A combined systems biology and genomics approach to the study of metabolism in *Kluyveromyces marxianus*

by

Du Toit Willem Petrus Schabort

Submitted in fulfilment of the requirements in respect of the

Doctoral degree qualification in Biochemistry

in the Department of Microbial, Biochemical and Food Biotechnology in the Faculty of Natural and Agricultural Sciences at the University of the Free State.

17 October 2016

Promoter: **J.C. du Preez**

Co-promoter: **S.G. Kilian**

Acknowledgements

I would like to thank Prof J.C. du Preez for his guidance as a mentor during the project. Through his hard work and dedication to perfection he is an inspiration to those around him.

To Prof S.G. Killian, thank you for fruitful conversations, and mentorship through tough times.

To Sarel Marais, Gabré Kemp and Laurinda Steyn, thank you for the analyses performed in the analytical laboratories and help with bioreactors.

To Honnours student Precious Letebele that has contributed with RNA-seq experiments, thank you and well done!

Thank you to Antonie Meyer with help on genome extraction.

To Prof Martie Smit and colleagues, thank you for the patience.

To my friends, thank you for being my friends.

To my loving family, Pa Johannes, Ma Lulu and Boet Charl, thank you for believing in me and supporting me through the toughest of times.

To my Lord and Saviour. I am grateful for this privilege, of which there are so many in my life.

Holy is He.

Contents

Summary

The yeast *Kluyveromyces marxianus* has become an important micro-organism for industrial applications, as have other non-conventional yeasts. It has the advantages over *Saccharomyces cerevisiae* (baker's yeast) in that it is more thermotolerant, has a much higher growth rate and can utilise a wider range of sugars, including the pentose D-xylose, which is found abundantly in lignocellulosic biomass. Although considerable advances have been made in engineering *S. cerevisiae* strains to ferment pentose sugars, their performance in this respect still does not approach that of glucose fermentation. *S. cerevisiae* is the model Crabtree positive yeast, meaning that it naturally ferments glucose even if oxygen is present at a high level. Crabtree negative yeasts, such as *K. marxianus,* have to be forced into a fermentative metabolism by imposing oxygen-limited conditions, which is impractical on industrial scale. Thus, a tremendous amount of knowledge needs to be gained regarding the regulation of metabolism in this non-conventional yeast before success could be expected in the re-programming of *K. marxianus* strains into xylose fermenting, Crabtree positive strains. The challenge of bringing a non-model species such as *K. marxianus* to the point of identifying key regulators affecting central metabolic pathways seems formidable. The aims of this work was to firstly harness the new technology of next-generation sequencing (NGS) to create a first draft genome for *K. marxianus* strain UFS-Y2791 and to generate high-quality RNA-seq differential transcriptome datasets, simultaneously capturing a tremendous amount of information. Efficient analytical methods and software implementations were also developed to explore these large datasets in an efficient manner, revealing new insights into the response of this species to glucose and xylose as carbon sources.

RNA-seq data revealed a striking resemblance with the pattern of glucose derepression in the xylose medium, with up-regulation of genes for alternative carbon source utilisation, especially in the peroxisomes. Subsequently, two independent approaches were taken to identify differentially active transcription factors regulating the response. The first was the enumerative method of heptamer frequency comparisons, revealing the most likely regulators of differentially expressed genes. Secondly, a likelihood statistical approach was designed that employs multiple sources of evidence, which resulted in the construction of the first genome-wide gene regulatory network for *K. marxianus*. The method bridges the gap between the new NGS-based methods, which can rapidly generate data on any non-model species, and the wealth of experimental data that exist for a model species such as *S. cerevisiae*. Gene set enrichment statistics of the transcription factor target sets showed a general pattern that the activities of differentially active transcription factors were regulated primarily by post-

translational modifications instead of gene regulation. The use of RNA-seq was further expanded to the elucidation of the kinases that regulate transcription factors. The chromosomal context of differential gene expression was also investigated. Clusters of genes were identified, similar to the sub-telomeric regions previously identified in *S. cerevisiae*, but not close to telomeres. These regions contain industrially important enzymes and the potential binding sites for differentially active transcription factors.

Finally, the possible roles of cofactor balances were investigated. Flux balance analysis was demonstrated here in rationalising the genetic response observed in RNA-seq transcriptomics and to understand the complex interplay between ATP, NADPH and NADH, the cofactor specificity of the oxidative pentose phosphate pathway, as well as the role of fructose-1,6-bisphosphatase. New roles are proposed for the latter enzyme, which differs from the currently accepted norm. A strategy for the metabolic engineering of a future xylose fermenting *K. marxianus* strain is also presented.

The integrated analysis presented here expands our knowledge base of this yeast species, which is set to become increasingly important in a future bio-economy.

Keywords

Kluyveromyces marxianus

Xylose

Transcription factors

Biochemical network analysis

Gene set enrichment

Metabolic Regulation Analysis

Bayesian network

Metabolism

Flux Balance Analysis

Fructose-1,6-bisphosphatase

Biofuel

Opsomming

Die gis *Kluyveromyces marxianus* het ’n belangrike mikroörganisme geword vir industriële toepassings, soos ook ander nie-konvensionele giste. Dit het die voordele bo *Saccharomyces cerevisiae* (bakkersgis) dat dit meer termotolerant is, ’n baie hoër groeitempo handhaaf en ’n groter verskeidenheid suikers benut, insluitend die pentose D-xilose, wat volop in lignosellulose-biomassa voorkom. Hoewel beduidende vordering al gemaak is in die genetiese manipulering van *S. cerevisiae*-stamme om pentoses te fermenteer, is hul prestasie in hierdie opsig steeds nie vergelykbaar met dié van glukose-fermentasie nie. *S. cerevisiae* is die model Crabtree-positiewe gis, wat beteken dat dit glukose natuurlik fermenteer, selfs al is ’n hoë suurstofvlak teenwoordig. Crabtree-negatiewe giste soos *K. marxianus* moet tot ’n fermentatiewe metabolisme gedwing word deur suurstof-beperkende toestande in te stel, wat op ’n industriële skaal onprakties is. ’n Geweldige hoeveelheid kennis oor die regulering van metabolisme in hierdie nie-konvensionele gis moet dus opgebou word voordat sukses verwag kan word met die herprogrammering van *K. marxianus*-stamme na xilose-fermenterende, Crabtree-positiewe stamme. Die uitdaging om ’n nie-modelspesie soos *K. marxianus* tot by die punt te bring waar sleutelreguleerders wat die sentrale metabolisme beïnvloed, geïdentifiseer kan word, blyk formidabel te wees. Die doelstellings van hierdie werk was, eerstens, om die nuwe tegnologie van volgende-generasie volgordebepaling (VGV) in te span om die eerste voorlopige genoom vir die *K. marxianus*-stam UFS-Y2791 te bepaal en hoë gehalte RNA-seq differensiële transkriptoomdatastelle te genereer, en om terselfdertyd ’n geweldige hoeveelheid data vas te vang. Doeltreffende analitiese metodes en programmatuur-implementerings is ook ontwerp om hierdie groot datastelle op ’n doeltreffende wyse te verken, wat nuwe insigte aan die lig gebring het ten opsigte van die respons van hierdie spesie tot glukose en xilose as koolstofbronne.

In die xilose-medium het die RNA-seq data ’n sterk ooreenkoms met die patroon van glukose-derepressie getoon, met die op-regulering van gene vir die benutting van alternatiewe koolstofbronne, veral in die peroksisome. Gevolglik is twee onafhanklike benaderings gevolg om die differensieël-aktiewe transkripsiefaktore wat die respons reguleer, te identifiseer. Die eerste was die numeriese metode van heptameerfrekwensie vergelykings, wat die mees waarskynlike reguleerders van differensieël-uitgedrukte gene onthul het. Tweedens, is ’n waarskynlikheids-statistiese benadering ontwerp wat veelvuldige bronne van bewyse inspan, wat gelei het tot die konstruksie van die eerste genoomwye geen-regulatoriese netwerk vir *K. marxianus*. Die metode oorbrug die gaping tussen die magdom eksperimentele data vir ’n modelspesie soos *S. cerevisiae* en die nuwe VGV-gebaseerde metodes, wat vinnig data van enige nie-model spesie kan genereer. Geen-stel verrykingstatistiek vir

die transkripsiefaktor teikenstelle het ’n algemene patroon aangedui dat die aktiwiteite van differensieël-aktiewe transkripsiefaktore primêr deur post-vertaling modifikasies gereguleer is, eerder as deur geen-regulering. Die gebruik van RNA-seq is verder uitgebrei na die toeligting van die kinases wat die transkripsiefaktore reguleer. Die chromosomale konteks van differensiële geen-uitdrukking is ook ondersoek. Groepe gene is geïdentifiseer, soortgelyk aan die sub-telomeriese streke wat voorheen in *S. cerevisiae* geïdentifiseer is, maar wat nie naby aan die telomere geleë was nie. Hierdie streke bevat industrieel-belangrike ensieme en die potensiële bindingsetels vir differensieel-aktiewe transkripsiefaktore.

Laastens, is die moontlike rolle van kofaktorbalanse ondersoek. Fluksbalans-analise is hier as ’n kragtige hulpmiddel gedemonstreer vir die rasionalisering van die genetiese respons wat met RNA-seq transkriptomika waargeneem word, en om die komplekse interaksie tussen ATP, NADPH en NADH, die kofaktor spesifisiteit van die oksidatiewe pentosefosfaat-weg, sowel as die rol van fruktose-1,6-bisfosfatase, te verstaan. Nuwe rolle word vir die laasgenoemde ensiem voorgestel, wat verskil van die tans aanvaarde norm. ’n Strategie vir die metaboliese manipulering van ’n toekomstige xilose-fermenterende *K. marxianus*-stam word ook aangebied.

Die geïntegreerde analise wat hier aangebied word, brei ons kennisbasis van hierdie gisspesie uit wat in ’n toekomstige bio-ekonomie toenemend belangrik gaan word.

Sleutelwoorde

Kluyveromyces marxianus

Xilose

Transkripsiefaktore

Biochemiese netwerk analise

Geen-stel verryking

Metaboliese Reguleringsanalise

Bayes netwerk

Metabolisme

Fluks-balans Analise

Fruktose-1,6-bisfosfatase

Bio-brandstof

Declaration

(i) "I, DuToit Schabort, declare that the Doctoral Degree research thesis or interrelated, publishable manuscripts / published articles, or coursework Doctoral Degree mini-thesis that I herewith submit for the Doctoral Degree qualification in **Biochemistry** at the University of the Free State is my independent work, and that I have not previously submitted it for a qualification at another institution of higher education."

(ii) "I, DuToit Schabort, hereby declare that I am aware that the copyright is vested in
the University of the Free State."

(iii) "I, DuToit Schabort, hereby declare that all royalties as regards intellectual property that was developed during the course of and/or in connection with the study at the University of the Free State, will accrue to the University."

Chapter 1

Introduction and Literature Review

Introduction

Global warming is by now established as a major threat to the long-term survival of the human race and could lead to the extinction of species at a global scale if solutions are not found to mitigate this problem. A major contributing factor is the burning of fossil fuels, leading to increased carbon dioxide concentrations in the atmosphere. Although technologies such as wind, solar and nuclear energy will become very important in this effort in the move to a greener future, these do not readily replace the liquid fuels powering ships, aeroplanes and motor vehicles. Liquid fuels such as petroleum derived from the refining of crude oil or the gasification of coal are energy dense and thus very convenient to use. However, they are non-renewable since they are based on a finite supply of fossil fuels. On the other hand, renewable biofuels in the form of ethyl alcohol (ethanol) or butanol might replace these, and can be produced at a large scale as the primary fermentation products of yeasts or bacteria. The best established biofuel is ethanol produced by baker's yeast, *Saccharomyces cerevisiae* [Hahn-Hägerdal et al. 2007]. Industrial scale biofuel production has been done on a large scale in the USA and in Brazil [Rosillo-Calle 2012]. In the USA, maize (corn) starch is converted to the monomer glucose, which is then fermented by *S. cerevisiae* which naturally produces ethanol, even in the presence of oxygen. In Brazil, the large-scale production of sugarcane yields the disaccharide sucrose, which is the feedstock used for bioethanol production. A significant problem is that the production cost of the sugar by either of these two methods is high. The cost of the sugar has been estimated at up to 70% of the final cost of bioethanol [Marini et al. 1997, Pfromm et al. 2010]. It is, therefore, imperative that strains should be optimised for maximal ethanol yields from the substrate. A more fundamental set of problems are those involving competition between biofuel production and food production for arable land, socio-economical injustices, and inequality [Rosillo-Calle 2012].

There now exists a huge opportunity for exploiting the abundant lignocellulosic biomass for the production of biofuels, recombinant proteins and biomaterials, and this resource cannot be ignored [Rosillo-Calle 2012]. Lignocellulosic biomass such as agricultural wastes and paper pulp currently goes to waste at a massive scale, which could potentially be used as a cheap carbon source as a second-

generation feedstock instead of glucose or sucrose from maize or sugarcane. This is dependent on effective technologies for depolymerisation of lignocellulose [Hahn-Hägerdal et al. 2007] and the development of microbial strains that could utilise and ferment the sugars that constitute lignocellulose. A major component in lignocellulose is the five-carbon sugar xylose as well as arabinose [Hahn-Hägerdal et al. 2007]. Natural strains of *S. cerevisiae*, however, do not utilise pentoses. Extensive efforts in metabolic engineering of *S. cerevisiae* for xylose fermentation over nearly two decades have resulted in a few strains of *S. cerevisiae* that can ferment xylose, but sensitivity of the strains to toxins in the lignocellulosic hydrolysate, among others, are problematic [Hahn-Hägerdal et al. 2007]. An alternative strategy would be to employ a non-traditional yeast which naturally utilises five-carbon sugars, such as *Kluyveromyces marxianus*, for ethanol production [Nonklang et al. 2008]. The main challenge is that these yeasts do not naturally produce high concentrations of ethanol. *S. cerevisiae* and other Saccharomycetes produce ethanol even in the presence of oxygen, an effect known as the Crabtree effect [Crabtree 1929, De Deken 1966, Postma et al. 1989]. Crabtree negative yeasts, on the other hand, have to be forced into fermentation by restricting oxygenation. The latter process is not only less productive as compared to the use of the Crabtree positive yeasts, but the requirement of controlled oxygen limitation [Kuloyo et al. 2014] would be expensive from an industrial perspective, which would also requires robustness of the process to control parameters.

The need for efficient pentose fermenting strains deserves consideration of genetic re-programming of non-conventional yeasts into Crabtree positive, xylose fermenting strains. A top candidate for this endeavour is *Kluyveromyces marxianus*. Apart from its ability to utilise xylose, arabinose and many six-carbon sugars, it is more thermotolerant than *S. cerevisiae* [Fonseca et al. 2008, Lane et al. 2010] and grows well on biomass hydrolysate [Akanni et al. 2015]. Moreover, it has an extremely high growth rate; in fact, the highest among all known eukaryotes [Groeneveld et al. 2009]. In order to guide such an ambitious endeavour, an in depth understanding of the genetic and metabolic regulation in *K. marxianus* at a genome-wide scale is essential. The latter is the main aim of this study. A powerful method of capturing knowledge of gene regulation is to construct models. These may be interrogated for elucidating the differentially active regulators, or be used in a predictive sense to guide genetic manipulations.

Before any models can be constructed, a blueprint is required. At the time this investigation commenced, no complete genome or reasonably complete draft genome existed for *K. marxianus*. The only one was a draft genome of 20% completion, which is not sufficient for the purpose, which was sequenced by the Genelovures Consortium [Llorente et al. 2000, Souciet 2011]. Next-generation

sequencing (NGS), however, is now an affordable option for generating a draft genome. Chapter 2 describes the sequencing and assembly of the first draft genome for strain UFS-Y2791. This strain was isolated from the juice of the arid zone succulent *Agave americana* by Carolina Pohl-Albertyn of the University of the Free State, and shows potential as a bioethanol producer [Kuloyo et al. 2014]. A draft genome of *Kluyveromyces marxianus* UFS-Y2791 is presented in this chapter. This formed the blueprint for subsequent investigations regarding bioinformatics, computational biology and transcriptome analyses presented in this thesis. Some innovations involving optimisation of assembly parameters, inclusion of genome annotations into this optimisation procedure, and a software programme that facilitates the process are also presented in this Chapter 2.

