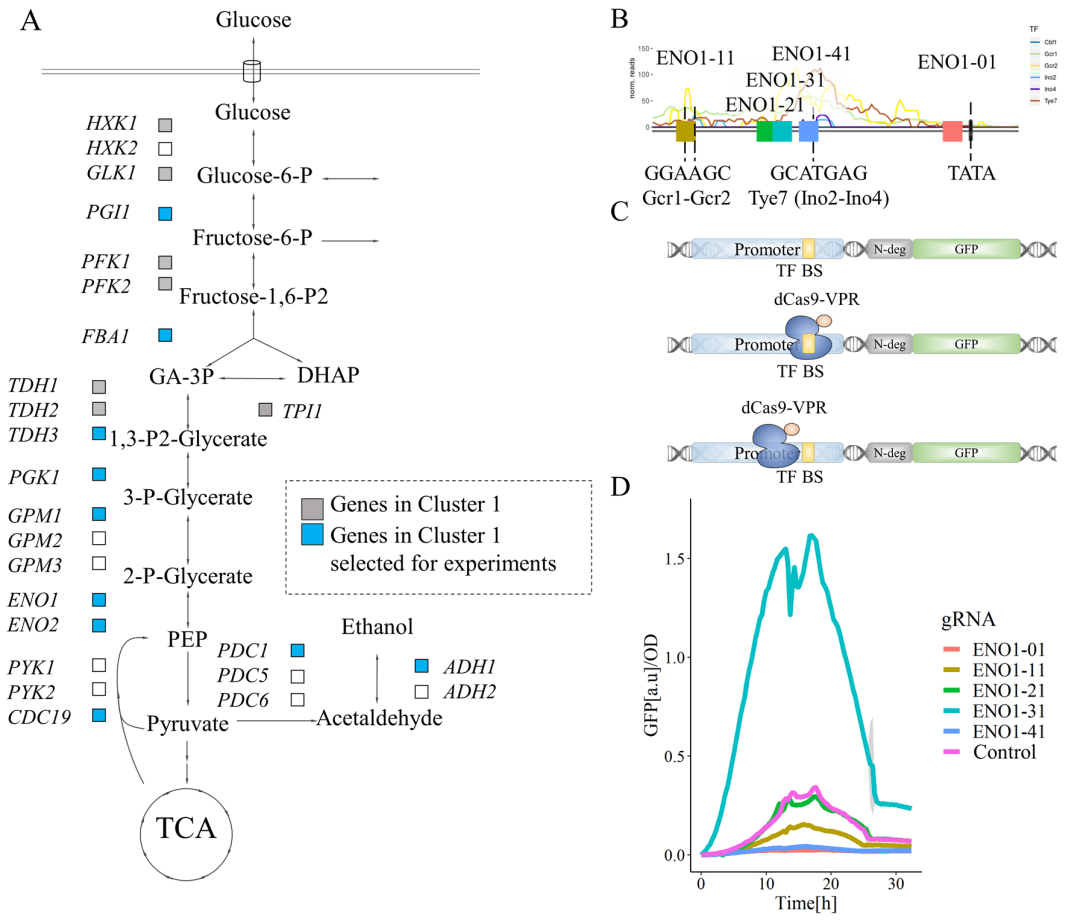


Figure 34 Design of the study. A) The 18 genes bound by Gcr1-Gcr2 and Tye7 where 10 genes were selected for experiments. B) The *ENO1* promoter with binding sites and binding profiles of several transcription factor, motifs marked with dotted line. gRNA sites are color coded to fit the GFP expression in D). C) How the gRNA sites are selected: either on top of a motif or next to a motif. D) The GFP expression of the gRNA expressing strains and the control.

8.2.1 Effect on dCas9-VPR and transcription factor positioning on gene expression.

To investigate the overall gRNA effect, GFP fluorescence data of the strains expressing gRNAs targeted to one of the motifs and strains expressing the non-motif targeted gRNAs were separately pooled. Results from glucose phase and ethanol phase was separated, here we demonstrate the results from ethanol phase. **Figure 35** A shows the fluorescence fold-change (FC) of the strains expressing gRNAs binding motifs and non-motif regions, respectively. A student's t-test was used on the log2FC of the GFP expression. When a gRNA was targeted to one of the motifs, the GFP expression level was either unchanged or decreased, while if bound to a non-motif region, the GFP expression was increased (**Figure 35** A). Since there seemed to

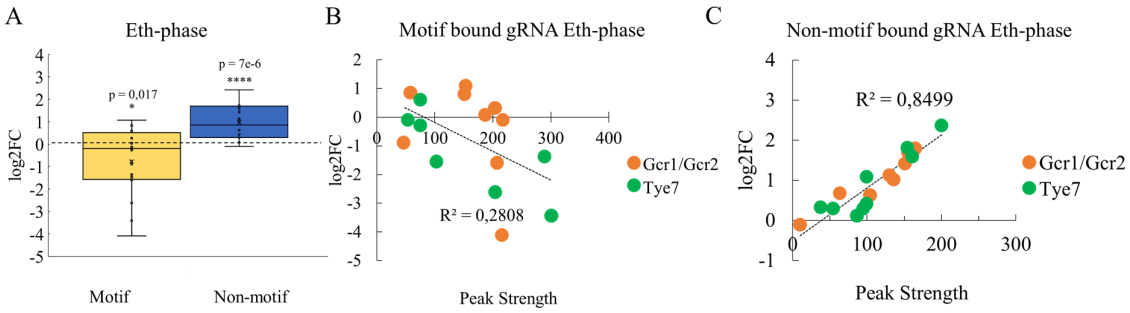


Figure 35 The log2 fold change (log2FC) of the GFP expression, is designated to either the motif or the non-motif group. A) A T-test was used to compare if the groups were either up- or downregulated compared to no change in GFP expression (dotted line). p designates the log2FC p-value of the students t-test. B) Correlation of the peak strength of the transcription factors and the fluorescence output in each case with binding of the gRNA to a motif or non-motif.

be different levels of both increased and decreased expression, we wondered if the strength of transcription factor binding was a determinant of the GFP expression.

8.2.2 Effect of adjacent transcription factor binding strength is a determinant of GFP expression

The transcription factor binding strength is the measure of probability that a TF will occupy a specific site during the ChIP-exo experiment and is measured as read count (i.e. peak height). We analyzed the two groups (motif bound gRNA and non-motif bound gRNA) separately and found a striking correlation. When a gRNA was bound on top of a motif with low transcription factor binding, the GFP expression was unchanged or only decreased slightly compared to control, and as the peak strength increases so does the decrease in GFP expression (**Figure 35** B). The opposite is true for no motif bound gRNA, where we connected the gRNA to the nearest peak. A low peak strength results in no change in GFP expression, as the peak strength increases so does the GFP expression (**Figure 35** C).