Another method that employs NGS is RNA-seq. This is a powerful method of obtaining RNA levels of all transcribed genes under a particular condition [Trapnell et al. 2010]. It has major advantages over microarray analysis in that it has an improved sensitivity, an increased dynamic range and is less influenced by various forms of bias. Further, it does not require the construction of a specialised array, making it ideal for the study of non-model species. In Chapter 3, RNA-seq is employed to explore the differential genetic response of *K. marxianus* to glucose and xylose as the respective carbon sources. The findings are compared to another transcriptomics data set published very recently [Lertwattanassakul et al. 2015].

Chapter 3 describes the in depth exploration of the RNA-seq datasets from a several perspectives, making use of Gene Ontology, metabolic pathway maps at various scales, biochemical network maps and gene set enrichment statistics. Several differentially regulated pathways and biochemical processes were identified. It was also investigated whether the yeast regulated its metabolic fluxes in central metabolism mostly at the enzyme kinetic level in response to changes in the concentrations of metabolites, or whether gene regulation played a significant role. The theoretical formalism of Metabolic Regulation Analysis (MRA) [Ter Kuile and Westerhoff 2001] was used for this purpose in this investigation. To perform each of these analyses, a comprehensive suite of software programmes was designed and coded in this study, called *Reactomica,* using the Wolfram Mathematica language. All subsequent modelling and data analyses performed in this work were also implemented in this manner. In each of the chapters, bioinformatics and computational biology algorithms were designed and written as part of *Reactomica*, and used in the analysis. In most cases, detailed explanations of the algorithms are provided in a condensed format in the Materials and Methods sections and addenda.

Two independent approaches were taken to elucidate the most likely candidate transcription factors (TFs) regulating genome-wide gene regulation. The first was the enumerative method of heptamer frequency comparisons used in Chapter 4. This method revealed the top candidate TFs. The second approach involved the construction of complete gene regulatory networks at the genome-wide scale, subsequently employing enrichment statistics to reveal differentially active TFs. A new method for construction of gene regulatory networks, based on likelihoods, is presented in Chapter 5. This method combines multiple sources of evidence, ranging from the evidence of DNA binding sites in the species of interest, to experimental evidence of interactions in the model species, which in this case was *S. cerevisiae*. Chapter 6 extended the idea and improved the process. These are the first gene regulatory networks for *K. marxianus*. Chapter 5 focusses on the methods that were required for network construction, including improvement of the draft genome and multiple genome alignment. Some algorithmic details are provided in Addenda 1-4. The network was subsequently used to elucidate kinases that might be the master regulators, affecting the activity of TFs, in a new approach presented in Chapter 7.

It is known that in *S. cerevisiae*, X and Y elements occur which are regions close to telomeres that regulate a series of adjacent genes by chromatin silencing [Smith et al. 2011]. The perspective of complete chromosomes is reported in Chapter 8, where clusters of differentially expressed genes were sought.

Finally, a computer simulation study of metabolism was made, using the theoretical framework of Flux Balance Analysis (FBA) Schilling et al. [1999] (Chapter 9). This study is not only relevant to the species, but also in a more general sense. For metabolic engineering, cofactor balances cannot be ignored and neither can they be interpreted separately. The complex relationship between the balances of NADH/NAD, ATP/ADP and NADPH/NADP, the fluxes in the oxidative pentose phosphate pathway (PPP), the pyruvate dehydrogenase bypass (PDB), ethanol production, electron transport chain (ETC) and glycerol production was investigated. In particular, fructose-1,6-bisphosphatase (FBP) is discussed as an enzyme which might have a special ('moonlighting') role, which might not be limited to its gluconeogenic role, as is usually assumed in the case of yeast. Finally, a possible engineering strategy is proposed for developing a future xylose fermenting strain of *K. marxianus*. The framework of FBA was built into *Reactomica* for these analyses.

RNA-seq and other NGS variants have recently been established in terms of the experimental protocols, and excellent software programmes have become available such as Bowtie [Langmead et

al. 2009, Langmead et al. 2012], TopHat [Trapnell et al. 2013, Kim et al. 2013] and CuffLinks [Trapnell et al. 2013]. However, it is important to note that these provide only core data processing of read data, up to the calculation of differential expression or elucidation of splice variants. Downstream analyses are much less standard and is an active field in Bioinformatics and Computational Biology. A major goal in this study was to develop convenient software programmes that could be used by the novice that require no programming skills. The "game changer" would be a programme or set of integrated programmes that takes as input, a genome, phenotypic data such as RNA-seq transcriptomics, and additional sources of evidence for biochemical interactions in model species and automatically create models of metabolism, gene regulation and signal transduction, and analyse the phenotypic data using these models as blueprint. Since the digital world and social media has brought with it a taste for visual communication, the outputs from these programmes need to be visually attractive. In particular, such a programme would need to allow the automated construction of biochemical networks, perform a variety of gene set enrichment statistics, render biochemical pathway maps at various scales and perform metabolic flux and other simulations. It would also need to construct genome-wide gene regulatory networks for a non-model species and make best use of the large databases of biochemical interactions in model species such as the *Saccharomyces* Genome Database (SGD). Some of these functionalities cannot be performed without core data processing ability such as DNA binding motif scans and heptamer frequency comparisons. The complex integrative process of constructing gene regulatory networks based on multiple sources of evidence, including multiple genome alignment, would also require a robust framework for genome-based data integration with a built-in genome locus coordinate system. This makes genomic track based visualisation not only convenient from a user perspective, but necessary for error checking.

Software with this number of diverse functionalities does not exist today, as far as the author is aware. Several other software programmes do exist that perform individual functionalities, including Galaxy [Afgan et al. 2015], PathwayTools [Karp et al. 2009], Raven [Agren et al. 2013], Cytoscape [Shannon et al. 2003], FiatFlux [Zamboni et al. 2005] and the genome browser at UCSC [Kent et al. 2002]. However, a large amount of code would still need to be developed for seamless integration, most likely in the Linux environment. A software programme that consists of multiple languages has the drawback of being difficult to maintain, and might become dependent on individual code repositories. Servers such as the genome browser at UCSC and Galaxy are run on Linux servers and developed by teams of skilled bioinformaticians. In this thesis, a unique suite of programmes was developed that performed the above functionalities in a single, platform independent programming language, while making use of outputs from standard, established programmes for basic NGS data processing. The *Wolfram*

Language, best known as *Mathematica*, has grown to be a highly sophisticated engine for scientific computations. This commercial language was chosen since it provides endless possibilities with powerful algorithms in mathematics, statistics, network analysis, visualisation, machine learning and other fields. Since it is, compared to other languages, relatively natural to programme, it might change the playing field in the sciences. There are, however, no well-known primary or third party bioinformatics packages in the *Wolfram Language* to date. *Reactomica*, which was developed in this study, is one of the first of these.

It should be noted that even though frameworks such as FBA [Schilling et al. 1999], MRA [Ter Kuile and Westerhoff 2001] and methods relating to the use of gene set enrichment [Patil and Nielsen 2005, Oliveira et al. 2008] have been developed by others, scope exists for exploration of their utility for certain modern data types such as RNA-seq, for refining their implementations and ultimately for their integration. In addition, a comment has to be made regarding the practical and economic feasibility of a comprehensive systems biology study applied to a non-model organism such as *K. marxianus*. Although ultimately a variety of high-throughput data types such as genome sequencing, RNA-seq, proteomics, phospho-proteomics, metabolomics and chromatin immunoprecipitation (ChIP) would be ideal for genome-scale systems biology studies, it is neither economically nor practically feasible to perform in a short time frame all of these methods. Yet, some of these are particularly rich in the relevant information for the purpose of studying genetic and metabolic regulation, and also more practically feasible and cost effective as opposed to others. In this study, genome sequencing and RNA-seq using NGS is demonstrated and promoted as a highly complementary and rich combination, and ideally suited for greatly expanding our knowledge of regulation in a species such as *K. marxianus*. To this end, a variety of analyses were performed on these data sets, resulting in many new observations and several new hypotheses.

Literature Review

Since a relevant specific literature review is provided in the introduction of each chapter, only some key aspects are discussed below.

Potential of *Kluyveromyces marxianus* for metabolic engineering

The yeast *K. marxianus* is homothallic, belongs to the hemiasocycetes and is related to *S. cerevisiae* and more closely to the dairy yeast *K. lactis* [Lachance 1998; Llorente et al. 2000]. Both species can

utilise lactose as sole carbon source [Lane et al. 2010]. *K. marxianus* has been adopted in industry due to its ability to utilise a variety of sugars, the ability of some strains to grow in temperatures as high as 53^{o}C, and because it has a high capacity for producing recombinant proteins [Fonseca et al. 2008, Lane et al. 2010]. Before molecular biology tools were employed for a systematic classification, yeasts were originally classified using physiological and morphological traits. Using molecular classification of the D1/D2 region of the 28S rRNA gene, and later multiple genes [Kurtzman 2003], the original *Kluyveromyces* genus was subsequently separated such that some of the original members such as *K. thermotolerans* and *K. polysporus* were excluded from the *Kluyveromyces* genus [Kurtzman et al. 1998, Kurtzman 2003, Lachance 2007, Lane et al. 2010]. The most closely related species to *K. marxianus*, in order of relatedness, are *K. dobzhanskii*, *K. lactis*, *K. wickerhamii*, *K. nonfermentans* and *K. aestauri*. Other closely related genera (in order of relatedness) are *Lachancea* (containing *L. thermotolerans*), *Torulaspora*, *Zygotorulaspora*, *Zygosaccharomyces*, *Vanderwaltozyma* (containing *V. polyspora*), *Tetrapisispora*, *Nakaseomyces*, *Saccharomyces*, *Naumovia* and *Kazachstania*. *K. marxianus* is now the new type species of the genus, whereas the previous type species, *K. polysporus,* was renamed to *Vanderwaltozyma polysporus*. The genus *Kluyveromyces* cannot be uniquely identified by a set of phenotypic traits shared by all species in the genus and not by other genera [Lachance 2007]. The *Kluyveromyces* genus is thus closely related to the *Saccharomyces* genus and both fall under the *Saccharomyces* complex. *Kluyveromyces* is separated from the *Saccharomyces sensu stricto* species by the fact that the latter underwent a whole-genome duplication event 100 million years ago [Wolfe and Shields 1997]. Additional copies of the genes allowed rapid evolution, resulting in Crabtree positive species in genera such as *Saccharomyces* and *Kazachstania*.

The ploidy of *Kluyveromyces marxianus* has been shown to be either haploid or diploid, depending on the species [Pecota et al. 2007, Hong et al. 2007, Nonklang et al. 2009]. This is an important consideration when attempting genetic engineering [Lane et al. 2010].

K. marxianus is a respiro-fermentative yeast, since it can perform both oxidative metabolism and fermentation. It is currently not established whether it is best characterised as a Crabtree positive or negative yeast, since some strains produce ethanol in the presence of oxygen [Lane 2010, Kuloyo et al. 2014, Bajpai and Margaritis 1982, Rocha et al. 2011]. The species is already important from an industrial perspective, since it is utilised to produce pectinases, β-galactosidases and inulinase (reviewed by Fonseca et al. 2008). For these applications, its proneness to Crabtree negative behaviour is useful. However, for the purpose of biofuel production, the Crabtree negative behaviour is detrimental; hence it is important to understand the mechanistic basis of fermentative behaviour

in terms of the metabolic network, the flux constraints, the cofactor balances and the signalling networks that govern metabolism.

A very convenient new technology for studying gene regulation at the genome scale is RNA-seq. The work by Lertwattanassakul et al. [2015] was the first genome-scale transcriptomics study in *K. marxianus*. These authors studied the differential response to glucose or xylose as the carbon source in a rich medium using RNA-seq. Dramatic differential expression was observed in peroxisomal metabolism, including β-oxidation. The suggestion was made that this was due to the presence of a trace amount of lipids in the complex medium as an additional carbon source that was simultaneously utilised with xylose. Most likely, however, the mechanistic basis for this response was the alleviation of glucose repression, which is a very important mechanism in *S. cerevisiae*. Up-regulation of other genes involved with utilisation of alternative carbon sources in the absence of glucose was also reminiscent of glucose derepression.

The first hypothesis that was addressed in this study was that in *K. marxianus* the same regulators governing glucose repression as in *S. cerevisiae* would be found. The regulators that govern the transcriptional response during glucose repression are briefly reviewed in the following sections, with a focus on the peroxisomal genes.

Response to fermentable and non-fermentable carbon sources in *S. cerevisiae*

The TFs involved in the utilisation of non-fermentable carbon sources are Hap2-5, Rtg1-3, Cat8, Sip4, Mig1, Adr1, Oaf1, Oaf3 and Pip2 [Schuller 2003, Wan et al. 2013, Broach 2012]. Broach [2012] wrote and excellent in-depth review recently on the control of growth and development in *S. cerevisiae*. An extensive review of these pathways fall outside the scope of this work. Whereas in mammals, cells perceive their environment (the bloodstream or interstitial fluid) by the influence of hormones, in microorganisms the nutrient molecules serve both as nutrients and as signalling molecules, in a sense making signalling more complicated as opposed to that in mammals [Broach 2012]. The most important of these molecules is glucose. In fact, one quarter of the 5 770 genes in *S. cerevisiae* are affected by glucose repression [Young et al. 2003]. It is likely that the genes reported to be differentially regulated in *K. marxianus* [Lertwattanasakul et al. 2015] are subject to glucose repression, as glucose was absent from the xylose medium. Five interlinked kinase signalling pathways are present in *S. cerevisiae*, which link the concentration of glucose to metabolism, growth and cell morphology via the TFs that regulate gene expression. These signalling pathways are (*a*) the

Ras/protein kinase A (PKA) pathway, (*b*) Sch9, (*c*) Snf1, (*d*) the HAP2/3/4/5 complex and (*e*) Rgt. The most influential is the Ras/PKA pathway, which may orchestrate as much a 90% of the differential expression in *S. cerevisiae* cells grown on glycerol after glucose addition [Zaman et al. 2009]. Responses such as pseudohyphal growth are dependent on the Ras/PKA pathway. The Snf1 pathway, on the other hand, is very important specifically for the regulation of TFs involved with alternative carbon source utilisation and peroxisomal metabolism. The HAP2/3/4/5 complex also plays a role in the TCA cycle which is, together with Snf1, required for aerobic catabolism of non-fermentable carbon sources [Broach 2012].

While both *S. cerevisiae* and *K. marxianus* display a yeast-like morphology in a rich medium when cell growth and replication is rapid, a fraction of the cells form pseudohyphae under a nutrient limitation or other stresses [Groeneveld et al. 2009]. In *S. cerevisiae*, phenotype switching is regulated by a number of TFs regulating several hundred target genes. These are the transcriptional activators Phd1, Mss11, Ash1, Flo8, Msn1, Haa1, Ste12, Tec1 and Mga1 and the transcriptional repressors Nrg1, Nrg2, Sfl1 and Sok2 [Broach 2012]. Phd1 and Mga1 are the main regulators of phenotype switching. Two different phenotypes (yeast and pseudohyphae) that occur simultaneously is a trademark of stochastic noise in gene expression; the levels of activity of these TFs would not be equal in all cells and a representative sample from the population shows the average gene expression levels among cells [Eldar and Elowitz 2010]. The coordinated action of these TFs activate the filamentous programme in which genes like the flocculation gene FLO11 is required [Broach 2012]. Depletion of the carbon source signals via both the Snf1 kinase and the Ras/PKA pathways, activating pseudohyphal growth and involving Yak1 [Robertson and Fink 1998; Malcher et al. 2011].

Snf1 has a homolog in mammals, namely the AMP-activated protein kinase (AMPK). When the intracellular energy levels decline, AMPK responds to the increased level of AMP by AMP binding to the regulatory γ-subunit Snf4 of the Snf1 complex. This leads to the phosphorylation of target TFs and the stimulation of glucose uptake, fatty acid β-oxidation and other reactions that generate ATP. At the same time it inhibits anabolism. AMPK has been described as a guardian of energy homeostasis [Hardie et al. 1998, 2011]. However, in *S. cerevisiae*, AMP does not stimulate Snf1 [Mitchelhill et al. 1994, Woods et al. 1994, Wilson et al. 1996, Broach 2012]. Instead, three activating kinases, Sak1, Tos3 and Elm1 activate Snf1 by phosphorylation of Thr210. Conversely, the protein phosphatase (PP1) complex Reg1/Glc7 complex inhibits Snf1 activity by dephosphorylation of the phosphorylated Thr210. The exact mechanism by which glucose interacts via the three kinases and the dephosphorylation complex is still unclear, but it seems that inactivation of the PP1 complex is the

most likely mechanism of activating Snf1 when the glucose concentration is low [Broach 2012]. Notably, the nitrogen starvation response is also relayed through Snf1. Snf1 is activated by the inactivation of Torc1 when cells are starved of nitrogen [Orlova et al. 2006]. Other stress factors also work through this important master kinase. Oxidative agents, a high sodium chloride concentration and an alkaline pH all lead to phosphorylation of Thr210, while heat shock and sorbitol do not [Hong and Carlson 2007].

Snf1 activates the transcriptional activators Cat8, Adr1 [Young et al. 2003] and Rds2 [Soontorngun et al. 2007]. Rds2 in turn activates gene expression of Hap4 [Broach 2012]. Cat8 and especially Adr1 has been implicated in the activation of genes associated with peroxisomal metabolism and specifically β-oxidation during glucose derepression [Young et al. 2003]. Rds2 is important in the activation of genes for gluconeogenesis by hyperphosphorylation [Soontorngun et al. 2007]. Conversely, phosphorylation of the transcriptional repressor Mig1 by Snf1 leads to its inactivation, resulting in the restoration of gene expression in Mig1 targets [Schuller 2003].

Although phosphorylation of proteins by Snf1 is mostly associated with the activation of gene expression, some examples of inhibition have also been found, for example HXT1 [Tomás-Cobos and Sanz 2002]. Activity of TF Gcn4 is also suppressed by Snf1 activity, but this is independent of transcription of the GCN4 gene [Shirra et al. 2008]. Inhibition of Gcn4 activity is rather affected at the translational level and the mechanism involves Gcn2 and Gcn20. Gcn4 is a transcriptional activator involved with a large number of genes encoding enzymes in the *de novo* synthesis of amino acids and nucleotides during conditions with low levels of these amino acids. The general theme of the cAMP-dependent Ras/PKA pathway, however, is that its heightened activity under high glucose concentrations leads to the inhibition of gene expression, as opposed to the case for Snf1 which usually leads to activation, but under low glucose conditions. In effect, both usually lead to activation of gene expression when the glucose concentration is low.

In higher organisms, β-oxidation of fatty acids occurs in peroxisomes and in mitochondria, while in *S. cerevisiae* it occurs exclusively in peroxisomes [Poirier et al. 2006]. Four TFs are involved, namely Oaf1, Oaf3, Pip2 and Adr1 (Ratnakumar and Young 2010). Several genes involved in peroxisomal proliferation have both the binding motif for the Oaf1/Pip2 TF complex, called the oleate-response element (ORE), and the binding motif for Adr1, called the upstream activating sequence (UAS) [Young et al. 2003]. These motifs are often in close proximity or overlapping. Peroxisomal proliferation and metabolism are both dependent on stimulation by fatty acids, which work via the Oaf1, Oaf3 and Pip2

binding, and by glucose derepression, working via Adr1 binding. Adr1 activity is indirectly stimulated by a high Snf1 activity under conditions of low glucose concentration [Simon et al. 1996] (discussed below).

In the presence of glucose, the addition of oleate results in the binding of the Oaf1/Pip2 TF complex to the genes of β-oxidation to activate gene expression [Wan et al. 2013]. Oaf1/Pip2 is understood as an activator of genes, while other configurations may suppress activity. It seems that the oleic acid response genes are regulated in two patterns. In the first pattern, Oaf1, Oaf3 and Adr1 bind to the DNA and shuts down a gene, and a second in which all four TFs bind to the DNA and cause up-regulation of expression [Smith et al. 2011]. The former set of genes seem to be stress-related, whereas the latter set of genes are those specific to fatty acid metabolism. Thus, regulation of these oleate response genes is highly combinatorial, and the classification of individual TFs as activators or repressors may be flawed in this setting. The complex dynamics have been explored by modelling [Smith et al. 2007, Ratushny et al. 2008]. Oaf1 binds oleate, and this results in the activation of target genes (as the Oaf1/Pip2 complex) which involves binding of the Mediator complex subunit Gal11 [Thakur et al. 2009]. The gene expression level of Pip2 is also stimulated by the heterodimer Oaf1/Pip2. Oaf1 is constitutively expressed, but its activity depends on binding to fatty acids. Both active forms are required to up-regulate Pip2 expression. Pip2 binds with Oaf1 and further increases its own expression. This positive feedback loop may be one reason for the dramatic, switch-like behaviour of the regulation of peroxisomal genes such as POT1.

Peroxisomal gene regulation is, however, not only stimulated by the presence of fatty acids but is also under transcriptional repression by glucose via dephosphorylation of Adr1, which is in some way dependent on Snf1. The current understanding, as reviewed by Ratnakumar and Young [2010], is that Adr1 is activated indirectly by Snf1, by removal of the phosphates on Ser-230 and Ser-98 on the Adr1 protein by some unknown interaction partner. Initially it was postulated that phosphorylation (inactivation) of Adr1 was carried out by the cAMP-dependent Ras/PKA pathway, inhibiting Adr1 activity [Cherry et al. 1989]. Later, it was found that a higher activity of the cAMP-dependent Ras/PKA pathway leads to a higher expression level of the ADR1 gene [Dombek et al. 1997]. Another kinase is, therefore, responsible for the phosphorylation and inhibition of Adr1. This kinase is as yet unidentified. Higher Snf1 kinase activity however, leads to lower phosphorylation of Adr1, and hence, this unknown protein is likely phosphorylated by Snf1. This unknown protein either increases the rate of dephosphorylation of Adr1, or decreases the rate of phosphorylation [Ratnakumar and Young 2010].

Increased activity of the Snf1 pathway thus seems to activate Adr1 dependent gene expression by lowering the phosphorylation state of Adr1. The cAMP-dependent Ras/PKA pathway apparently causes activation of Adr1-dependent genes by a mechanism that leads to the up-regulation of the expression of the ADR1 gene and does not affect the phosphorylation state of the Adr1 protein.

Adr1 is able to activate peroxisomal genes without stimuli by fatty acids via Oaf1 or Pip2. *In vivo* studies using ChIP showed that binding of a constitutively active form of Adr1, Adr1^{c}, which cannot be phosphorylated and thereby inactivated, was improved compared to wild-type Adr1 [Ratnakumar and Young 2010]. However, the increased binding did not always correlate with increased expression levels of the target genes. Thus, the regulating domain of Adr1 that carries the phosphorylation, and perhaps even the charge of the phosphate group itself, probably has some effect that alters chromatin structure or the formation of the pre-initiation complex. Phosphorylation has multiple effects on the activity of Adr1, as was observed with the Adr1^{c} strain. It has effects on (*a*) the recruitment of the SAGA co-activator complex, (*b*) on the remodelling of chromatin by alternative histone variants, and (*c*) on the phosphorylation state and activity of RNA polymerase II [Ratnakumar and Young 2010]. Using ChiP, the authors showed that that more co-activating proteins Ada1 and Gcn5 were recruited to promoters by the Adr1^{c} variant compared to the wild type, although the degree of improved recruitment varied among genes. Under glucose repression, the constitutive activity of the Adr1^{c} protein was not dependent on the involvement of the Mediator complex or the SWI/SNF complex, but was dependent on involvement of the SAGA complex. Deletion of the acetyl transferase Gcn5 and scaffold Ada1 was deleterious to high constitutive gene expression in the Adr1^{c} strain, while deletion in Mediator and SNF/SWI complexes did not make a difference [Ratnakumar and Young 2010].

Recently, Adr1 was shown to cooperate with a histone variant, H2A.Z (histone variant Htz1), which was required to expel the repressing activity of Oaf3 in response to oleate [Wan et al. 2013]. Adr1 also facilitated the insertion of the H2A.Z histone variant into chromatin. The alternative histone H2A.Z can replace H2A, resulting in anti-silenced chromatin. As mentioned above, the combination of Oaf1, Oaf3, Pip2 and Adr1 binding sites is reflective of the combinatorial nature of eukaryotic gene regulation. The combination of binding sites for Oaf1, Oaf3 and Adr1, without Pip2, is a configuration that promotes repression of a gene in response to oleate, whereas if a Pip2 site is present, the gene would respond positively to oleate [Wan et al. 2013]. The binding sites for Oaf1, Oaf3 and Adr1 have been shown to be enriched at regions within 10 kb of the telomeres of chromosomes in *S. cerevisiae* [Smith et al. 2011]. These so-called subtelomeric regions are prone to chromatin silencing in response to

environmental signals, and Oaf1, Oaf3 and Adr1 have been described as regulators of subtelomeric silencing [Smith et al. 2011]. The positioning of oleate activated genes (when a Pip2 binding site was present) is not biased towards subtelometric regions, but randomly distributed. Oaf1 seems to be a main player in silencing at X-elements close to subtelomeric regions in response to oleate. However, a number of other TFs also bind X-elements [Smith et al. 2011]. In summary, not only is the epigenetic code combinatorial, but may also be regional – at least in the case of the subtelometric regions of *S. cerevisiae*.

Oaf1 and Pip2 also bind to some non-peroxisomal genes such as the citrate synthase gene of the citric acid cycle, CIT1 [Karpichev et al. 1998]. ADH2, the alcohol dehydrogenase that is understood to be responsible for the utilisation of ethanol, is also strongly induced by the addition of oleate [Ratnakumar and Young 2010]. Physical interaction and synergy between Adr1 and Cat8 in the ADH2 promoter may be a reason for the strong response to glucose derepression. In addition to binding sites for Adr1 and Cat8, there are also two half-sites where Oaf1 may bind. Oaf1 might similarly cooperate with Adr1 in the ADH2 gene promoter when the signal was the presence of fatty acids [Ratnakumar and Young 2010].

In summary, alleviation of glucose repression is governed by master kinases in which cAMP-dependent Ras/PKA and Snf1 are predominant. These generally activate gene expression at low glucose concentrations by their effects on TFs, where Snf1 kinase activity is increased and Ras/PKA activity is decreased. Glucose derepression leads to phenotype switching to pseudohyphae in a fraction of the cells and this response is initiated by the TFs Phd1, Mss11, Ash1, Flo8, Msn1, Haa1, Ste12, Tec1, Mga1, Nrg1, Nrg2, Sfl1 and Sok2, governed by Phd1 and Mga1 [Broach 2012]. Alternative carbon source utilisation is activated by a separate set of TFs, namely Hap2-5, Rtg1-3, Cat8, Sip4, Mig1, Adr1, Oaf1, Oaf3 and Pip2 [Broach 2012, Ratnakumar and Young 2010]. Peroxisomal gene expression depends on the TFs Adr1, Oaf1, Oaf3 and Pip2. While the combination of Oaf1 and Pip2 binding stimulates gene expression in response of oleate addition, regardless of the concentration of glucose, Adr1 activates gene expression in response to low a glucose concentration. Snf1 indirectly activates Adr1 via some unknown mechanism of dephosphorylation of Adr1, whereas Ras/PKA activates Adr1 activity by some unknown mechanism of up-regulating expression of the Adr1 gene. Snf1 is thus a convergence point for stimuli not only from glucose limitation, but also from stress factors [Orlova et al. 2006, Hong and Carlson 2007]. These stress factors, including nitrogen starvation (working via Torc1), lead to a higher activity of Snf1, activating stress responses. In the case of nitrogen starvation, activated Snf1 rather leads to a lower protein level of the target TF Gcn4 [Shirra et al. 2008]. Hence, glucose limitation and

nitrogen limitation would both cause an increase in the alternative carbon source utilisation capacity and a decrease in *de novo* biosynthetic capacity through Snf1, in a balanced manner.

Since the main aim in this thesis was to establish the regulatory design and the master regulators in the non-model species *K. marxianus*, it was necessary to find evidence for the differential expression of genes as well as for TF binding sites in the DNA of *K. marxianus*. The methods involved in discovering TF binding sites are discussed briefly below, along with NGS and selected methods required for the study of metabolism.

Transcription factors and DNA binding motifs

Transcription factors bind to DNA at specific sequence patterns called motifs, which can be discovered and described in a number of ways [Sinha and Tompa 2000, Lawrence et al. 1993, Stormo et al. 1982, Mathelier and Wasserman 2013, Zeng et al. 2016]. Consensus sequences describe the most likely single sequence in a group of related sequences bound by the DNA binding protein. If the binding of a TF is highly precise in terms of the target sequence, such consensus motifs may be found easily by a string matching algorithm. Consensus strings are, however, often too restrictive to be realistic, since a DNA binding protein would typically bind a variety of related sequences [Stormo et al. 1982]. A more accurate motif would allow degeneracy. A type of string expression that allows degeneracy is known as a regular expression. Since DNA binding motifs for TFs usually are short, a relatively restrictive regular expression could suitably be used to find putative TF binding sites. The longer and more restrictive the regulator expression is, the lower the probability of obtaining false positives. (In Chapter 8 regular expressions are used to model the motifs for the TFs Mig1, Adr1 and Aft1.) Regular expressions allow a degenerate description, but does not capture the fact that certain sequences are better bound by the TF than others. To the contrary, position specific probability matrices (PPMs) capture the probability of each base at a specific position being matched by the motif model [Stormo et al. 1982]. It is the most commonly used description and is the type most often stored in motif databases such as JASPAR [Mathelier et al. 2014]. It cannot account for interactions between bases in the motif, however, which determines the three-dimensional shape of the DNA. Markov models or Hidden Markov Models improve on weight matrices by capturing these interactions [Mathelier and Wasserman 2013]. However, usually there are not sufficient data to represent motifs as Markov Models. Artificial neural networks is another such approach that may even capture long-range interactions, but is not frequently used for representing TF binding motifs [Zeng et al. 2016].

When scanning a PPM against a stretch of DNA, a motif score is assigned at each position along the DNA for each base in the PPM, and since the probability of observing each base along the PPM is considered independent from other bases, all the probabilities are multiplied [Mathelier and Wasserman 2013]. The higher this score, the better the match. To account for the background frequencies of nucleotide bases, a background model is also defined. In the case of PPMs, this could simply be a multiplication of background frequencies of the four nucleotides. As the background frequencies may also change depending on the distance from the coding region of a gene, the model should also incorporate this. (The effect of this distance on background frequencies is explored in Chapters 5 and 6.) A motif likelihood ratio is then calculated by normalising the motif score with the background score, as below.

$$L\mathrm{m} = \frac{\prod_i^n m\,[i]}{\prod_i^n b\,[i]}$$

The symbol *Lm* is used throughout this thesis to signify the motif likelihood ratio, which is calculated at any position along the length of the DNA region defined as the regulatory region. The symbol *m* is the entry in the PPM at position *i* in the subsequence that stretches from 1 to *n*, and *b* is the background frequency of the relevant nucleotide base at a distance *d* from the transcription start site. An *Lm* value of 1 signifies a completely random match which carries no significance, whereas a large number indicates a good match to the PPM compared to the background model. PPMs differ both in their length and in their degeneracy. Longer PPMs are more restrictive and are less degenerate (more precise) PPMs. Unfortunately, whatever the method of description of motifs, the degeneracy of DNA binding preference of TFs together with the very short nature of most DNA binding motifs result in many false positives during their *de novo* prediction. False positives in the prediction of DNA binding sites is a major concern [Hannenhalli 2008]. Additional sources of evidence need to be used to increase the accuracy of assigning TF-target gene interactions. (This is the topic of Chapters 4-6.)