8.2.3 Competition and cooperativity

The results indicate that there is a direct connection between the binding strength of a transcription factor and the expression change resulting from the CRISPRi/a. As we see two different states of the CRISPRa system, both repression and activation, we looked into competition and cooperativity. dCas9 can stay bound to DNA up to 3 h in eukaryotic cells (Ma et al. 2016), thus having a slow dissociation, whereas transcription factors stay bound for milliseconds to seconds, thus having a fast dissociation (Swift and Coruzzi 2017). dCas9 acts as a competitor to the transcription factor that through steric hindrance blocks the transcription factor from binding. This means that the dCas9-VPR acts as a de facto repressor the stronger

the occupancy of the TF binding is, and it drives out the endogenous transcriptional activator and cannot compensate for this loss if the transcription factor is a strong activator.

However, in the case of activation the effect seems to be due to cooperativity. Transcription factors can act together in a cooperative manner (see section 1.3) and dCas9 has shown to act similarly (Perez-Pinera et al. 2013a). It is then possible that the transcription factors have a profound impact on the surrounding and thereby impact the dCas9 binding in a positive cooperative manner. These results indicate the importance of choosing the gRNA site in relation to other binding sites, as dCas9-VPR targeting activator binding sites has a higher probability to decrease expression levels while dCas9 binding outside transcription factor binding sites has a higher probability in activating transcription levels (**Figure 36**).

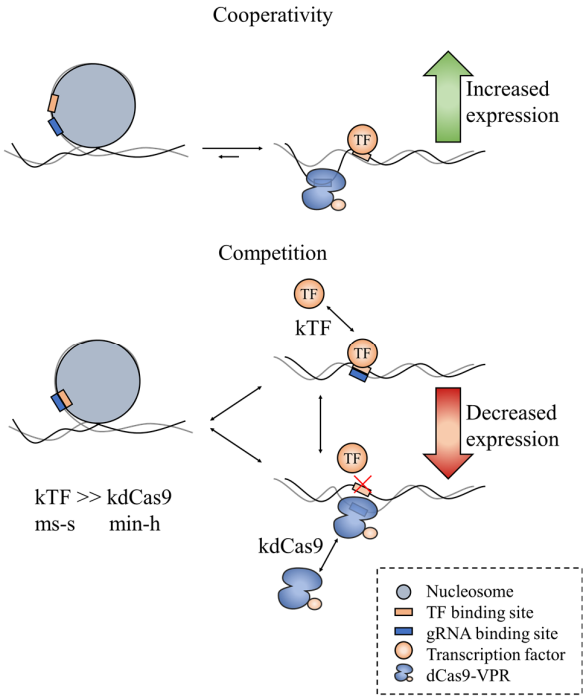


Figure 36 Cooperativity and competition between dCas9 and activating transcription factors. Binding of dCas9 close to a TF results in increased expression where also the binding strength influences the resulting increase. Binding of dCas9 on top of a TF motif results in competition as the dissociation of dCas9 is much weaker than that of the TF, the TF is outcompeted, and the expression decreases.

In summary, T-rEx is a versatile toolbox that can be used for both in detail promoter study, co-localization studies, identification of common targets, computational modelling as a prediction

tool for gene expression and to identify regulatory modules using clustering functions. The high-resolution binding that we can display in T-rEx can aid us in the design of gRNAs. We have shown that expression levels of a target gene are in direct connection with the chosen gRNA site and the transcription factor binding both in terms of overlap and in binding strength.

9 INTO THE FUTURE

Here I will summarize the work presented in this thesis and provide an outlook into what might be next in the endeavor of unravelling the transcriptional regulatory network.

9.1 CONCLUSIONS

This thesis contributes to the scientific field by providing a holistic view on a subset of the transcriptional regulatory networks in *S. cerevisiae*. We developed a small bioreactor system for high-throughput strain characterization. The bioreactor system proved to be comparable to a commercially available system. The multiplexing and miniaturizing increased the throughput and at the same time reduced the need for materials (such as media) and the complexity for setup (**Paper I**). Setting up a new method, ChIP-exo, for studying protein-DNA interactions proved to be a laborious but worthwhile task. The generated data were both of high quality and high resolution, and therefore we could identify stress response as a novel functionality for the transcription factor Cst6 (**Paper II**). Working with this high-quality data also required a robust system for the analysis. We developed a bioinformatics pipeline with high emphasis on quality control that is applicable to any ChIP-exo data set (**Paper III**). Next, we expanded the set of transcription factors to some of the major contributors to lipid metabolism, namely Ino2, Ino4, Oaf1, Pip2 and Hap1. By using gene set enrichment analysis, we could identify novel pathways for the studied transcription factors and expand their TRNs (**Paper IV**). We focused on an important key regulator of the pentose phosphate pathway, Stb5, to identify its regulatory role. Here, the different growth conditions played an important role in the regulation in different metabolic states. We found that NADPH became a limiting factor only when glycolysis was active, and not in gluconeogenesis (**Paper V**). We employed statistical methods and regression models to understand and predict regulatory pathways. For this analysis, we needed more transcription factors and therefore we included the following transcription factors that had been implicated in central carbon metabolism: Cat8, Cbf1, Ert1, Gcn4, Gcr1, Gcr2, Hap4, Lue3, Rds2, Rgt1, Rtg1, Rtg3, Sip4, Sut1 and Tye7. Linear models allowed good predictive power to relevant subsets of the central carbon metabolism. This modelling approach could also accurately describe some previously known biological functions which provides strength to the overall model (**Paper VI**). The vast amounts of generated data and the many implications and utilizations they can have on the research field encouraged us to create a toolbox for transcription factor visualization and analysis. The toolbox can be used for in-depth promoter studies or to identify subsets of regulatory pathways (**Paper VII**). The newest technique for engineered gene regulation (CRISPRa/i) has shown promise in metabolic engineering. It is however problematic to generate gRNAs with the desired effect. To address this, we used our high-resolution transcription factor data to design gRNAs that bind either on or adjacent to motifs. We found that the strength of transcription factor binding has a profound effect on the gene expression levels and that the transcription factor and the dCas9-VPR either compete or cooperate with each other (**Paper VIII**).

9.2 WHERE DO WE GO FROM HERE?

I started out pondering about life itself, so let's continue there. A yet largely unanswered question is: Can we understand life well enough to design it to our purpose? My firm belief is yes. Earth has existed for 4.5 billion years, humans only for 200 000. But in these years, we have accomplished so much! And now, 7 billion people live on the planet, and each year offers 7 billion human-years of ingenuity and advances. Humans and human-built AIs will most probably get us there. We have the question, but can we see the full extent of it? Probably not, our brains are not made for containing all that information. At the same time, an AI cannot ask the questions, yet, and therefore both are required to find the answer. Today we are still far, far away, and on our journey life as we have defined it is changing. How do we define life, when the border of what we can design and what nature designed is fading? Robots that work autonomously and that can reproduce by building new robots, are they alive? When we build bio-robots that use biological fragments to function, are they alive? When will biology become "just another" technology?