The principle of calculating likelihood ratios is very useful and was applied throughout this thesis in different contexts. A very useful aspect regarding likelihoods, is that they are convenient for combining multiple sources of evidence. The idea behind the use of likelihoods as a Bayesian classifier is explained next, before discussing the data generation methods used in this work. This idea originated in the analysis of protein-protein interaction datasets and will be discussed in this context, although this was adapted to the case of gene regulatory networks in subsequent chapters. Others have also followed a Bayesian networks approach [Heckerman et al. 1995, Friedman et al. 2003, Isci et al. 2011].

A probabilistic approach to combining datasets

Modern high-throughput experimental methods have transformed the manner in which scientific discoveries are being made. These are especially relevant in areas such as the elucidation of protein-protein interactions by methods such as yeast-two-hybrid [Fields and Song 1989] and variants of affinity chromatography coupled to mass spectrometry (reviewed in Smits and Vermeulen 2016). It has been estimated that, by using any single large high-throughput protein-protein interaction dataset on its own, a large fraction of the observed interactions would be false positives (low specificity) due to a number of experimental artefacts [Collins et al. 2007]. Also, a significant fraction of interactions are not observed in a particular experiment due to a low true positive rate (low sensitivity). However, multiple experiments are also conducted on the same organism, using different types of experiment and possibly in different laboratories, and also possibly repeats of the same experiment. Observing the same interaction multiple times should increase the confidence in a certain interaction, especially if different experimental methods were used. For instance, Gavin et al. [2002] used the TAP system of affinity capture mass spectrometry, while Ho et al. [2002] used a single purification step of the captured proteins, but in which the targets were over-expressed. The TAP system has the advantage of fewer false positives due to fewer non-specific interactions. In the system in which proteins with a low natural abundance are over-expressed, competition by proteins with a high natural abundance is mitigated. Combining such datasets should thus be especially useful for altogether different experimental types, which each have different strengths and weaknesses. The consensus approach to combining datasets would be to use only the overlapping set of observations among experiments. However, using only the consensus set would result in keeping only very few interactions, as different experiments tend to focus on different types of proteins and interactions. As an important innovation, Collins et al. [2007] developed a probabilistic strategy to combine such datasets to end up with increased likelihoods for some interactions. A naïve Bayesian classifier states the likelihood ratio of two opposing hypotheses: in the numerator, the probability (hypothesis) that any given experiment yielded a true positive, and in the denominator, the probability (hypothesis) that any given experiment yielded a false positive. The likelihood ratio of an experiment, *Le* (below), is related to the confidence one could have in a certain type of experiment.

$$sensitivity = \frac{TP}{TP + FN}$$

$$false\ positive\ rate = \frac{FP}{FP + TN}$$

$$Le = \frac{sensitivity}{false\ positive\ rate}$$

One beauty of this approach is that, for each interaction observed, the likelihoods of the multiple data types can be multiplied to end up with a final likelihood as below.

$$L = La \times Lb \times Lc \times Ld \times \ldots$$

In the above likelihood equation, the likelihood ratios *La*, *Lb*, …*Ln* are specific to the type of experiment, and more than one dataset of the same type may be used. Some experiments will make a greater contribution to the final likelihood than other methods, as they have a better sensitivity or specificity, or both. At this point it is important to state that the final likelihood is not exactly the interpretation of "how many times more likely is an observation a true positive as opposed to a false positive". The equation originates from the Bayesian classifier, which requires also knowing how many true interactions there are in the first place, which is unknown. Hence, this *naïve* Bayesian formulation here is more suitably called a likelihood *rank* ratio. All observations are ranked by the final likelihood rank, and the best interactions are chosen, where the number may be some estimate of the total number of interactions, such as 10 000.

The (naïve) Bayesian method requires Gold Standard (GS) positive and negative datasets to estimate the individual likelihood ratio values for the different experimental types. The GS positive is some dataset with a set of high-confidence observations, based on some high-accuracy, low throughput experiment. For protein-protein interactions it may be done by co-crystallization or other methods which make it obvious that the interactions are highly specific. The GS negative is a dataset that contains interactions that *do not* exist between proteins. There are very few cases reporting that proteins do not interact, and hence such as dataset has to be deduced by logic. The expression of proteins in a complex is often correlated, as they are required at the same time. Finding proteins with anti-correlated expression levels is a logical way to derive a GS negative set.

The calculation of likelihood rank ratio values for an experiment type is demonstrated graphically in Figure 1. The goal is to discover the interactions that are not in the GS positive set (already known) by using some high-throughput method. Of the observations in the high-throughput dataset, the fraction in the GS positive interactions that are observed (true positives) is calculated, as well as the fraction that fall in the GS negative (false positives) interactions. The ratio is then the likelihood rank ratio. Note that even if a test is not very sensitive to detect interactions, its likelihood rank ratio may be high if it was very specific (very few false observations).

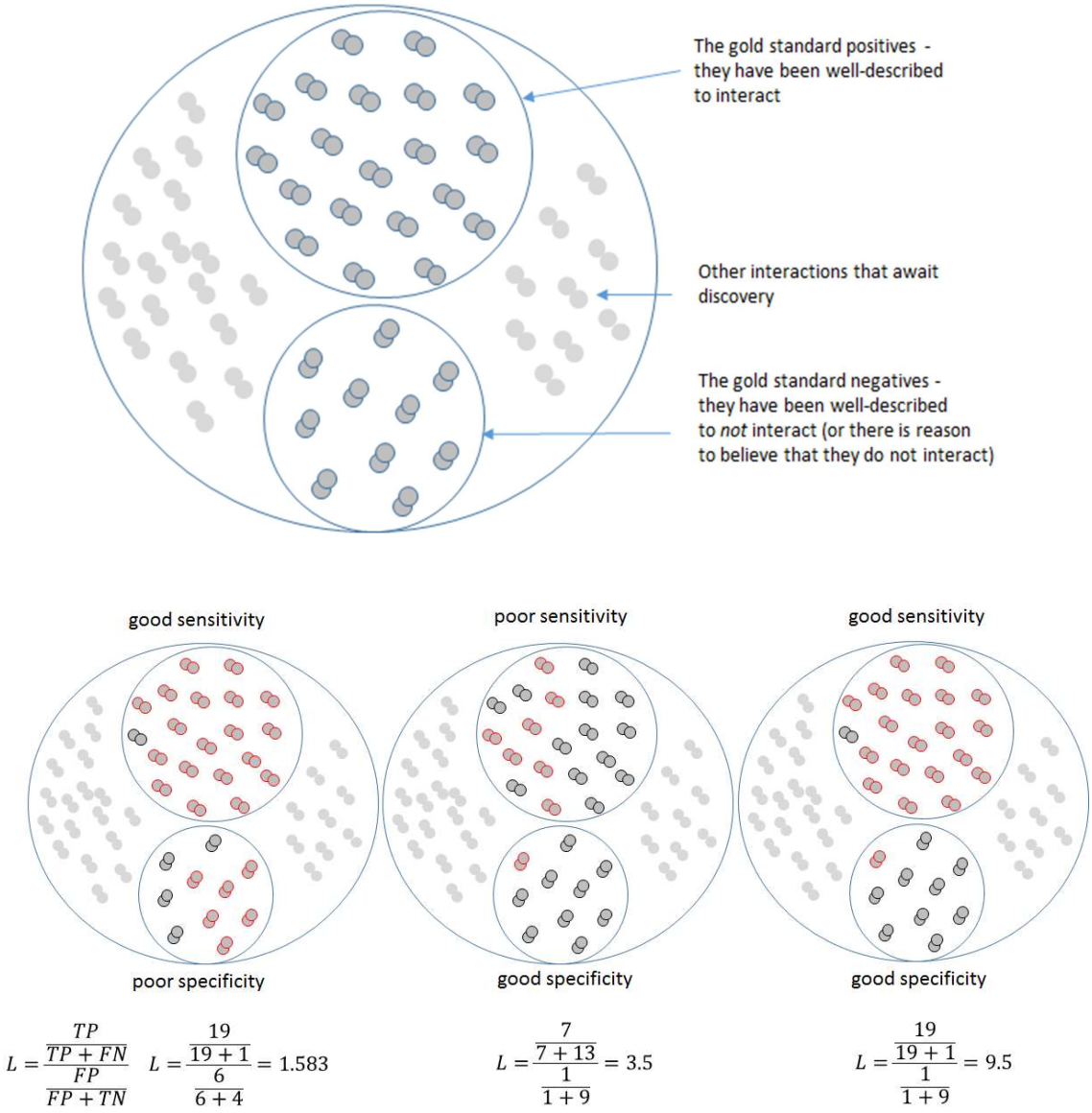


Figure 1. Top: Gold Standard datasets and discoveries to be made. Bottom: Calculation of likelihood rank ratios. Red: observed in high-throughput dataset. Black: not observed in high-throughput dataset.

Although the rank ratio approach has first been demonstrated on protein-protein interaction datasets of affinity capture MS [Collins et al. 2007], it has tremendous potential in combination with a variety of other types of data, including signal transduction networks and gene regulatory networks, and even metabolic signalling networks. It is important to realise that the analytical technologies characteristic of the 21st century necessitates the use of such theoretical frameworks for the integration of multiple sources of data. From a practical and economic perspective, it is also becoming increasingly important

to harness these large datasets that are freely available in public databases for applicability in any biotechnology-related product, and not limited to those involving model organisms. (This idea will be expanded on in Chapters 5 and 6 involving gene regulatory networks and in Chapter 7 involving kinase signalling networks.)

Exploratory data analysis, gene regulatory networks and statistical enrichment

With the advent of the technologies of microarray analysis and high-end MS/MS proteomics around the turn of the century, and again with the disruptive advances in NGS applications recently, such as RNA-seq, there is an increased interest in exploratory data analysis of these high-content datasets. Whereas there has been a strong emphasis on detailed enzyme kinetic studies and the mechanistic modelling of metabolism, the focus has changed substantially towards the genetic level in recent years. Microarrays and RNA-seq both provide a tremendous amount of information using standardised methods, and compared to enzyme kinetic studies, metabolomics and mass spectrometry based proteomics, provides a much easier entry point into systems studies. Tools for basic data analyses of NGS data, such as *de novo* assembly of genomes and RNA transcripts, read-mapping, variant analysis and differential expression testing of RNA-seq, are now widely available and easy to use in online programmes such as Galaxy [Afgan et al. 2015], as are the tools for microarray analysis, including the many algorithms in BioConductor [Gentleman et al. 2004] and Chipster [Kallio et al. 2011]. However, utilisation of these data for a systems level understanding lags far behind the current capabilities for raw data processing. Some of the key concepts deserve a brief review, which are the cornerstones in moving from large datasets to a meaningful understanding of cause and effect in biochemical systems. These are ontologies, enrichment statistics, *in silico* networks, the reverse engineering approach and probabilistic networks. The principles are not limited to gene expression, and have been used throughout this thesis. Importantly, as is explained in Chapters 5 and 6, the mechanistic biochemical details in such analyses will become increasingly important in the years to come.

An important development in the bioinformatics community was the idea of ontologies, in particular Gene Ontology (GO) [Ashburner et al. 2000], to make sense of large microarray datasets. An ontology is a structured vocabulary of terms ("GO terms"), each with a clear biological, biochemical or physiological meaning and related to each other in a tree structure (see Figure 2). GO actually consists for three ontologies: GO cellular location, GO molecular function and GO biological process. For the example of GO cellular location in Figure 2, the general term 'vacuole' maps to various more specific

types of vacuoles, which again maps to even more specific types, some of which may only be found in some types of organisms. In gene and protein annotation databases such as the UniProt protein database [Boutet et al. 2016], these GO terms are associated with molecular sequences. UniProt consists of two types of sequences. Firstly, the Swiss-Prot sequences, in which all examples and the GO terms are manually curated. Secondly, the Trembl sequences which are automatically computationally annotated, based on their sequence similarity with other well-known (Swiss-Prot) sequences. Protein sequences enter UniProt as Trembl entries and progress to Swiss-Prot. Multiple proteins may contain the same GO term; for instance, a variety of genes will occur in lytic vacuoles and hence carry the term 'lytic vacuole'. For each term in the ontology, the links (mappings) can be followed from the GO term to all of the protein entries in the database carrying that term, and subsequently their differential expressions statistics in a microarray or RNA-seq dataset may be looked up. The group of genes mapping to an ontology term is termed a gene set. Instead of performing this lookup process manually, algorithms can do this efficiently.

A gene set enrichment can subsequently be calculated for each term to test whether the particular set of genes is differentially expressed more significantly than is expected from a randomly picked gene set of the same number of genes [Ideker et al. 2002]. Enrichment statistics have been reviewed by Maciejewski [2013]. More than one way to perform enrichment exists, including the use of discrete statistical distributions such as the hypergeometric distribution, in which the number of up or down-regulated genes are used and compared to the background. The most sensitive method may be the one developed by the Nielsen group [Patil and Nielsen 2005, Oliveira et al. 2008], which will be referred to as the Z-score method throughout this thesis. In this method, the probability values (p-values) or their corrected form for multiple comparisons (q-values), originating from differential expression calculation of a microarray dataset (adapted for RNA-seq in Chapter 3), are converted to Z-scores. In simple terms, the Z-score refers to the number of standard deviations away from a mean. Z-scores can be summed for the whole gene set and normalised to the average of the background as well as the standard deviation of the background, for a specific number of genes [Patil and Nielsen 2005, Oliveira et al. 2008]. The calculation of the enrichment score is described by the formula below.

$$\mathrm{S} = \frac{\mathrm{Z}(\mathrm{total}, \mathrm{Test}) - \mathrm{Mean}(\mathrm{Z}, \mathrm{Background})}{\mathrm{Standard\ deviation}(\mathrm{Z}, \mathrm{Background})}$$

As this is in effect a bootstrapping method, the background sets are simulated by randomly sampled gene sets containing a specific number of genes, compiling the total Z scores into a list and subsequently calculating the average and standard deviation for each number of genes. This should be performed for a variety of gene set sizes, covering all gene set sizes so that the enrichment value

of any gene set can be calculated after looking up the background mean and standard deviation in a table.