A question I often get is "why do you want to do this? Why not just use the mRNA (transcriptomics data) from different chemostat conditions, then you know what the promoter does and how it responds". Yes, in part this is true. But what we want to do is to map all interactions on the promoter. From the mapping we would obtain a vocabulary, telling us which interactions occur on the promoter. From this we can build models that can translate this vocabulary and finally we will have models powerful enough to tell us how a promoter responds in different conditions.

What we have discovered is something that I think no one else has seen. By combining the information from the different chemostat conditions that we have studied, one conclusion becomes evident: If a motif exists on a promoter, at some point and in a certain environment, the transcription factor that recognizes that motif will bind. The environmental condition for this might be highly specific, such as in the Cst6 case, but nevertheless it will be there. This is something we also showed in **Paper XIV,** where we engineered promoters for higher expression in acidic environments. We did so by introducing more binding sites of a specific transcription factor on the promoter. By using ChIP-qPCR we could show that the introduced binding sites were in fact bound by the transcription factor of interest.

However, to identify all these conditions and to study all transcription factors is just not feasible. Still, this work clearly shows that careful selection of a small number of different conditions can greatly enhance our ability to understand transcription factor regulation, and the variation in binding a certain transcription factor can have.

One thought on how to move the field forward is demonstrated in **Figure 37**. It starts by using ATAC-seq (but maybe upgraded to ATAC-exo) (Buenrostro et al. 2013). This method allows to find all transcription factor bindings that occur throughout the genome and at the same time identifies all the nucleosome binding events. In the data, we can then remove the nucleosome binding and thereby provide a profile of all binding events due to DNA binding

proteins (Li et al. 2019). Unfortunately, as with any method there are drawbacks. We cannot identify proteins that bind to other proteins nor if they are bound to nucleosomes, neither can we identify which protein that is bound. We therefore do not know the identity of the proteins or even if they are transcription factors. One solution to this problem could be DNA-binding domain motif models, which through statistical and thermodynamical modelling predict all the binding sites of a DNA-binding protein based on the protein structure. So far, the models improve based on the available data, such as ChIP-seq (Zamanighomi et al. 2017). This will however not allow us to find new binding sites when studying new conditions, but progress has been made towards predictive models without preexisting data (Farrel and Guo 2017). Identifying the DNA binding protein in ATAC-seq data could be possible by generating a transcription factor binding score based on ChIP-exo data and DNA binding domain motif.

In this work we have used GO-terms to group genes within metabolic processes in order to investigate effects of transcription factors in a more specific manner compared to looking at the whole genome. Although this clearly provided advantages for example to identify which general processes a transcription factor regulates, there are clearly limitations. For instance, Oaf1-Pip2 regulates genes within β-oxidation, but also regulates surrounding processes such as fatty acid metabolism and malate metabolism. In fact, this transcription factor pair also regulates individual genes, such as *ZWF1* whose gene product produces a metabolite (NADPH) that is required for regeneration of thioredoxin/glutathione that are in turn needed for the detoxification of H_2O_2 generated in the peroxisomal β-oxidation. This indicates that GO-terms are not sufficient for identifying TRNs. Integration of all TRNs into GEMs, which in recent years have improved in predictive power (Cardoso et al. 2018; Sanchez et al. 2017), will hopefully in time allow us to predict outcomes on a cellular level.

This was an initial goal when I started my PhD. However, the TRNs turned out to be more complex than anticipated, and there was not enough time to reach to this desirable goal. Currently, only the metabolic enzymes are included, GEMs therefore need additional levels of integration. The yeast GEM model contains roughly 800 genes, but over 2000 genes were found to be bound by any of the 21 transcription factors that we have studied. Gcn4 has in itself over 1000 gene targets, where over 100 genes are connected to transcription, and many encode themselves transcription factors i.e. *CST6*, *HAP4*, *INO2*, *LEU3* and *PIP2* (to mention a few). Also, Ino2 and Ino4 have around 1000 gene targets. The role of transcription factors with these many targets is likely to maintain a basal level of expression, and they only constitute one of many transcription factors regulating each target gene. Cbf1 is another transcription factor with over a 1000 gene targets. However, in a similar ChIP-exo study using YPD as growth media, only 102 targets were identified for Cbf1 (Rossi et al. 2018a), again proving that multiple conditions are needed for identifying the true nature of a transcription factors regulatory network.

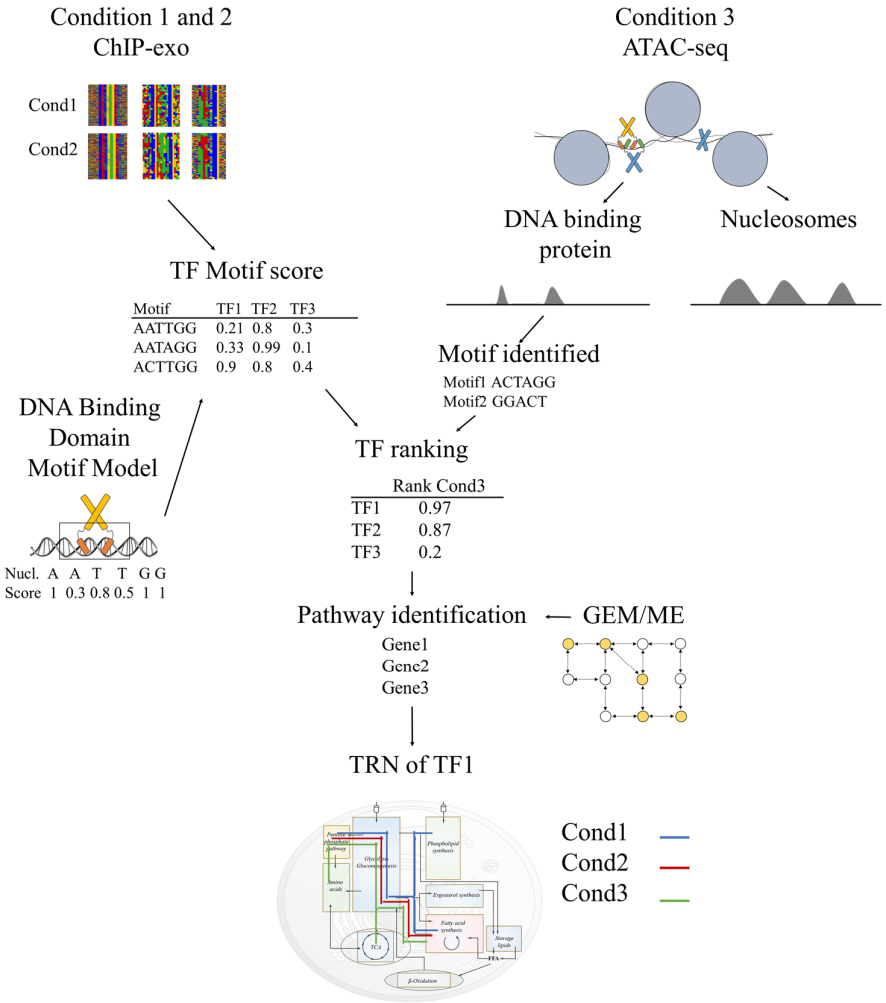


Figure 37 Illustration of TRN identification in newly studied conditions. Integration of ChIP-exo data that reveals the many binding sites in studied conditions with DNA binding domain motif models allows for the identification of all the possible binding sites a transcription factor can have. ATAC-seq allows for the identification of all DNA-binding events throughout the genome. By detecting the DNA binding protein regions, we can identify motifs. Combining the TF Motif Score with the identified motif we can rank which transcription factor is the most probably bound. Thereafter we use GEM/ME models to predict the new TRN.