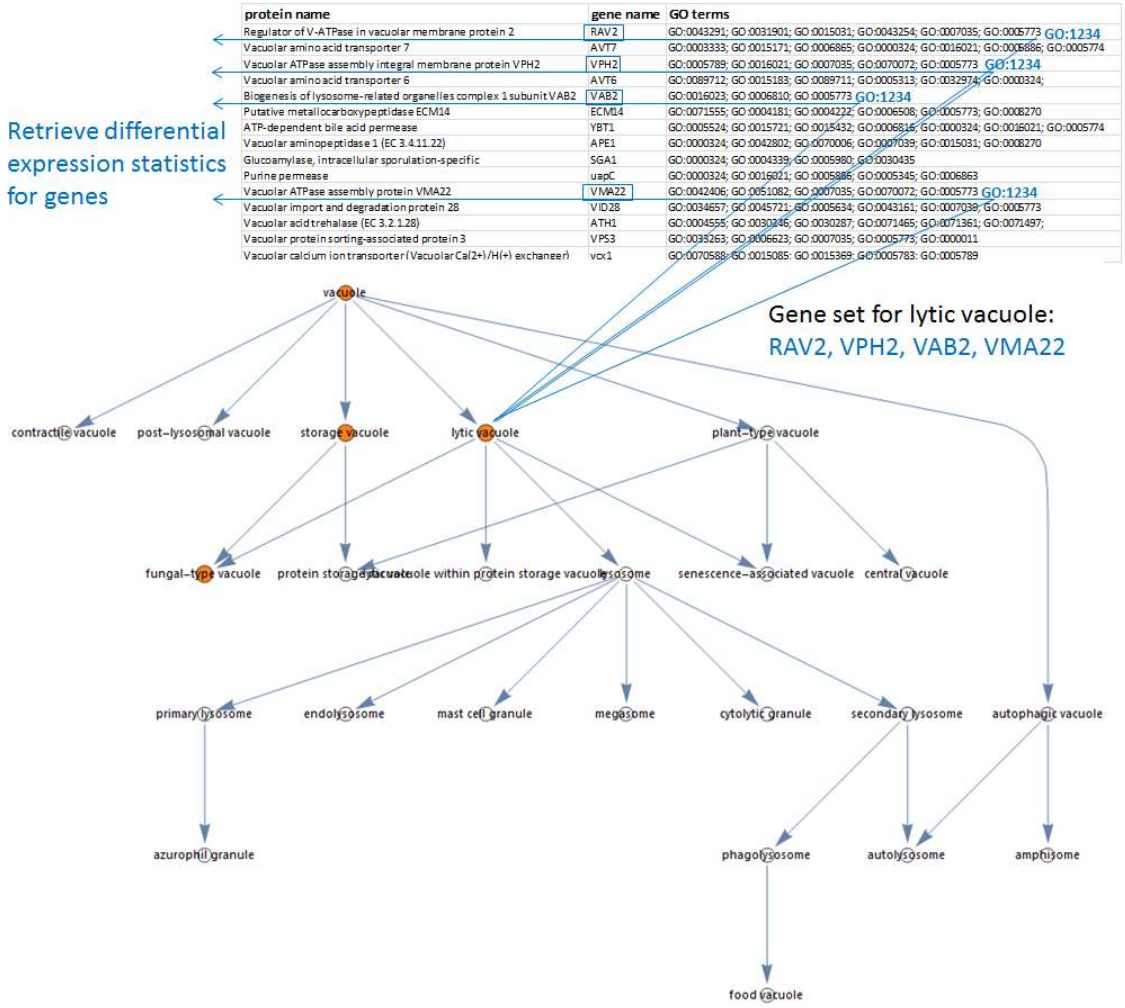


Figure 2. The structure of the GO cellular location ontology below 'vacuole'. The tree was generated with software designed as part of this project, named *Reactomica*. (See Chapter 3 for more examples of GO enrichment and visualization.)

Not only is it challenging to implement such an algorithm in a streamlined form in an integrated systems analysis software (part of this work), but it is also a substantial challenge to get a grasp of the output. GO [Ashburner et al. 2000] currently has 40 143 terms (July 2016). A table with this many terms is not easy to navigate and too large to use efficiently in spreadsheets. The use of visualization as a tree or network structure could greatly assist in obtaining an overview of enriched terms, as some terms have largely the same meaning. Miniature versions of the full GO ontology are also helpful,

especially for visualization. A good example is GO_slim for yeast, which contains most of the important terms to gain an initial overview of gene regulation. However, it falls short in more detailed analyses and has to be reverted to the elaborate full GO ontology (see Chapter 3). As far as the author is aware, currently there is no software that can perform the combination of navigating large ontologies, carry out enrichment statistics, visualise enrichments as a tree structure and create streamlined versions, although some of these aspects are found in ontology editors such as OBOEdit [Day-Richter et al. 2007] and more recently as a plugin for the network visualisation programme Cytoscape [Shannon et al. 2003].

While some of the abovementioned shortcomings related to gene set enrichment could still be overcome by combining a variety of programmes, other specialised potential applications for ontologies still require development. Three examples are given below, which were the main reasons why a substantial amount of algorithm development was done on ontology related work in this investigation. In the relatively simple case of trying to find all genes that have been associated with a certain subcellular localization such as the lytic vacuole, a substantial amount of searching may have to be done to find the correct GO term, as it is embedded within a large database of terms in which a text search might not be very efficient. Visualization of the relationships in a small window (Figure 2) is required to pick the correct term, such that the term could be used to find the relevant proteins encoded by the genome of interest. A more sophisticated use of GO is to automatically create compartmentalised versions of metabolic pathway maps, which may be used to separate RNA-seq data into the different subcellular compartments. This application has much potential and may reveal novel insights (see Chapter 3). Another application is to apply the co-assignment of GO terms between a transcription factor and a potential target as a likelihood ratio. This was used in Chapters 5 and 6 in creating the first gene regulatory network for *K. marxianus* UFS-Y2791. Also, this may have other applications such as improving the assignments of protein-protein interactions [Collins et al. 2007].

The concept of gene set enrichment is not limited to ontologies. The concept of metabolic pathways as gene sets provides an efficient method to reveal the cellular state or the global differential gene expression, since pathways have an explicit meaning and the relationship between metabolic pathways is well understood by biologists. Pathway Tools, which operates on the MetaCyc database, provides mappings from EC numbers or GO terms to a pathway genome database (PGDB) via the PathoLogic algorithm [Karp et al. 2009], while UniProt has recently started to implement pathway mappings for proteins. These terms could be used similarly to GO terms for pathway enrichments. A significant limitation currently is the representation of complete metabolic charts on a single

rendering. Metabolic reactions are hypergraphs (more than one reactant or product), thus any automated rendering is bound to be very complex. An innovative approach was taken in this work by representing pathway enrichments as nodes in a pathway-to-pathway network, while representing textbook-style metabolic pathway maps of individual key pathways individually, painted with expression levels or differential expression statistics (see Chapter 3). This complements the cellular overview approach followed by a visualization tool such as is found in Pathway Tools [Karp et al. 2009], which currently is only for the microarray format.

Apart from ontologies and pathways, any sensible grouping of genes could, in principle, be used as a gene set which could be tested for enrichment and visualised in some form. Recently, the concept was generalised to 'reporters' [Oliveira et al. 2008]. Reporter metabolites are metabolites that are likely differentially regulated by the enzymes that produce or consume them [Oliveira et al. 2008]. This concept, used in Chapter 3, predicts that the concentration of a metabolite is likely altered, but it cannot predict whether the concentration may be higher or lower, as is often the case with enrichment statistics. Notably, it predicts changes in the concentrations of metabolites by using differential gene expression data, and is thus a complement or even a replacement for the more technically challenging metabolomics done by using mass spectrometry methods. Protein complexes may also be considered as reporters.

A particularly interesting idea is that of reporter TFs [Oliveira et al. 2008]. The gene set for a TF is the set of targets that a TF affects by gene regulation. Finding a high enrichment statistic for such a TF (reporter) would suggest that it explains the differential expression of its targets, and hence that it is, at least in part, the reason for their differential expression. Knowledge of the differentially active TF in a response provides substantial insight into the signalling pathways that are active, but can also provide a route to manipulating the cell at a higher regulatory level to cause a dramatic change in cellular phenotype. Consider the case in which it might be found that a single TF regulates the switch between aerobic and fermentative metabolism, or between the normal phenotype and cancer. Genetic manipulation (in case of a microorganism) or treatment with a targeted drug (for humans) could then be performed to alter the phenotype by targeting the relevant TF. Finding such a major switch in potential biofuel producers such as *K. marxianus* that could lead to an industrially important phenotype, is highly attractive.

A major strength of the enrichment (or reporter) approach is that it is robust towards individual potential erroneous assignments of interactions. The knowledge gained on the differential activity of

the TF is more important than the correctness of essentially all the assigned interactions. Since the sample (gene set) is sufficiently large, the total observed change (enrichment) stands out above the background enrichment; hence this is also a sensitive method in the detection of altered biochemical activity. The major constraint currently with reporter TFs specifically, however, is that a genome-scale gene regulatory network is required for the species of interest. Gene regulation is one of the most relevant topics in the biosciences today, but the elucidation of a genome-wide gene regulatory network is a major challenge. A substantial effort was invested into reconstructing the first *in silico* genome scale gene regulatory network for *K. marxianus*, which is presented in Chapters 5 and 6. A special likelihood based method was developed which includes multiple sources of evidence, building on the idea of a naïve Bayesian classifier, which was described earlier in this review. This gene regulatory network was used to obtain the target gene sets for TFs, which were used to calculate enrichment statistics using both the Z-score method and the Hypergeometric distribution [Chapter 6].

The gene regulatory network also turns out to be an important link between transcriptomics data and the phosphorylation signal transduction network. In Chapter 7, a method is presented by which the activity of important kinases is predicted by RNA-seq data, which by itself does not measure the phosphorylation state of proteins.

Although we are decades from grasping the full complexity of eukaryotic gene regulation due to its combinatorial nature [Voet and Voet 2011], suitable and robust analytical frameworks such as the enrichment statistics and the likelihood based network approaches used in this work, should lead to deep insight into cellular regulation of both model and non-model species. This is an exciting time for computational and high-throughput biology studies in eukaryotic gene regulation. The key catalyst that is now available to allow major gains in knowledge at a rapid pace for non-model species, is the advent of NGS. Some of the applications of NGS will be discussed briefly.

Next-generation sequencing

Below is a brief overview of the sequencing technologies, the types of NGS experiments and important algorithms.

Sequencing technologies

Sanger sequencing enabled molecular biology [Sanger and Coulson 1975]. It provided high quality reads, but could only be performed in low throughput. In order to complete sequencing the human

genome, large sequencing centres were set up with multiple sequencers. This multi-national effort led to the first human genome sequence, which required several years. The first high-throughput sequencers, such as the Roche 454, revolutionised sequencing by parallelising the process in miniature wells and using PCR amplification of fragments of DNA [Wheeler et al. 2008]. However, the next generation of sequencers caused another revolution, in which the throughput dramatically increased, which coincided with a substantial cost reduction [Bentley et al. 2008]. The Illumina systems currently leads the NGS market. The technology works by hybridising single-stranded DNA fragments onto a specialised transparent slide. Each fragment is amplified, resulting in a spot of identical fragments. In the next step, the complementary strands are synthesised using specialised deoxyribonucleotides that can be detected by an optical reader during each round of nucleotide incorporation. Each nucleotide is detected as a different colour. The sequence of appearance of colours corresponding to each spot is then converted into a nucleotide sequence. The most popular Illumina instrument currently is the Illumina MiSeq, which provides approximately 15 Gb of sequence per run, with read lengths of up to 300 bp. The larger Illumina systems such as the HiScanSQ produce shorter reads of 75-100 bp but at a larger scale. A competing technology is the Ion Torrent/Ion Proton instruments. The detection of base pair inclusion is, instead of optical detection, based on the detection of small changes in pH originating from the inclusion of nucleotide bases, which saves on cost [Rusk 2011]. The process is called ion semiconductor sequencing. With both these next-generation sequencing technologies, the major constraint is the read length. Since the read quality deteriorates with an increase in read length, the longest reads of sufficient quality are approximately 300 bp. Many applications in NGS requires reads to be as long as possible. This is especially important when genome assemblies have to be done where repetitive regions have to be spanned. Short reads cannot span repetitive regions and hence assemblies become inaccurate if the read lengths are shorter than the length of the repeating regions [Compeau et al. 2011]. Another revolution in sequencing is expected in the form of single molecule sequencing. In this method, no synthesis or PCR is applied, which also eliminates bias towards certain types of sequence. Instead, the method directly reads the base pairs in individual DNA or even RNA molecules [van Dijk et al. 2014]. Moreover, the read length promises to be very long.

An innovation in sequencing technology that is available on platforms such as Illumina is paired-end sequencing, which was explained by Bentley et al. [2008] and is summarised below. After the suitable number of PCR cycles has been performed in which nucleotides are incorporated onto the single-stranded reads and the quality drops to a non-usable level, a process called bridge amplification is performed. In this step, the DNA is bent over such that the free end binds with the slide, forming a bridge. Several steps are then carried out to effectively turn the fragment around such that the same

number of PCR steps (sequencing) can be done from the other end. The result is a dual set of reads, each pair corresponding to the sequences of the ends of a fragment. The data pair is then stored in two separate *.fastq* files and the pairs are kept in the same order such that they could be used together during read-mapping and other steps in data analysis. Paired-end sequencing effectively doubles the amount of data, and since the approximate distance between the two reads on the opposite ends of a fragment is known, this information could be used to improve assemblies. A variation on this is long mate-pair sequencing [Bentley et al. 2008]. Long fragments of up to 25 Kb are selected, the ends of the fragments labelled with a biotin tag and the fragments circularised by joining the ends of a fragment. These circular DNAs are then fragmented and the part containing the two original ends of the long fragment, with the two biotin tags in the centre, purified by making use of the affinity of biotin for avidin. These shorter fragments, containing the inverted ends, are then ligated to the adapters for paired-end sequencing, followed by the standard sequencing protocol. Mate-pair reads are especially useful for *de-novo* assembly of genomes, or genome finishing by which contigs from *de novo* assemblies can be sewn together by using the approximate distance between the reads on the ends of long fragments.

Genome assembly

Two methods of genome assembly can be distinguished. The first is read-mapping of NGS data to a reference genome, which would mostly be that of a different strain or individual of the same species. This process is useful for detecting genetic differences among individuals in a population, such as for performing genome-wide association studies in humans, or to detect the differences between strains of pathogenic or non-pathogenic *Escherichia coli*. A good algorithm for read-mapping is Bowtie [Langmead et al. 2009, Langmead and Salzberg 2012].

Another approach is *de novo* assembly, which does not require a reference genome. The algorithm effectively looks for overlaps among NGS reads and sew these together into contigs. The best established method is *de novo* assembly based on the de Bruijn graph method [Compeau et al. 2011]. This algorithm was coded to demonstrate the process of *de novo* genome assembly using de Bruijn graphs (Figure 3). Briefly, the algorithm first separates each NGS read (*1* in the figure) into overlapping k-mers (k-letter words) (*2* in the figure). All the adjacent k-mers in a read are marked as adjacent by constructing a connected graph where the nodes are the k-mers in a read. Next, all identical, or nearly identical k-mers between reads are mapped to one another to result in a graph containing all the data (*3*). Subsequently, these identical nodes (*4*) are collapsed into single k-mer nodes. In this manner, the arcs (edges) become increasingly dense with the evidence of connections between adjacent k-mers

(*5*). Finally, the graph is traversed to find the path along k-mers with the most supporting evidence (*6*), which is then sewn together into the final sequence and coverage at each base (*7*).

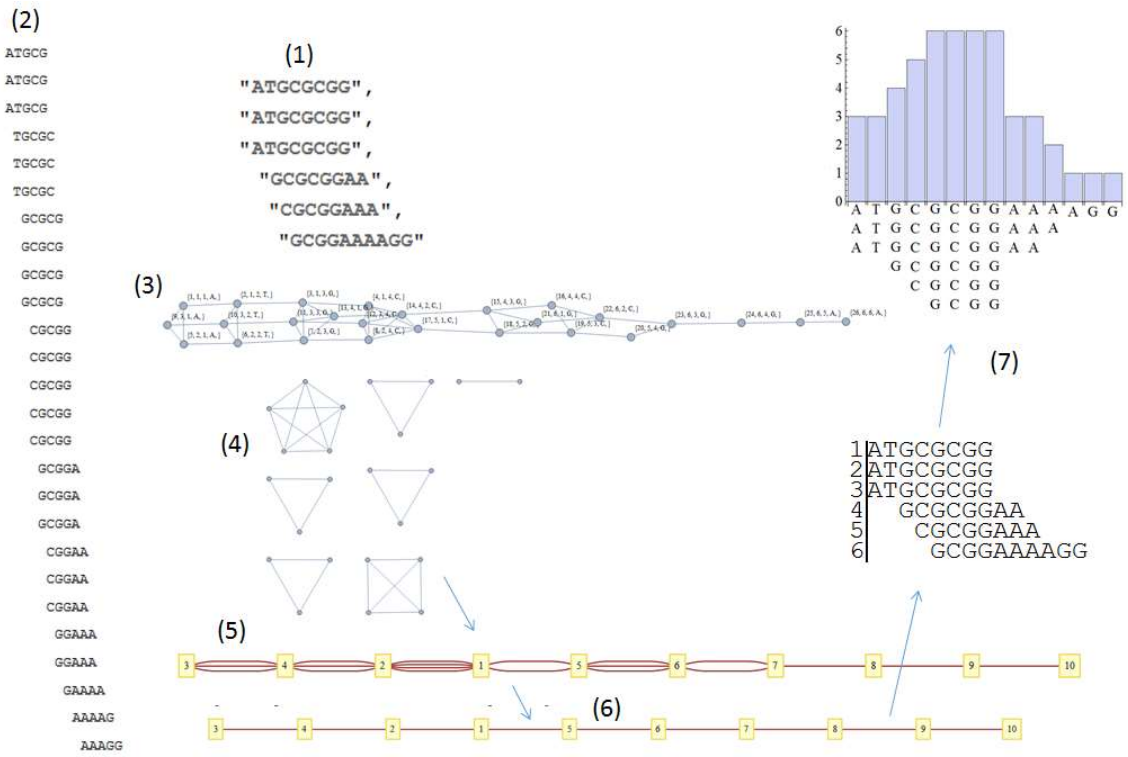


Figure 3. The process of *de novo* assembly using the de Bruijn graph method. Images are the output from an assembler coded as part of the project, for demonstration purposes. The process is explained in the text.

The de Bruijn graph algorithm uses a large amount of computer memory, but can be parallelised, as was done for Abyss [Simpson et al. 2009]. Another popular *de novo* assembler is Velvet [Zerbino 2008]. Assembly statistics are useful to characterise the quality of a *de novo* assembly. The values of N50, N75 and N90 are respectively defined as 50% of the genome residing in contigs equal or longer than the N50 value, 75% of the genome in contigs equal or longer than the N75 value, and so forth.

Transcriptomics and RNA-seq

Transcriptomics involves the measurement of RNA levels in the cell at the genome-wide scale. The development of DNA microarrays marked the beginning of functional genomics [Schena et al. 1995]. Typically, samples are taken from cells of different types or that have been cultivated under different

conditions, the RNA purified, reverse transcribed, differentially labelled using fluorescent labels, and the labelled cDNA hybridised to a slide containing spots of oligonucleotides that represent each of the genes in a genome [Schena et al. 1995]. Usually two samples are mixed, allowing RNA molecules from the two samples to compete for the same sites on the array. Subsequently, differential expression of each gene is calculated from the ratio of the binding of the two samples, and the experiment usually performed in triplicate to calculate p-values. Several steps of normalisation are also required in the data analysis, and it is important to take the same amount of RNA from the two samples. It would also be highly detrimental if the sample contained RNA from a contaminant species; thus mixed cultures cannot be studied. Findings of differentially expressed genes are often confirmed by using a low-throughput method such as quantitative, real-time PCR and the standard of Minimum Information About a Microarray Experiment (MIAME) has been established for the publication of such data [Brazma et al. 2001]. Nevertheless, expression microarray has become so popular that there are by now a tremendous number of these datasets published. Another revolution in the field of functional genomics has resulted from the advent of NGS applied in various forms. NGS has been applied to the measurement of RNA levels, which is referred to as RNA-seq. The experimental and data analytical protocols have been described by Trapnell et al. [2010] and are summarised below.

In the experimental protocol for RNA-seq, RNA is extracted, fragmented, reverse transcribed, PCR amplified, and a library constructed for sequencing by NGS. During the library construction process, platform specific sequences such as adapters and PCR primers are ligated to the fragments. RNA-seq data are then interpreted by one of three methods. The first is to map the reads to a reference genome. This creates a pile-up in which the number of reads that map to a gene are counted and the proper statistics used to calculate the abundances of RNA originating from each annotated gene in a sample (discussed below). These abundances are then compared between two or more samples to calculate fold changes and p-values for differential expression. The second approach is to first assemble RNA transcripts *de novo*, independent from the genome, and then use this set of transcripts as the reference to which reads would be mapped quantitatively to calculate differential expression. This is the only method that can be performed if a genome is not available, typical of species with very large genomes such as plants and other higher eukaryotes. This method also reveals alternative splice forms (splice variants) of transcripts. The third approach combines the steps from the first and second approaches. Reads are mapped first to a genome in which putative splice junctions have been annotated, which are the exon-intron boundaries of eukaryotes. Alternative splice forms of RNAs and their abundances are then simultaneously resolved in a probabilistic method lines [Trapnell et al. 2010].

The benefits and constraints of the different methods of RNA-seq data analysis have been reviewed by Martin and Wang [2011]. The first method is the simplest and most suitable for bacteria, archaea (which have no introns), as well as for lower eukaryotes such as yeasts, in which only a small number of genes have introns. Since data are mapped to a reference genome, the maximal number of reads can be utilised for differential expression testing, rendering the method very sensitive. The second and third methods are required for high eukaryotes and have the benefit of elucidating splice variants and their abundances in different cell lines, under different conditions, or in diseased versus healthy states such as cancer. The third method is more sensitive than the second *de novo* assembly method, since the maximum number of reads can be used due to the availability of the reference genome, but the accuracy depends on the accuracy of the genome annotation [Martin and Wang 2011]. The latter approach has, in a single experiment, revealed thousands of novel transcripts in mammalian cell lines [Trapnell et al. 2010], demonstrating the potential of RNA-seq to transform the functional genomics landscape.

Theoretically, RNA-seq can detect the presence of a single transcript and is thus similar in its sensitivity to quantitative real-time PCR and is superior to microarray analysis. As there is no theoretical upper limit, since the detector cannot be saturated, the dynamic range is also superior to microarrays. The transcript abundances for paired-end data are reported as FPKM, which is an abbreviation for fragments per kilobases of a gene, per million reads sequenced [Trapnell et al. 2010]. This unit implies that two normalisations have been performed on the data. Since two samples would not contain exactly the same amount of RNA due to differences in extraction efficiency, the abundances are normalised by the total numbers of reads sequenced (per million reads), which circumvents the need for diluting or concentrating a sample. Also, longer genes would potentially have more reads mapping to them compared to shorter genes; hence the number of reads mapping to a gene is normalised by the length of the gene (per kilobase of nucleotides in the gene). RNA-seq is also very flexible in that mixed cultures can be studied without interference, assuming that the reads could be sufficiently mapped to the genome of either species, which would work best if complete genomes of both (or multiple species) are available and the species are not very closely related. There is also no need for investment into the development of a microarray and it is accordingly very suitable for projects on non-model species. Hence, several aspects make RNA-seq superior to microarray analysis, although it is more expensive. [Trapnell et al. 2010]

Several software programmes for the raw data processing of RNA-seq data have become available. TopHat [Kim et al. 2013] is an excellent read-mapping tool for reads, which employs Bowtie [Langmead et al. 2009, Langmead and Salzberg 2012], allowing it to span splice junctions. CuffLinks is used for resolving splice variants using a reference genome [Trapnell et al. 2013], while CuffDiff is used for calculating transcript abundances and differential expression among samples [Trapnell et al. 2010, Trapnell et al. 2013]. Galaxy is an excellent online resource for NGS data analysis and provides an easy entry into NGS raw data analysis [Afgan et al. 2015]. Improvements could, however, still be made to RNA-seq experimental and data analytical protocols, since different algorithms provide somewhat different results [Su et al. 2014].

Methods to study metabolism

Metabolism is central to cellular physiology and hence a focus both in biotechnology research and in medical science. Elucidation of metabolic pathways pre-dated the study of signalling pathways and genetic regulation [Voet and Voet 2011]. Like all systems in biology, its regulation is complex and non-linear. Mathematical modelling and computer simulation can provide a deeper understanding of regulation and control of these processes as opposed to experimentation alone. Models of metabolism can elucidate the rate-limiting steps in a metabolic pathway, which evidently is useful from a biotechnology perspective for the purpose of metabolic engineering [Kacsar and Burns 1973]. Another is the exploration of the metabolic capacity under a variety of conditions, which might not feasibly be tested experimentally [Famili et al. 2003, Schilling et al. 1999]. Also, detailed models of metabolism may be used to predict concentrations or reaction rates that are not readily measurable with existing technologies at the time. Detailed models of metabolism became popular early on. This line of research was greatly supported by the strong emphasis on enzyme kinetic studies which lead to elaborate rate equations for a variety of mechanisms, like the well-known Michaelis-Menten [Michaelis and Menten 1913] and Hill models [Hill 1910]. These are readily built into a set of differential equations and a computer simulation is then used to solve the set of equations over the time variable, often leading to a steady state of fluxes and concentrations [Garfinkel et al. 1970, Heinrich and Schuster 1998]. Some of the most elaborate mechanistic models of metabolism with relevance to biotechnology is that of the glycolytic and fermentative pathways in the yeast *S. cerevisiae* [Teusink et al. 2000]. A key benefit of using a mechanistic, bottom-up approach is that the predictions have a direct relationship to parameters that can be manipulated by the experimenter. For instance, the maximal activities in enzyme kinetics relate to protein concentrations, which may be manipulated by over-expression or knock-out of genes. Another benefit of the extraordinary detail of

such models is that these might be thought of as condition-independent. Since the central metabolites are mostly treated as variables and not constants, the simulations should still be relevant, and hence making the model extrapolatable to other conditions. In most cases, however, these of models based on detailed kinetics do not include any signalling or genetic regulation, rendering them again dependent on the context under which they were constructed for, unless enzyme activities are measured for each condition.

A strong foundation for the understanding of the regulation of metabolism is Metabolic Control Analysis (MCA), which has been developed by a number of contributors over several decades [Kacsar and Burns 1973, Heinrich and Rapoport 1974, Fell and Sauro 1985, Reder 1988, Conradie et al. 2006]. The mathematical formulation can be used to calculate control coefficients that describe the contribution of an enzyme to the control of the flux through any reaction in the system, including the reaction catalysed by its own activity. To make a simplification, the flux control coefficient describes the fractional change in the flux that could be expected by a one percent change in the concentration of that enzyme. The higher the flux control coefficient, the more responsive the flux should be to a change in the concentration of that enzyme. This theoretical framework has also been extended to various other forms, such as supply-demand analysis, which groups several enzymes into reaction blocks to study the control by the reaction block via a single linking intermediate [Kroukamp 2003], or any number of intermediates [Rohwer and Hofmeyr 2008], on any reaction or reaction block. It is very attractive to perform MCA for the purpose of metabolic engineering, as it predicts which enzymes to focus on for genetic engineering. To unlock the potential of MCA, at least two main approaches exist. The first, more traditional approach, is to characterise the individual enzymes in the pathway by enzyme kinetic assays, and make simplifications where needed. Good examples of such studies were performed by Teusink et al. [2000], who created a detailed model of fermentative glycolysis in *S. cerevisiae*. The MCA formulation calculates the elasticity coefficients (similar to reaction order, or percentage-wise slopes on a graph of rate versus concentration) at the simulated concentrations and fluxes close to the reference experimental condition, and then calculates the flux control coefficients. Another approach requires the measurement of both fluxes and concentrations of key metabolic intermediates [Hofmeyr and Cornish-Bowden 1996].

Unfortunately, applicability of MCA by both approaches is limited by our ability to perform realistic experimental measurements. For the detailed enzyme characterisation approach, the labour intensiveness of detailed enzyme kinetics, in addition to possible experimental artefacts introduced by isolation of the enzymes, limit the approach to a few pathways of key interest. For the experimental

MCA approach, our ability to perform accurate, quantitative metabolomics measurements is a serious constraint. It is, however, attractive to measure changes in intracellular metabolite levels over time after a sudden perturbation of metabolism, such as a pulse of glucose, and fit enzyme kinetics [Visser et al. 2002]. This could be termed *in vivo* kinetics. A variety of mass spectrometry based methods such as GC-MS currently exist that can detect hundreds or even thousands of small molecules in a biological sample, but the quantitation accuracy is much worse than required to perform accurate quantitative experimental MCA throughout the cell, or to accurately fit enzyme kinetics at a series of steady states, or after a substrate pulse. LC-MS is more quantitative, since no chemical derivatization is required. However, for the sensitive measurement mode of multiple reaction monitoring (MRM), a chemical standard is required for each compound of interest. For the majority of metabolites, the standards are not available and those that are, are usually very expensive. From own experience, a tremendous amount of effort is required in performing a pulse experiment involving sampling and sample preparation (the invasive method), including rapid sampling [Visser et al. 2002]. For rapidly growing microorganisms, samples ideally need to be quenched within a fraction of a second to stop metabolism and minimise changes in the intracellular metabolite levels, which are extremely low in most cases. This requires specialised equipment [Visser et al. 2002] which currently is not widely available. By using *in vivo* nuclear magnetic resonance spectroscopy (NMR), the need for rapid sampling could be circumvented [Gillies et al. 1981, Crous 2011]. An additional improvement is to use NMR to measure the metabolite and cofactor levels using permeabilised cells [Smith 2010]. Since the cells are non-growing, longer acquisition times could be used, increasing the sensitivity of NMR detection. Fitting *in vivo* kinetics has the advantage over experimental MCA in that the mechanistic details are captured, which gives context independence to the model, allowing predictions of control to be make for conditions that may differ from the experimental setup used for fitting [Smith 2010].

Detailed models of metabolism usually capture only a very small fraction of the reactions and molecules in the cell. Omics technologies on the other hand, capture a much larger fraction, and thus there has been a wide gap between these two sides of systems biology. Metabolic flux analysis at the steady state is a powerful set of tools to elucidate the cellular state, at least of central carbon metabolism (see a discussion of methods below). Schabort [2007] combined the data from ^{13}C-Metabolic Flux Analysis (^{13}C-MFA) with a detailed kinetic model of fermentative glycolysis in yeast [Teusink et al. 2000]. By including the effects of the larger flux network as hard constraints in a detailed kinetic model, the neglected reactions could be filled in for the purpose of rigorous model validation at a condition that was different to that for which the model was constructed. In this "^{13}C-constrained kinetic modelling" method, the simulated fluxes were forced to the measured values, revealing the

individual enzymes for which enzyme activities or other parameters were incorrect. This is a complementary approach to *in vivo* kinetics measuring intracellular metabolites.

A different approach to the study of metabolism was recently demonstrated by Canelas et al. [2011] for *S. cerevisiae*. The authors focused only on glucose as the carbon source and varied the dilution rate in a chemostat culture, measuring the intracellular metabolite concentrations and the fluxes throughout central carbon metabolism, and calculated the disequilibrium ratio as a function of intracellular flux. The idea was to classify reactions into those that function close to equilibrium or far from equilibrium. This knowledge can be very useful, since reactions that function close to equilibrium would always simply translate the signals from changes in their own substrate and product concentrations into changes in the flux. They would also have very little control over the metabolic flux, as is true for reactions that function close to equilibrium, which usually have high maximal activities compared to the neighbouring reactions. Moreover, it results in a vast simplification of cellular modelling, as the majority of reactions could be modelled using only equilibrium constants, which are generally applicable to all species. On the other end of the spectrum, the enzymes that function far from equilibrium are the ones that require detailed characterisation with enzyme kinetics, as well as measurements of the enzyme activity levels [Canelas et al. 2011]. An in-between class contain those reactions that showed a linear response of the flux with the disequilibrium ratio. These reactions need to be modelled using both the equilibrium constants and the empirically determined slope, which captures the combined effects of protein expression changes (enzyme activity) and the kinetic effects of metabolite level changes on the enzyme. It is a good example of a reduction strategy that has potential for genome scale modelling, complementary to *in vivo* kinetics and ^{13}C-constrained kinetic modelling.

Rapid equilibrium or near equilibrium enzymes should not require gene regulation, since their activity levels would typically be very high compared to the neighbouring enzymes and only very large changes would result in a change in flux (in the case when they are substantially down-regulated, and then become a rate-limiting enzyme). The classification of some of these reactions in *S. cerevisiae* were put to the test in Chapter 3, using a different species (*K. marxianus*), and asking whether those classified as rapid equilibrium enzymes would indeed be constitutively expressed between conditions in which the fluxes through them would change. These different conditions were imposed by using glucose or xylose as the carbon source. It is expected that the genes of lower glycolysis and the translaldolase and transketolase reactions of the non-oxidative pentose phosphate pathway would be constitutively expressed, while the fluxes would be affected by the two different carbon sources.