In human cells, transcription factors have shown to have a dominant role in the control of specific cell states and that they are capable of reprogramming cell states when expressed in various cell types (Lee and Young 2013). For instance, reprogramming of somatic cells to embryonic stem cells is done by expression of four transcription factors, namely Oct4, Sox2, Klf4, and c-Myc (Orkin and Hochedlinger 2011). The human bHLH transcription factor TAL1 is an oncogenic transcription factor that is overexpressed in 40-60% of the cells in leukemia (Sanda et al. 2012). The human transcription factor c-Myc is overexpressed in many tumor cells where it accumulates in the promoter regions of most active genes, recruiting the transcription elongation factor P-TEFb, and causes transcriptional amplification (Lin et al.

2012). Loss of function of the human transcription factor AIRE can lead to autoimmune diseases where only a fraction of the tissue antigens are expressed (Akirav et al. 2011). In the ageing cell, the Forkhead transcription factor FOXO (in yeast Fkh1 and Fkh2), plays a key role in stress response and autophagy. FOXO is shown to regulate lifespan in many species such as *H. vulgaris*, *C. elegans* and *D. melanogaster* (Martins et al. 2016) as well as in *S. cerevisiae* (Postnikoff et al. 2012). Translating what we have learnt from yeast to humans will in time help us treat diseases in a way that previously was not possible. Understanding how the transcription factors act on the promoters and how groups of transcription factors work together might give a new level to disease treatment by controlling the regulatory function that is underlying the problem.

When it comes to understanding the life of cells, transcription factors play an essential role and is a central part of how life is defined for that cell. So far, we are getting glimpses, snapshots, of how life works at specific conditions and time points. But the picture becomes clearer for each transcription factor we study. Once the network evolves, it will no longer be a picture, we will have a movie, with life in action right before our eyes. And it is our job to decode this movie so that everyone can watch it.

I see a bright future ahead, where advancements in biology, mathematics, computer technology and AI will join forces to solve our biggest questions in designing and changing life.

10 ACKNOWLEDGMENTS

First and foremost, I would like to thank all the people that has been around me through these years and without you this thesis would not have seen the light of day.

Jens, my main supervisor, for being so open to my ideas and that you always support me, even if I go a bit outside the scope of what I am "supposed" to do. I am so grateful to have had this opportunity to work with you, your positivity and curiousness inspires. Verena, my co-supervisor, for being a great mentor, friend and for keeping me on the right track and maintaining my scientific research solid. Christer, my examiner, who has read and approved most of the things I have done as a PhD student but also for great fika discussions. This work would not have been possible without my co-workers and collaborators and so I want to give a special thanks to: Michael Gossing taking me in as a fresh PhD student and who taught me much about the scientific field, Guodong Liu for being a great mentor, Christoph Börlin my rubber duck, Raphael Ferreira for our fun and educational entrepreneurial attempts, David Hansson for making D2 Biotech happen, Petter Holland, Yongjun Wei, Liming Ouyang, Yasaman Dabirian, Arun Rajkumar and Florian David. I want to thank the SysBio support team for making my life easier both in the lab and outside the lab: Martina, Erica, Anne-Lise, Angelica, Emelie, Fredrik, Marie, Shaghayegh and Joakim.

In the SysBio community there are plenty of people that has made my working life fantastic through discussions, lunches, social activities and so you are all an inspiration for how a scientific community should be: Dina, Stefan T, Yassi, Benjamin, JensC, Martin, Bouke, Leonie, Paulo, Yun, Jonathan, Alexej, Ivan, Stefan H, Yating, Jichen, John, Xin, Rui, Kate, Sylvain, Promi, Zongjie, Yongjin, Zhiwei, Rosemary, Feiran, Max, Ela, Linnea Ö, Ivan D, Tao, Ievgeniia, Tyler, Louis, Dimitra, Eugene, Lucy, Olena, Pinar, Jan, Avlant, Carl, Johan, John, Filip, Veronica, Oliver, and many more that I have met throughout my studies.

I have had the great opportunity to make many friends who I have shared science, skiing, climbing, partying, singing, dancing, traveling and many other adventures with and all of you deserve a special thanks as you will forever have a place in my heart: Ana, Alex, Flå, Michi, Petri, Gattino, Leif, Isa, Ed, Elle, Christoph and Rapha.

To my family. Mum, Dad and Niklas, thanks for the encouragement, support and for allowing me to seek my own adventures to discover who I am. My extended family, Aliz, Melker, Hannes, Ebba, Magnus, Frida, Tilli, Rune, Helga, Thomas, Kristina, Maria and Peter. And to Lurvas, our happy, fluffy, cuddly companion.

Adrian, I haven't known you for that long yet, but I know how much joy you already bring to my life. Isabelle, my crazy, happy little wildling you make my life filled with joy and inspiration of how wonderful the world is. Linnea, it's hard to put in words what you have done for me and what you mean to me. Your love, enthusiasm, positivity, scientific advices, and your eager strive for making our live's better. You truly bright up my sky and you make me into the best person I can be.

11 REFERENCES

Adams CC, Workman JL. 1995. Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. Mol Cell Biol 15(3):1405-21.

Akirav EM, Ruddle NH, Herold KC. 2011. The role of AIRE in human autoimmune disease. Nature Reviews Endocrinology 7(1):25-33.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biol 11(10):R106.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT and others. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25-9.

Baker HV. 1986. Glycolytic gene expression in Saccharomyces cerevisiae: nucleotide sequence of GCR1, null mutants, and evidence for expression. Mol Cell Biol 6(11):3774-84.

Bergman A, Vitay D, Hellgren J, Chen Y, Nielsen J, Siewers V. 2019. Effects of overexpression of STB5 in Saccharomyces cerevisiae on fatty acid biosynthesis, physiology and transcriptome. FEMS Yeast Res 19(3):foz027.

Bernards R. 1995. Transcriptional Regulation: Flipping the Myc switch. Current Biology 5(8):859-861.

Black RA, Blosser MC. 2016. A Self-Assembled Aggregate Composed of a Fatty Acid Membrane and the Building Blocks of Biological Polymers Provides a First Step in the Emergence of Protocells. Life (Basel) 6(3):33.

Bohm S, Frishman D, Mewes HW. 1997. Variations of the C2H2 zinc finger motif in the yeast genome and classification of yeast zinc finger proteins. Nucleic Acids Res 25(12):2464-9.

Bourot S, Karst F. 1995. Isolation and characterization of the Saccharomyces cerevisiae SUT1 gene involved in sterol uptake. Gene 165(1):97-102.

Breaker RR. 2012. Riboswitches and the RNA world. Cold Spring Harbor perspectives in biology 4(2):a003566.

Brindle PK, Holland JP, Willett CE, Innis MA, Holland MJ. 1990. Multiple factors bind the upstream activation sites of the yeast enolase genes ENO1 and ENO2: ABFI protein, like repressor activator protein RAP1, binds cis-acting sequences which modulate repression or activation of transcription. Mol Cell Biol 10(9):4872-85.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature methods 10(12):1213-1218.

Cafferty BJ, Fialho DM, Hud NV. 2018. Searching for Possible Ancestors of RNA: The Self-Assembly Hypothesis for the Origin of Proto-RNA. In: Menor-Salván C, editor. Prebiotic Chemistry and Chemical Evolution of Nucleic Acids. Cham: Springer International Publishing. p 143-174.

Calkhoven CF, Ab G. 1996. Multiple steps in the regulation of transcription-factor level and activity. Biochem J 317 (Pt 2)(Pt 2):329-42.

Cardoso JGR, Jensen K, Lieven C, Laerke Hansen AS, Galkina S, Beber M, Ozdemir E, Herrgard MJ, Redestig H, Sonnenschein N. 2018. Cameo: A Python Library for Computer Aided Metabolic Engineering and Optimization of Cell Factories. ACS Synth Biol 7(4):1163-1166.

Chavez A, Scheiman J, Vora S, Pruitt BW, Tuttle M, P R Iyer E, Lin S, Kiani S, Guzman CD, Wiegand DJ and others. 2015. Highly efficient Cas9-mediated transcriptional programming. Nature methods 12(4):326-328.

Chelstowska A, Butow RA. 1995. RTG Genes in Yeast That Function in Communication between Mitochondria and the Nucleus Are Also Required for Expression of Genes Encoding Peroxisomal Proteins. Journal of Biological Chemistry 270(30):18141-18146.

Chen M, Hancock LC, Lopes JM. 2007. Transcriptional regulation of yeast phospholipid biosynthetic genes. Biochim Biophys Acta 1771(3):310-21.

Chen M, Lopes JM. 2007. Multiple basic helix-loop-helix proteins regulate expression of the ENO1 gene of Saccharomyces cerevisiae. Eukaryot Cell 6(5):786-96.

Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M and others. 1998. SGD: Saccharomyces Genome Database. Nucleic Acids Research 26(1):73-79.

Chirala SS, Zhong Q, Huang W, al-Feel W. 1994. Analysis of FAS3/ACC regulatory region of Saccharomyces cerevisiae: identification of a functional UASINO and sequences responsible for fatty acid mediated repression. Nucleic Acids Res 22(3):412-8.

Comer FI, Hart GW. 1999. O-GlcNAc and the control of gene expression. Biochim Biophys Acta 1473(1):161-71.

Cottier F, Raymond M, Kurzai O, Bolstad M, Leewattanapasuk W, Jimenez-Lopez C, Lorenz MC, Sanglard D, Vachova L, Pavelka N and others. 2012. The bZIP transcription factor Rca1p is a central regulator of a novel CO(2) sensing pathway in yeast. PLoS Pathog 8(1):e1002485.
Crocker J, Preger-Ben Noon E, Stern DL. 2016. Chapter Twenty-Seven - The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution. In: Wassarman PM, editor. Current Topics in Developmental Biology: Academic Press. p 455-469.
Cullen PJ, Xu-Friedman R, Delrow J, Sprague GF. 2006. Genome-wide analysis of the response to protein glycosylation deficiency in yeast. FEMS Yeast Res 6(8):1264-73.
Dang W, Sutphin GL, Dorsey JA, Otte GL, Cao K, Perry RM, Wanat JJ, Saviolaki D, Murakami CJ, Tsuchiyama S and others. 2014. Inactivation of yeast Isw2 chromatin remodeling enzyme mimics longevity effect of calorie restriction via induction of genotoxic stress response. Cell Metab 19(6):952-66.
Davie JK, Trumbly RJ, Dent SY. 2002. Histone-dependent association of Tup1-Ssn6 with repressed genes in vivo. Mol Cell Biol 22(3):693-703.
de Boer CG, Hughes TR. 2011. YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. Nucleic Acids Research 40(D1):D169-D179.
Deaner M, Mejia J, Alper HS. 2017. Enabling Graded and Large-Scale Multiplex of Desired Genes Using a Dual-Mode dCas9 Activator in Saccharomyces cerevisiae. ACS Synthetic Biology 6(10):1931-1943.
Duarte NC, Herrgard MJ, Palsson BO. 2004. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. Genome Res 14(7):1298-309.
Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95(25):14863-8.
Erickson HP. 2009. Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy. Biol Proced Online 11:32-51.
Farrel A, Guo JT. 2017. An efficient algorithm for improving structure-based prediction of transcription factor binding sites. BMC Bioinformatics 18(1):342.
Fazio A, Jewett MC, Daran-Lapujade P, Mustacchi R, Usaite R, Pronk JT, Workman CT, Nielsen J. 2008. Transcription factor control of growth rate dependent genes in Saccharomyces cerevisiae: a three factor design. BMC Genomics 9:341.
Fernandes L, Rodrigues-Pousada C, Struhl K. 1997. Yap, a novel family of eight bZIP proteins in Saccharomyces cerevisiae with distinct biological functions. Mol Cell Biol 17(12):6982-93.
Furukawa K, Heinzle E, Dunn IJ. 1983. Influence of oxygen on the growth of Saccharomyces cerevisiae in continuous culture. Biotechnol Bioeng 25(10):2293-317.
Garcia-Gimeno MA, Struhl K. 2000. Aca1 and Aca2, ATF/CREB activators in Saccharomyces cerevisiae, are important for carbon source utilization but not the response to stress. Mol Cell Biol 20(12):4340-9.
Gietz RD, Woods RA. 2001. Genetic transformation of yeast. Biotechniques 30(4):816-20, 822-6, 828 passim.
Gitter A, Siegfried Z, Klutstein M, Fornes O, Oliva B, Simon I, Bar-Joseph Z. 2009. Backup in gene regulatory networks explains differences between binding and knockout results. Mol Syst Biol 5:276.
Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M and others. 1996. Life with 6000 genes. Science 274(5287):546, 563-7.
Guo Y, Mahony S, Gifford DK. 2012. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. PLoS Comput Biol 8(8):e1002638.
Hampsey M. 1997. A review of phenotypes in Saccharomyces cerevisiae. Yeast 13(12):1099-133.
Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J and others. 2004. Transcriptional regulatory code of a eukaryotic genome. Nature 431(7004):99-104.
He B, Tan K. 2016. Understanding transcriptional regulatory networks using computational models. Curr Opin Genet Dev 37:101-108.
Hedges D, Proft M, Entian KD. 1995. CAT8, a new zinc cluster-encoding gene necessary for derepression of gluconeogenic enzymes in the yeast Saccharomyces cerevisiae. Mol Cell Biol 15(4):1915-22.
Hickman MJ, Winston F. 2007. Heme levels switch the function of Hap1 of Saccharomyces cerevisiae between transcriptional activator and transcriptional repressor. Mol Cell Biol 27(21):7414-24.
Hiltunen JK, Mursula AM, Rottensteiner H, Wierenga RK, Kastaniotis AJ, Gurvitz A. 2003. The biochemistry of peroxisomal beta-oxidation in the yeast Saccharomyces cerevisiae. FEMS Microbiol Rev 27(1):35-64.
Hinnebusch AG. 1988. Mechanisms of gene regulation in the general control of amino acid biosynthesis in Saccharomyces cerevisiae. Microbiol Rev 52(2):248-73.
Hu L, Grosberg AY, Bruinsma R. 2008. Are DNA transcription factor proteins maxwellian demons? Biophys J 95(3):1151-6.
Hughes TR, de Boer CG. 2013. Mapping yeast transcriptional networks. Genetics 195(1):9-36.