Very recently, a novel approach was developed by the Rabinowitz group [Hackett et al 2016], which aims at elucidation of *in vivo* kinetics and making extensive use of chemostats for yeast, like that done by Canelas et al. [2011], but making use of proposed Michaelis-Menten type rate equations in fitting procedures. Both the fluxes and metabolite concentrations are measured under multiple conditions, and the rate equations are tested as the causal links between the observed changes in metabolite levels and fluxes. The use of multiple steady states and especially the accurate measurement of fluxes are important for this combined data-driven and hypothesis driven approach. The combination of fluxes and proteome data sets by the Milo group has also recently suggested relationships between enzyme kinetics and the circuitry design of pathways at the network level [Barenholz et al. 2017]. The measurement, or rather, the calculation of metabolic fluxes evidently is becoming increasingly important in systems biology and is a non-trivial task.

Metabolic Flux Analysis and related methods

The metabolic reaction rate at a steady state is called the metabolic flux. A group of related methods calculates or predicts fluxes, depending on what data are available. Methods have been thoroughly reviewed by Wiechert [2001] and Schilling et al. [1999], and will only be described below as relevant to the application in this thesis. Metabolic Flux Analysis (MFA) usually refers to using the metabolite balance model as a measurement tool for calculating fluxes, providing a snapshot of metabolism at a given condition. Flux Balance Analysis (FBA) uses the flux model in a predictive sense, and usually as a function of various simulated growth conditions. This line of research on genome-scale FBA was pioneered by the Palsson group [Famili et al. 2003, Schilling et al. 1999]. The basis for most of these methods is the stoichiometric matrix, *S*, which describes the metabolite balances in differential equations captured in matrix form. The change in concentrations over time is formulated as

$$\frac{dx}{dt} = S.v$$

The stoichiometric matrix *S* has *m* rows describing the metabolite balances and *n* columns, representing the *n* reactions, and v is the vector containing *n* fluxes. In MFA and FBA, *S* is separated into two matrices. The first describes the intracellular fluxes, while the second is an identity matrix that maps those metabolites that can enter or leave the system via the exchange reactions [Schilling et al. 1999]. The formulation is described in Chapter 9. FBA uses linear optimisation to calculate a single flux distribution (pattern) in the convex space defined by the null space of flux constraints defined by *S* [Schilling et al. 1999]. As objective function for linear optimisation, the maximised growth

rate is often used, which proposes that the flux distribution in an organism would be such that it captures as much as possible of the nutrients in the form of biomass such that the *in silico* organism would multiply as rapidly as possible. This makes intuitive sense for microorganisms such as bacteria, which likely evolved to compete based on growth rate [Schuetz et al. 2007]. Yeasts that are Crabtree negative would also follow this pattern, since wasteful ethanol production is not prevalent in these yeasts, including *K. marxianus*.

^{13}C-Metabolic Flux Analysis (^{13}C-MFA), like MFA, uses the metabolite balance model to calculate fluxes, but uses also isotope labelling patterns to impose further constraints on the intracellular flux distribution. Two main approaches exist. The first is ^{13}C-Constrained Metabolic Flux Analysis, in which the additional constraints imposed by isotope labelling patterns from gas chromatography coupled to mass spectrometry (GC-MS) or from nuclear magnetic resonance (NMR) are captured in the form of flux ratios and combined with the metabolite balance model [Fischer et al. 2004, Zamboni et al. 2005]. The second approach is a simulation approach in which a balance model is set up for all possible labelling patterns (isotopomers), and the simulation outputs compared with measured isotope labelling patterns [Wiechert 2001]. Whatever the method of calculation, the application of ^{13}C-MFA is limited by identifiability, which is the theoretical possibility to calculate the set of fluxes, given the labelling pattern of the substrate(s), and by the availability of such a labelled substrate. Often, ^{13}C-MFA is not applicable for calculating all fluxes for growth from certain substrates, or the most suitable labelling pattern may not be available or may be prohibitively expensive. For example, the complete set of fluxes in central metabolism in *K. marxianus* would not be identically identifiable on glucose or xylose as growth substrates, and suitable ^{13}C-labelled xylose is very expensive. Hence, in Chapter 3, fluxes were estimated by FBA.

Metabolic Regulation Analysis

MCA can be used to calculate explicitly the hierarchical effects of kinases and gene expression on the metabolic level. However, the kinetic parameters and binding constants for interactions outside of metabolism are rarely measured or measurable at this point in time. A sensible alternative to attempting to answer the question of *which components have the most control* over a pathway, is to ask *which components most likely were the causative agents leading to a change in the flux of interest*, considering their observed differential expression. Metabolic Regulation Analysis (MRA) was developed to quantitatively separate the metabolic and hierarchical levels of regulation [Ter Kuile and Westerhoff 2001]. The metabolic level of regulation involves the changes in metabolite concentrations

between two conditions, which is the effect of changes in the environment, transduced as a signal through the pathways. It is logical that an increase in a concentration of the substrate of a reaction would lead to an increase in the flux, while the opposite is true for the product. The hierarchical level of regulation describes the effect that a change in the activity of an enzyme would have on the flux through that reaction. Activity of an enzyme is usually directly related to the concentration of an enzyme, and may also be affected by post-translational modifications. While metabolite levels are very challenging to measure, enzyme activity is simpler to measure accurately, as is the concentration of protein levels. In fact, only the relative fold changes of activities or protein levels need to be measured in order to determine the hierarchical component, while the metabolic component is calculated from the hierarchical component, using the theorem for MRA [Ter Kuile and Westerhoff 2001]. This formulation is described in Chapter 3. Notably, the MRA approach could potentially use the genome-wide data generation methods from proteomics and transcriptomics, effectively extending functional genomics to a quantitative description of metabolic regulation. In Chapter 3 the idea is extended to the case of RNA-seq, which may gain in popularity since it is a practical approach to genome-wide studies in regulation, especially if flux analytical methods such as ^{13}C-MFA or FBA would be extended to the larger scale. Indeed, the Nielsen group very recently elucidated that, although there is almost no correlation between mRNA levels and their protein levels, reflecting differences in translation efficiency, there was a strong correlation when considering their differential expression values [Lahtvee et al 2017]. Hence, this new finding supports the use of differential RNA-seq as a practical and cost-effective replacement for differential proteomics.

Alternative roles for fructose-1,6-bisphosphatase

Derepression of genes involved in the utilisation of alternative carbon sources, such as alternative sugar transporters, could be considered as a strategy to both “sense” the presence of supplementary nutrients in the absence of the preferred hexoses, and to enable the capacity for their utilisation, in case they were present. Such a strategy would prepare a cell for an alternative lifestyle, although the gene products may not actually be involved in any metabolic activity if the relevant nutrients were not present. Of particular interest is fructose-1,6-bisphosphatase (FBP), and the FBP1 gene is under the control of Mig1 and glucose repression in *S. cerevisiae* [Klein et al. 1998]. However, its centrality to energy metabolism distinguishes it from other genes under glucose repression. If FBP is expressed and active, it might have a profound effect on central metabolism, whereas transporters and β-oxidation could be considered as peripheral to central metabolism and hence would have no effect if the relevant nutrients were not present. Hence, the up-regulation of the FBP1 gene might not be for the

same purpose as alternative transporters and peripheral pathways. The question arises whether there might be other reasons for its up-regulation in the absence of glucose, as is the case in *S. cerevisiae*. The potential roles of FBP are reviewed below.

Fructose-1,6-bisphosphatase, gluconeogenesis and NADPH

The phosphofructokinase (PFK) reaction of glycolysis is a highly irreversible reaction *in vivo*, therefore a bypass is required during gluconeogenesis, which must have a different chemistry to allow operation in the different direction [Voet and Voet 2011]. This reaction is catalysed by FBP and, accordingly, FBP activity is usually associated with gluconeogenesis. Previously, a microarray study of the transcriptomic response of a recombinant *S. cerevisiae* strain that could utilise xylose indicated a four-fold increase in the FBP1 transcript levels in a xylose medium [Runquist et al. 2009]. The authors concluded that, due to the additional requirement for NAPDH for xylose utilisation, the oxidative pentose phosphate pathway flux had to increase, which would also require a reverse 'gluconeogenic' flux in upper glycolysis, requiring the FBP reaction. This latter hypothesis might be relevant also to *K. marxianus* when cultivated in a xylose medium, since additional NADPH is required for xylose utilisation by xylose reductase. However, the flux constraints and the interconnected nature of metabolism through several cofactors poses the question of whether the proposed purpose as an NADPH generating mechanism through some cyclic form of PPP flux is indeed possible.

Fructose-1,6-bisphosphatase, ATP substrate cycles and cancer

A simultaneous flux in both PFK and FBP would result in a substrate cycle that would hydrolyse and thus waste ATP. Thus it was assumed in early studies that the two enzymes functioned under separate conditions, especially since it was found that the allosteric activators of PFK, AMP and fructose-2,6-bisphosphate, also inhibited FBP [Voet and Voet 2011]. These substrate cycles, termed 'futile cycles', were presumed to be avoided. It is, therefore, a logical assumption that gene regulation may also follow the same pattern by up-regulating FBP only during gluconeogenesis. Subsequently, however, it was found that both these reactions functioned simultaneously in at least mammalian muscle cells [Voet and Voet 2011, Newsholme et al. 1984]. It was also demonstrated that this substrate cycle actually had a regulatory purpose in the mammalian muscle cell working in a fermentative mode, rendering the net glycolytic flux highly responsive to a sudden decrease in ATP concentration during muscle contraction (Figure 4) [Voet and Voet 2011, Newsholme et al. 1984]. Adenylate kinase responds to a sudden increase in ADP via rapid equilibrium, increasing the level of AMP four-fold from its originally very low concentration during the dynamic response, which acts as a signal amplifier for a small decrease in ATP. AMP then stimulates PFK and inhibits FBP allosterically, thereby effectively

increasing the dynamic range of the response. Fermentative glycolytic flux in mammalian muscles can thus be regulated up to 89-fold from a resting state, during which the flux through FBP is significant and approaches that of PFK, to an active state in which the FBP flux is nine-fold lower and the PFK flux nine-fold higher than in the resting state. Such allosteric regulation does not require genetic regulation, but rather the proteins are constitutively expressed and associated with a short-term dynamic response.

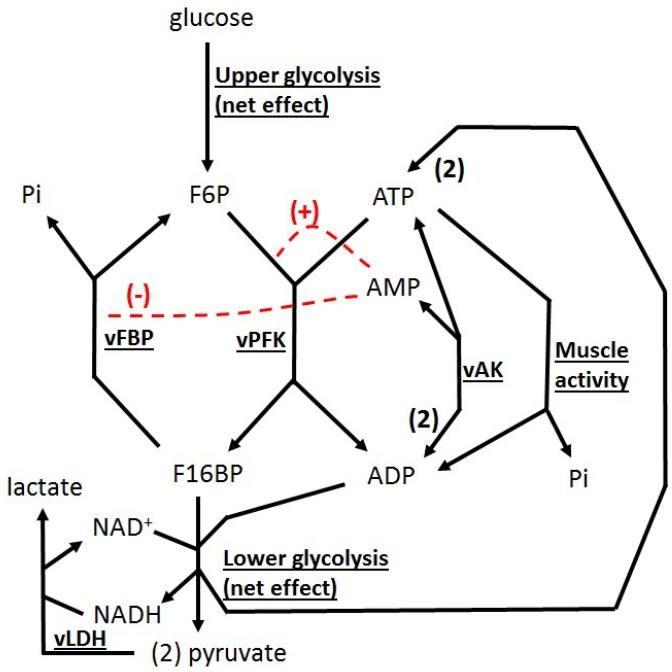


Figure 4. Fermentative metabolism in mammalian muscle cells. Adenylate kinase acts as a rapid equilibrium enzyme that transduces the increase in ADP after muscle contraction into an increase in AMP in a dynamic response, stimulating PFK and inhibiting FBP, resulting in a drastic increase in net glycolysis flux to restore homeostasis of ATP concentration. (Diagram based on data from Voet and Voet [2011]).

Another interesting example is found in the flight muscles of bumblebees, in which PFK and FBP are highly expressed and at similar activity, resulting in a highly active substrate cycle that may contribute to generate heat for take-off during low temperatures [Newsholme 1972]. This heating effect is termed “non-shivering thermogenesis” through substrate cycling which hydrolyses ATP.

Very recently it was discovered that in every one of more than six hundred clear cell renal cell carcinoma tumours analysed in humans, a decrease in the expression of the FBP1 gene was evident [Li et al. 2014]. The dis-regulation of ATP homeostasis clearly may have a dramatic effect on overall

cellular physiology. Tumours, including clear cell renal cell carcinoma, display the Warburg effect in which rapid glycolytic flux occurs with concomitant lactate production [Crabtree 1929], strongly resembling the Crabtree effect in yeast.

Avoidance of ATP production by uncoupling mechanisms

Another type of non-shivering thermogenesis is present in mammals, including humans, in which brown fat (the brown colour originating from the many mitochondria) catabolises acetyl CoA from lipids to generate heat [Voet and Voet 2011]. As the TCA cycle produces NADH and electron transport pumps protons into the mitochondrial intermembrane space, the increasing electro-osmotic potential across the inner mitochondrial membrane would lead to excessive ATP production via proton channelling through F_1F_0 ATPase if it were not for the presence of a natural uncoupling mechanism. This function is performed through thermogenin (UCP1), a transmembrane channel for protons in the inner mitochondrial membrane that channels protons without generating ATP [Jarmuszkiewics 2001]. Uncoupling is under control of the hormones released by the thyroid gland. It is reasonable to believe that only animals would have such internal heating mechanisms, but even plants were found to possess uncoupling proteins such as the plant uncoupling mitochondrial protein (PUMP) in potatoes [Jarmuszkiewics 2001] and in fruits which undergo a respiratory burst in the ripening process [Jezek et al. 1998]. It was later discovered that as much as five uncoupling protein genes exist in mammals and that they are tissue-specific, and therefore could have various physiological roles [Luévano-Martínez 2012]. In the amoeboid protozoan *Acanthamoeba castellanii* an elusive uncoupling activity was discovered that responded in the same manner as thermogenin to fatty acids and nucleotides, and was even up-regulated by cold stress [Jarmuszkiewicz 1999].

These findings suggested that uncoupling proteins are conserved throughout the domains of life. In the yeast *Yarrowia lipolytica* a mitochondrial carrier for oxaloacetate and sulphate was found that resembled thermogenin in its proton transporting function [Luévano-Martínez 2010]. Some authors also reported proton translocation by metabolite transporters [Jarmuszkiewicz 2000]. However, to date no proteins have been isolated in yeasts that seem to function as general regulated uncoupling mechanisms similar to uncoupling proteins of mammals, a protein family that seems to be absent in the fungi [Luévano-Martínez 2010].

The P/O ratio (the number of moles of ATP produced per NADH molecule, or per half a molecule of O_2) in most eukaryotes is about 2.5. Early estimates of the P/O ratio in *S. cerevisiae* grown on glucose gave a surprisingly low value of about 1.0 [Verduyn et al. 1990, Sheldon 1996]. This implied that more

than half of the potential energy was dissipated by some unknown mechanism. Later it was discovered that the theoretical or mechanistic P/O ratio of *S. cerevisiae* was only 1.5, since the NADH dehydrogenase Ndi1p does not translocate protons [Bakker et al. 2001, Famili et al. 2003]. Other yeasts have a theoretical P/O ratio closer to 2.5.

In summary, FBP could potentially play a role in the NADPH balance, although flux constraints in the PPP and glycolysis might render this a mechanism to be condition specific. Further, simultaneous FBP and PFK activity could directly impact on the ATP balance. As such, it might serve as an ATP avoidance mechanism. In yeasts, which seemingly lack dedicated uncoupling proteins, this might be an alternative to the uncoupling proteins found in mammals and plants. The complex interplay between cofactor balances, energy generating pathways and the PPP is explored in Chapter 9, where the potential role of FBP is specifically explored in the simulation framework of FBA. Also in Chapter 9, a strategy for metabolic engineering of a future xylose fermenting strain of *K. marxianus* is developed.