Huisinga KL, Pugh BF. 2004. A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in Saccharomyces cerevisiae. Mol Cell 13(4):573-85.
Inukai S, Kock KH, Bulyk ML. 2017. Transcription factor-DNA binding: beyond binding site motifs. Curr Opin Genet Dev 43:110-119.
Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. Journal of Molecular Biology 3(3):318-356.
Jensen MK. 2018. Design principles for nuclease-deficient CRISPR-based transcriptional regulators. FEMS Yeast Research 18(4).
Jewett MC, Workman CT, Nookaew I, Pizarro FA, Agosin E, Hellgren LI, Nielsen J. 2013. Mapping condition-dependent regulation of lipid metabolism in Saccharomyces cerevisiae. G3 (Bethesda) 3(11):1979-95.
Juan LJ, Walter PP, Taylor IC, Kingston RE, Workman JL. 1993. Nucleosome cores and histone H1 in the binding of GAL4 derivatives and the reactivation of transcription from nucleosome templates in vitro. Cold Spring Harb Symp Quant Biol 58:213-23.
Juhnke H, Krems B, Kötter P, Entian K-D. 1996. Mutants that show increased sensitivity to hydrogen peroxide reveal an important role for the pentose phosphate pathway in protection of yeast against oxidative stress. Molecular and General Genetics MGG 252(4):456-464.
Kachroo AH, Laurent JM, Yellman CM, Meyer AG, Wilke CO, Marcotte EM. 2015. Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. Science 348(6237):921-5.
Karpichev IV, Durand-Heredia JM, Luo Y, Small GM. 2008. Binding characteristics and regulatory mechanisms of the transcription factors controlling oleate-responsive genes in Saccharomyces cerevisiae. J Biol Chem 283(16):10264-75.
Karpichev IV, Small GM. 1998. Global regulatory functions of Oaf1p and Pip2p (Oaf2p), transcription factors that regulate genes encoding peroxisomal proteins in Saccharomyces cerevisiae. Mol Cell Biol 18(11):6560-70.
Kouzarides T. 2007. Chromatin modifications and their function. Cell 128(4):693-705.
Lahtvee PJ, Kumar R, Hallstrom BM, Nielsen J. 2016. Adaptation to different types of stress converge on mitochondrial metabolism. Mol Biol Cell 27(15):2505-14.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25.
Larochelle M, Drouin S, Robert F, Turcotte B. 2006. Oxidative stress-activated zinc cluster protein Stb5 has dual activator/repressor functions required for pentose phosphate pathway regulation and NADPH production. Mol Cell Biol 26(17):6690-701.
Larsson C, von Stockar U, Marison I, Gustafsson L. 1993. Growth and metabolism of Saccharomyces cerevisiae in chemostat cultures under carbon-, nitrogen-, or carbon- and nitrogen-limiting conditions. J Bacteriol 175(15):4809-16.
Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I and others. 2002. Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298(5594):799-804.
Lee TI, Young RA. 2013. Transcriptional regulation and its misregulation in disease. Cell 152(6):1237-1251.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078-9.
Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG. 2019. Identification of transcription factor binding sites using ATAC-seq. Genome Biol 20(1):45.
Liang H, Gaber RF. 1996. A novel signal transduction pathway in Saccharomyces cerevisiae defined by Snf3-regulated expression of HXT6. Mol Biol Cell 7(12):1953-66.
Lin Charles Y, Lovén J, Rahl Peter B, Paranal Ronald M, Burge Christopher B, Bradner James E, Lee Tong I, Young Richard A. 2012. Transcriptional Amplification in Tumor Cells with Elevated c-Myc. Cell 151(1):56-67.
Ljungdahl PO, Daignan-Fornier B. 2012. Regulation of amino acid, nucleotide, and phosphate metabolism in Saccharomyces cerevisiae. Genetics 190(3):885-929.
Lopes JM, Henry SA. 1991. Interaction of trans and cis regulatory elements in the INO1 promoter of Saccharomyces cerevisiae. Nucleic Acids Res 19(14):3987-94.
Lu H, Li F, Sanchez BJ, Zhu Z, Li G, Domenzain I, Marcisauskas S, Anton PM, Lappa D, Lieven C and others. 2019. A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nat Commun 10(1):3586.
Ma H, Tu LC, Naseri A, Huisman M, Zhang S, Grunwald D, Pederson T. 2016. CRISPR-Cas9 nuclear dynamics and target recognition in living cells. J Cell Biol 214(5):529-37.

Mackiewicz P, Kowalczuk M, Mackiewicz D, Nowicka A, Dudkiewicz M, Laszkiewicz A, Dudek MR, Cebrat S. 2002. How many protein-coding genes are there in the Saccharomyces cerevisiae genome? Yeast 19(7):619-29.
MacPherson S, Larochelle M, Turcotte B. 2006. A fungal family of transcriptional regulators: the zinc cluster proteins. Microbiol Mol Biol Rev 70(3):583-604.
Martins R, Lithgow GJ, Link W. 2016. Long live FOXO: unraveling the role of FOXO proteins in aging and longevity. Aging Cell 15(2):196-207.
Miller SL. 1953. A production of amino acids under possible primitive earth conditions. Science 117(3046):528-9.
Minard KI, McAlister-Henn L. 1999. Dependence of peroxisomal beta-oxidation on cytosolic sources of NADPH. J Biol Chem 274(6):3402-6.
Mortimer RK. 2000. Evolution and variation of the yeast (Saccharomyces) genome. Genome Res 10(4):403-9.
Murillo-Sanchez S, Beaufils D, Gonzalez Manas JM, Pascal R, Ruiz-Mirazo K. 2016. Fatty acids' double role in the prebiotic formation of a hydrophobic dipeptide. Chem Sci 7(5):3406-3413.
Neely KE, Hassan AH, Brown CE, Howe L, Workman JL. 2002. Transcription activator interactions with multiple SWI/SNF subunits. Molecular and cellular biology 22(6):1615-1625.
Ness F, Bourot S, Regnacq M, Spagnoli R, Berges T, Karst F. 2001. SUT1 is a putative Zn[II]2Cys6-transcription factor whose upregulation enhances both sterol uptake and synthesis in aerobically growing Saccharomyces cerevisiae cells. European Journal of Biochemistry 268(6):1585-1595.
Nielsen J. 2003. It is all about metabolic fluxes. J Bacteriol 185(24):7031-5.
Nielsen J. 2019. Yeast Systems Biology: Model Organism and Cell Factory. Biotechnol J 14(9):e1800421.
Nishi K, Park CS, Pepper AE, Eichinger G, Innis MA, Holland MJ. 1995. The GCR1 requirement for yeast glycolytic gene expression is suppressed by dominant mutations in the SGC1 gene, which encodes a novel basic-helix-loop-helix protein. Mol Cell Biol 15(5):2646-53.
Nissen TL, Schulze U, Nielsen J, Villadsen J. 1997. Flux distributions in anaerobic, glucose-limited continuous cultures of Saccharomyces cerevisiae. Microbiology 143 (Pt 1)(1):203-18.
Novick A, Szilard L. 1950. Description of the chemostat. Science 112(2920):715-6.
Oh CS, Toke DA, Mandala S, Martin CE. 1997. ELO2 and ELO3, homologues of the Saccharomyces cerevisiae ELO1 gene, function in fatty acid elongation and are required for sphingolipid formation. J Biol Chem 272(28):17376-84.
Orkin Stuart H, Hochedlinger K. 2011. Chromatin Connections to Pluripotency and Cellular Reprogramming. Cell 145(6):835-850.
Osterlund T, Bordel S, Nielsen J. 2015. Controllability analysis of transcriptional regulatory networks reveals circular control patterns among transcription factors. Integr Biol (Camb) 7(5):560-8.
Ozcan S, Johnston M. 1999. Function and regulation of yeast hexose transporters. Microbiol Mol Biol Rev 63(3):554-69.
Ozonov EA, van Nimwegen E. 2013. Nucleosome Free Regions in Yeast Promoters Result from Competitive Binding of Transcription Factors That Interact with Chromatin Modifiers. PLOS Computational Biology 9(8):e1003181.
Page RA, Okada S, Harwood JL. 1994. Acetyl-CoA carboxylase exerts strong flux control over lipid synthesis in plants. Biochim Biophys Acta 1210(3):369-72.
Perez-Pinera P, Kocak DD, Vockley CM, Adler AF, Kabadi AM, Polstein LR, Thakore PI, Glass KA, Ousterout DG, Leong KW and others. 2013a. RNA-guided gene activation by CRISPR-Cas9-based transcription factors. Nat Methods 10(10):973-6.
Perez-Pinera P, Kocak DD, Vockley CM, Adler AF, Kabadi AM, Polstein LR, Thakore PI, Glass KA, Ousterout DG, Leong KW and others. 2013b. RNA-guided gene activation by CRISPR-Cas9–based transcription factors. Nature Methods 10:973.
Phelps C, Gburcik V, Suslova E, Dudek P, Forafonov F, Bot N, MacLean M, Fagan RJ, Picard D. 2006. Fungi and animals may share a common ancestor to nuclear receptors. Proc Natl Acad Sci U S A 103(18):7077-81.
Polish JA, Kim JH, Johnston M. 2005. How the Rgt1 transcription factor of Saccharomyces cerevisiae is regulated by glucose. Genetics 169(2):583-94.
Postnikoff SDL, Malo ME, Wong B, Harkness TAA. 2012. The Yeast Forkhead Transcription Factors Fkh1 and Fkh2 Regulate Lifespan and Stress Response Together with the Anaphase-Promoting Complex. PLOS Genetics 8(3):e1002583.
Proft M, Gibbons FD, Copeland M, Roth FP, Struhl K. 2005. Genomewide identification of Sko1 target promoters reveals a regulatory network that operates in response to osmotic stress in Saccharomyces cerevisiae. Eukaryot Cell 4(8):1343-52.
Proft M, Struhl K. 2002. Hog1 Kinase Converts the Sko1-Cyc8-Tup1 Repressor Complex into an Activator that Recruits SAGA and SWI/SNF in Response to Osmotic Stress. Molecular Cell 9(6):1307-1317.

Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R and others. 2005. Global analysis of protein phosphorylation in yeast. Nature 438(7068):679-84.
Qi Lei S, Larson Matthew H, Gilbert Luke A, Doudna Jennifer A, Weissman Jonathan S, Arkin Adam P, Lim Wendell A. 2013. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. Cell 152(5):1173-1183.
Querol A, Fleet GH. 2006. Yeasts in food and beverages.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841-2.
Rahner A, Hiesinger M, Schuller HJ. 1999. Deregulation of gluconeogenic structural genes by variants of the transcriptional activator Cat8p of the yeast Saccharomyces cerevisiae. Mol Microbiol 34(1):146-56.
Regenberg B, Grotkjaer T, Winther O, Fausboll A, Akesson M, Bro C, Hansen LK, Brunak S, Nielsen J. 2006. Growth-rate regulated genes have profound impact on interpretation of transcriptome profiling in Saccharomyces cerevisiae. Genome Biol 7(11):R107.
Rhee HS, Pugh BF. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell 147(6):1408-19.
Rhee HS, Pugh BF. 2012. Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature 483(7389):295-301.
Robinson KA, Koepke JI, Kharodawala M, Lopes JM. 2000. A network of yeast basic helix-loop-helix interactions. Nucleic Acids Res 28(22):4460-6.
Robinson KA, Lopes JM. 2000. SURVEY AND SUMMARY: Saccharomyces cerevisiae basic helix-loop-helix proteins regulate diverse biological processes. Nucleic Acids Res 28(7):1499-505.
Rodriguez-Martinez JA, Reinke AW, Bhimsaria D, Keating AE, Ansari AZ. 2017. Combinatorial bZIP dimers display complex DNA-binding specificity landscapes. Elife 6:e19272.
Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. 2010. Origins of specificity in protein-DNA recognition. Annu Rev Biochem 79:233-69.
Rossi MJ, Lai WKM, Pugh BF. 2018a. Genome-wide determinants of sequence-specific DNA binding of general regulatory factors. Genome Res 28(4):497-508.
Rossi MJ, Lai WKM, Pugh BF. 2018b. Simplified ChIP-exo assays. Nat Commun 9(1):2842.
Roth S, Schuller HJ. 2001. Cat8 and Sip4 mediate regulated transcriptional activation of the yeast malate dehydrogenase gene MDH2 by three carbon source-responsive promoter elements. Yeast 18(2):151-62.
Salazar AN, Gorter de Vries AR, van den Broek M, Wijsman M, de la Torre Cortes P, Brickwedde A, Brouwers N, Daran JG, Abeel T. 2017. Nanopore sequencing enables near-complete de novo assembly of Saccharomyces cerevisiae reference strain CEN.PK113-7D. FEMS Yeast Res 17(7).
Sanchez B, Li,F., Lu,H., Kerkhoven,E. and Nielsen,J. 2016. Yeast-GEM: yeast 7.6.0. .
Sanchez BJ, Zhang C, Nilsson A, Lahtvee PJ, Kerkhoven EJ, Nielsen J. 2017. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. Mol Syst Biol 13(8):935.
Sanda T, Lawton LN, Barrasa MI, Fan ZP, Kohlhammer H, Gutierrez A, Ma W, Tatarek J, Ahn Y, Kelliher MA and others. 2012. Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. Cancer cell 22(2):209-221.
Santiago TC, Mamoun CB. 2003. Genome expression analysis in yeast reveals novel transcriptional regulation by inositol and choline and new regulatory functions for Opi1p, Ino2p, and Ino4p. J Biol Chem 278(40):38723-30.
Scherer S, Davis RW. 1979. Replacement of chromosome segments with altered DNA sequences constructed in vitro. Proc Natl Acad Sci U S A 76(10):4951-5.
Shahzad K, Loor JJ. 2012. Application of Top-Down and Bottom-up Systems Approaches in Ruminant Physiology and Metabolism. Curr Genomics 13(5):379-94.
Solomon MJ, Varshavsky A. 1985. Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. Proceedings of the National Academy of Sciences 82(19):6470-6474.
Stukey JE, Mcdonough VM, Martin CE. 1990. The Ole1 Gene of Saccharomyces-Cerevisiae Encodes the Delta-9 Fatty-Acid Desaturase and Can Be Functionally Replaced by the Rat Stearoyl-Coa Desaturase Gene. Journal of Biological Chemistry 265(33):20144-20149.
Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES and others. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102(43):15545-50.
Swift J, Coruzzi GM. 2017. A matter of time - How transient transcription factor interactions create dynamic gene regulatory networks. Biochim Biophys Acta Gene Regul Mech 1860(1):75-83.
Teixeira MC, Monteiro PT, Palma M, Costa C, Godinho CP, Pais P, Cavalheiro M, Antunes M, Lemos A, Pedreira T and others. 2017. YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. Nucleic Acids Research 46(D1):D348-D353.

Toke DA, Martin CE. 1996. Isolation and characterization of a gene affecting fatty acid elongation in Saccharomyces cerevisiae. J Biol Chem 271(31):18413-22.
Trzcinska-Danielewicz J, Ishikawa T, Miciałkiewicz A, Fronk J. 2008. Yeast transcription factor Oaf1 forms homodimer and induces some oleate-responsive genes in absence of Pip2. Biochemical and Biophysical Research Communications 374(4):763-766.
Turcotte B, Liang XB, Robert F, Soontorngun N. 2010. Transcriptional regulation of nonfermentable carbon utilization in budding yeast. FEMS Yeast Res 10(1):2-13.
Uemura H, Fraenkel DG. 1990. gcr2, a new mutation affecting glycolytic gene expression in Saccharomyces cerevisiae. Mol Cell Biol 10(12):6389-96.
Uemura H, Jigami Y. 1992. Role of GCR2 in transcriptional activation of yeast glycolytic genes. Mol Cell Biol 12(9):3834-42.
Varemo L, Nielsen J, Nookaew I. 2013. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Research 41(8):4378-4391.
Verduyn C, Postma E, Scheffers WA, van Dijken JP. 1990. Physiology of Saccharomyces cerevisiae in anaerobic glucose-limited chemostat cultures. J Gen Microbiol 136(3):395-403.
Westholm JO, Nordberg N, Muren E, Ameur A, Komorowski J, Ronne H. 2008. Combinatorial control of gene expression by the three yeast repressors Mig1, Mig2 and Mig3. BMC Genomics 9:601.
Vik A, Rine J. 2001. Upc2p and Ecm22p, dual regulators of sterol biosynthesis in Saccharomyces cerevisiae. Molecular and cellular biology 21(19):6395-6405.
Wing MK. 2010. bamUtil [cited Dec 2015].
Winston F, Carlson M. 1992. Yeast SNF/SWI transcriptional activators and the SPT/SIN chromatin connection. Trends Genet 8(11):387-91.
Workman CT, Mak HC, McCuine S, Tagne JB, Agarwal M, Ozier O, Begley TJ, Samson LD, Ideker T. 2006. A systems approach to mapping DNA damage response pathways. Science 312(5776):1054-9.
Workman JL, Kingston RE. 1998. Alteration of nucleosome structure as a mechanism of transcriptional regulation. Annu Rev Biochem 67(1):545-79.
Wu FX. 2008. Genetic weighted k-means algorithm for clustering large-scale gene expression data. BMC Bioinformatics 9 Suppl 6(Suppl 6):S12.
Yan C, Chen H, Bai L. 2018. Systematic Study of Nucleosome-Displacing Factors in Budding Yeast. Mol Cell 71(2):294-305 e4.
Zamanighomi M, Lin Z, Wang Y, Jiang R, Wong WH. 2017. Predicting transcription factor binding motifs from DNA-binding domains, chromatin accessibility and gene expression data. Nucleic acids research 45(10):5666-5677.
Zampar GG, Kummel A, Ewald J, Jol S, Niebel B, Picotti P, Aebersold R, Sauer U, Zamboni N, Heinemann M. 2013. Temporal system-level organization of the switch from glycolytic to gluconeogenic operation in yeast. Mol Syst Biol 9:651.
Zaret KS, Carroll JS. 2011. Pioneer transcription factors: establishing competence for gene expression. Genes Dev 25(21):2227-41.
Zhang Z, Dietrich FS. 2005. Mapping of transcription start sites in Saccharomyces cerevisiae using 5' SAGE. Nucleic Acids Res 33(9):2838-51.
Zhou KM, Bai YL, Kohlhaw GB. 1990. Yeast regulatory protein LEU3: a structure-function analysis. Nucleic Acids Res 18(2):291-8.