References

Afgan E, Sloggett C, Goonasekera N, Makunin I, Benson D, Crowe M, Gladman S, Kowsar Y, Pheasant M, Horst R, Lonie A. Genomics Virtual Laboratory: a practical bioinformatics workbench for the cloud. PLoS One. 2015;10(10): e0140829. doi: 10.1371/journal.pone.0140829.

Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J. The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. PLoS Comput Biol. 2013;9(3): e1002980.

Akanni GB, du Preez JC, Steyn L, Kilian SG. Protein enrichment of an *Opuntia ficus-indica* cladode hydrolysate by cultivation of *Candida utilis* and *Kluyveromyces marxianus*. J Sci Food Agric. 2015;95(5): 1094–1102.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry M, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1): 25-29.

Bajpai P, Margaritis A. Ethanol inhibition kinetics of *Kluyveromyces marxianus* grown on Jerusalem artichoke juice. Appl Environ Microbiol. 1982;44(6): 1325-1329.

Bakker BM, Overkamp KM, van Maris AJA, Kötter P, Luttik MAH, van Dijken JP, Pronk JT. Stoichiometry and compartmentation of NADH metabolism in *Saccharomyces cerevisiae*. FEMS Microbiol Rev. 2001. 25: 15–37.

Barenholz U, Davidi D, Reznik E, Bar-On Y, Antonovsky N, Noor E, Milo R. Design principles of autocatalytic cycles constrain enzyme kinetics and force low substrate saturation at flux branch points. eLIFE. 2017. doi.org/10.7554/eLife.20667.001.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218): 53–59.

Blank LM, Lehmbeck F, Sauer U. Metabolic-flux and network analysis in fourteen hemiascomycetous yeasts. FEMS Yeast Res. 2005;5: 545–558.

Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: How to use the entry view. Methods Mol Biol. 2016;1374: 23-54.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. Nat Genet. 2001;29(4): 365-71.

Broach JR. Nutritional control of growth and development in yeast. Genetics. 2012;192: 73-105.

Canelas AB, Ras C, Pierick A, van Gulik WM, Heijnen JJ. An *in vivo* data-driven framework for classification and quantification of enzyme kinetics and determination of apparent thermodynamic data. Metab Eng. 2011;13: 294-306.

Cherry JR, Johnson TR, Dollard C, Shuster JR, Denis CL. Cyclic AMP-dependent protein kinase phosphorylates and inactivates the yeast transcriptional activator ADR1. Cell. 1989;56(3): 409-419.

Collins SR, Kemmeren P, Zhao X, Greenblatt JF, Spencer F, Holstege FCP, Weissman JS, Krogan NJ. Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. Mol Cell Proteomics. 2007;6: 439-450.

Compeau PEC, Pevsner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Comput Biol. 2011;29(11): 987-991.

Conradie R, Westerhoff HV, Rohwer JM, Hofmeyr JH, Snoep JL. Summation theorems for flux and concentration control coefficients of dynamic systems. Syst Biol. 2006;153(5): 314-317.

Crabtree, H. G. Observations on the carbohydrate metabolism of tumours. Biochem J. 1929;23: 536–545.

Crous C. Supply-demand analysis of anaerobic free-energy metabolism in *Zymomonas mobilis*. MSc thesis. Stellenbosch University, South Africa. 2011.

Day-Richter J, Harris MA, Haendel M. OBO-Edit—an ontology editor for biologists. Bioinformatics. 2007;23(16): 2198-2200.

De Deken. The Crabtree effect: a regulatory system in yeast. J Gen Microbiol. 1966;44: 149-156.

Dombek KM, Young ET. Cyclic AMP-dependent protein kinase inhibits ADH2 expression in part by decreasing expression of the transcription factor gene ADR1. Mol Cell Biol. 1997;17(3): 1450-1458.

Eldar A, Elowitz MB. Functional roles for noise in genetic circuits. Nature. 2010;467: 167–173.

Famili I, Forster J, Nielsen J, Palsson BO. *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. Proc Natl Acad Sci USA. 2003;100(23): 13134-13139.

Fell DA, Sauro HM. Metabolic control and its analysis: additional relationships between elasticities and control coefficients. Eur J Biochem. 1985;148: 555-561.

Fields S, Song O. A novel genetic system to detect protein–protein interactions. Nature. 1989. 340(6230): 245–246.

Fischer E, Zamboni N, Sauer U. High-throughput metabolic flux analysis based on gas chromatography–mass spectrometry derived ^{13}C constraints. Anal Biochem. 2004;325: 308–316.

Fonseca GG, Heinzle E, Wittmann C, Gombert AK. The yeast *Kluyveromyces marxianus* and its biotechnological potential. Appl Microbiol Biotechnol. 2008;79: 339-354.

Friedman N, Koller D. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. Machine Learning. 2003;50(1): 95–125.

Garfinkel D, Garfinkel L, Pring M, Green SB, Chance B. Computer applications to biochemical kinetics. Annu. Rev. Biochem. 1970;39: 473–498.

Gavin A, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon A, Cruciat C, Remor M, Höfert C, Schelder M, Brajenovic M, Ruffner H, Merino M, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley R, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. 2002;415: 141-147.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10): R80.

Gillies RJ, Ugubril K, Den Hollander JA, Shulman RG. ^{31}P NMR studies of intracellular pH and phosphate metabolism during cell division cycle of *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA. 1981;78(4): 2125-2129.

Groeneveld P, Stouthamer AH, Westerhoff HV. Super life – how and why 'cell selection' leads to the fastest-growing eukaryote. FEBS J. 2009;276: 254–270.

Hahn-Hägerdal B, Karhumaa K, Fonseca C, Spencer-Martins I, Gorwa-Grausland MF. Towards industrial pentose-fermenting yeast strains. Appl Microbiol Biotechnol. 2007;74: 937–953.

Hackett SR, Zanotelli VRT, Xu W, Goya J, Park JO, Perlman DH, Gibney PA, Botstein D, Storey JD, Rabinowitz JD. Systems-level analysis of mechanisms regulating yeast metabolic flux. Science. 2016;354(6311): aaf2786.

Hannenhalli S. Eukaryotic transcription factor binding sites — modeling and integrative search methods. Bioinformatics. 2008;24(11): 1325–1331.

Hardie DG, Carling D, Carlson M. The AMP-activated/SNF1 protein kinase subfamily: metabolic sensors of the eukaryotic cell. Annu Rev Biochem. 1998;67: 821–855.

Hardie DG, Carling D, Gamblin SJ. AMP-activated protein kinase: also regulated by ADP. Trends Biochem Sci. 2011;36: 470–477.

Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning. 1995;20(3): 197–243.

Heinrich R, Rapoport TA. A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. Eur J Biochem. 1974;42: 89-95.

Heinrich R, Schuster S. The modelling of metabolic systems. Structure, control and optimality. BioSystems. 1998;47: 61–77.

Hill AV. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. J. Physiol. 1910. 40(Suppl): iv–vii. doi:10.1113/jphysiol.1910.sp001386.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature. 2002;415, 180–183.

Hofmeyr JS, Cornish-Bowden A. Co-response analysis: a new experimental strategy for metabolic control analysis. J Theor Biol. 1996;182: 371-380.

Hong SP, Carlson M. Regulation of snf1 protein kinase in response to environmental stress. J Biol Chem. 2007;282: 16838–16845.

Hong J, Wang Y, Kumagai H, Tamaki H. Construction of thermotolerant yeast expressing thermostable cellulase genes. J. Biotechnol. 2007;130: 114–123.

Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signaling circuits in molecular interaction networks. Bioinformatics. 2002;18(1): 233-240.

Isci S, Ozturk C, Jones, J, Out HH. Pathway analysis of high-throughput biological data within a Bayesian network framework. Bioinformatics. 2011;27(12): 1667–1674.

Jarmuszkiewicz W. Uncoupling proteins in mitochondria of plants and some microorganisms. Acta Biochim Pol (Engl Transl). 2001;48: 145-155.

Jarmuszkiewicz W, Milani G, Fortes F, Schreiber AZ, Sluse FE, Vercesi AE. First evidence and characterization of an uncoupling protein in fungi kingdom: CpUCP of *Candida parapsilosis*. FEBS Lett. 2000;467: 145-149.

Jarmuszkiewicz W, Sluse-Goffart, CM, Hryniewiecka L, Sluse FE. Identification and characterization of a protozoan uncoupling protein in *Acanthamoeba castellanii*. J Biol Chem. 1999;274: 23198–23202.

Jezek P, Engstová H, Saèková M, Vercesi AE, Costa DT, Arruda P, Garlid P. Fatty acid cycling mechanism and mitochondrial uncoupling proteins. Biochim Biophys Acta. 1998;1365: 319–327.

Kacser H, Burns JA. The control of flux. Symp Soc Expr Biol. 1973;27: 65-104.

Kallio MA, Tuimala JT, Hupponen T, Klemelä P, Gentile M, Scheinin I. Chipster: user-friendly analysis software for microarray and other high-throughput data. BMC Genomics. 2011;12: 507.

Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, Kaipa P, Gilham F, Spaulding A, Popescu L, Altman T, Paulsen I, Keseler IM, Caspi R. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Briefings Bioinf. 2009;11(1): 40-79. doi:10.1093/bib/bbp043.

Karpichev IV, Small GM. Global regulatory functions of Oaf1p and Pip2p (Oaf2p), transcription factors that regulate genes encoding peroxisomal proteins in *Saccharomyces cerevisiae*. Mol Cell Biol. 1998;18(11): 6560–6570.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002;12(6): 996-1006.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg S. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4): R36. doi:10.1186/gb-2013-14-4-r36.

Klein CJ, Olsson L, Nielsen J. Glucose control in *Saccharomyces cerevisiae*: the role of Mig1 in metabolic functions. Microbiology. 1998;144(1): 13-24.

Kroukamp M. Supply-demand analysis of free-energy metabolism in *Saccharomyces cerevisiae*. PhD thesis. Stellenbosch University, South Africa. 2003.

Kuloyo, O.O., du Preez, J.C., García-Aparicio, M.P., Kilian, S.G., Steyn, L. & Görgens, J. *Opuntia ficus-indica* cladodes as feedstock for ethanol production by *Kluyveromyces marxianus* and *Saccharomyces cerevisiae*. World J Microbiol Biotechnol. 2014;30(12): 3173-3183.

Kurtzman CP. Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the *Saccharomycetaceae*, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygotorulaspora*. FEMS Yeast Res. 2003;4: 233–245.

Kurtzman CP, Robnett CJ. Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. Antonie van Leeuwenhoek. 1998;73: 331–371.

Lachance M. Current status of *Kluyveromyces* systematics. FEMS Yeast Res. 2007;7: 642–645.

Lane MM, Morrissey JP. *Kluyveromyces marxianus*: a yeast emerging from its sister's shadow. Fungal Biol Rev. 2010;24: 17–26.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Meth. 2012;9(4): 357–359. doi:10.1038/nmeth.1923.

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3): R25. doi:10.1186/gb-2009-10-3-r25.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science. 1993;262(5131): 208-214.

Lertwattanasakul N, Kosaka T, Hosoyama A, Suzuki Y, Rodrussamee N, Matsutani M, Murata M, Fujimoto M, Suprayogi S, Tsuchikane K, Limtong S, Fujita N, Yamada M. Genetic basis of the highly efficient yeast *Kluyveromyces marxianus*: complete genome sequence and transcriptome analyses. Biotechnol Biofuels. 2015;8(47). doi: 10.1186/s13068-015-0227-x

Li B, Qiu B, Lee DSM, Walton ZE, Ochoki JE, Mathew LK, Mancuso A, Gade TP, Keith B, Nissim L, Simon MC. Fructose-1,6-bisphosphatase opposes renal carcinoma progression. Nature. 2014;514: 251-255.

Llorente B, Malpertuy A, Blandin G, Artiguenave F. Winker P, Dujon B. Genomic exploration of the hemiascomycetous yeasts: 12. *Kluyveromyces marxianus* var. *marxianus*. FEBS Lett. 2000;487: 71-75.

Luévano-Martínez LA. Uncoupling proteins (UCP) in unicellular eukaryotes: true UCPs or UCP1-like acting proteins? FEBS Lett. 2012;586: 1073-1078.

Luévano-Martínez LA, Moyano E, de Lacoba MG, Rial E, Uribe-Carvajal S. Identification of the mitochondrial carrier that provides *Yarrowia lipolytica* with a fatty acid-induced and nucleotide-sensitive uncoupling protein-like activity. Biochim Biophys Acta. 2010;1797: 81–88.

Maciejewski, H. Gene set analysis methods: statistical models and methodological differences. Briefings Bioinf. 2013;15: 504-518. doi:10.1093/bib/bbt002.

Malcher M, Schladebeck S, Mösch H. The Yak1 protein kinase lies at the center of a regulatory cascade affecting adhesive growth and stress resistance in *Saccharomyces cerevisiae*. Genetics. 2011;187(3): 717-30.

Marini AM, Soussi-Boudekou S, Vissers S, Andre B. A family of ammonium transporters in *Saccharomyces cerevisiae*. Mol Cell Biol. 1997;17: 4282–4293.

Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12: 671-682.

Mathelier A, Wasserman WW. The next generation of transcription factor binding site prediction. PLoS Comput Biol. 2013;9(9): e1003214.

Mathelier A, Zhao X, Zhang AW, Percy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H, Lim J, Shyr C, Tan G, Zhou M, Lenhard B, Sandelin A, Wasserman WW. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 2014:42(D1): D142-D147.

Michaelis L, Menten ML. Die kinetik der invertinwirkung. Biochemische Zeitschrift. 1913. 49: 333–369.

Mitchelhill, KI, Stapleton D, Gao G, House C, Michell B, Katsis F, Witters LA, Kemp BE. Mammalian AMP-activated protein kinase shares structural and functional homology with the catalytic domain of yeast Snf1 protein kinase. J Biol Chem. 1994;269: 2361–2364.

Newsholme EA, Challiss RAJ, Crabtree B. Substrate cycles: their role in improving sensitivity in metabolic control. Trends Biochem Sci. 1984;9: 277-280.

Newsholme EA, Crabtree B, Higgins SJ, Thornton SD, Start C. The activities of fructose diphosphatase in flight muscles from the bumble-bee and the role of this enzyme in heat generation. Biochem J. 1972;128: 89-97.

Lahtvee P, Sánchez BJ, Smialowska A, Kasvandik S, Elsemman IE, Gatto F, Nielsen J. Absolute quantification of protein and mRNA abundances demonstrate variability in gene-specific translation efficiency in yeast. Cell Systems. 2017;4: 495-504.

Nonklang S, Abdel-Banat BMA, Cha-aim K, Moonjai N, Hoshida H, Limtong S. High-temperature ethanol fermentation and transformation with linear DNA in the thermotolerant yeast *Kluyveromyces marxianus* DMKU3-1042. Appl Environ Microbiol. 2008;74(24): 7514–7521.

Nonklang S, Ano A, Abdel-Banat BM, Saito Y, Hoshida H, Akada, R. Construction of flocculent *Kluyveromyces marxianus* strains suitable for high-temperature ethanol fermentation. Biosci Biotechnol Biochem. 2009;73: 1090–1095.

Oliveira AP, Patil KR, Nielsen J. Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. BMC Syst. Biol. 2008;2(17). doi:10.1186/1752-0509-2-17.

Orlova M, Kanter E, Krakovich D, Kuchin S. Nitrogen availability and TOR regulate the Snf1 protein kinase in *Saccharomyces cerevisiae*. Eukaryot Cell. 2006;5: 1831–1837.

Patil KR, Nielsen J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc Natl Acad Sci USA. 2005;102(8): 2685-2689.

Pecota DC, Rajgarhia V, Da Silva NA. Sequential gene integration for the engineering of *Kluyveromyces marxianus*. J Biotechnol. 2007;127: 408–416.

Pfromm PH, Amanor-Boadu V, Nelson R, Vadlani P, Madl R. Bio-butanol vs. bio-ethanol: a technical and economic assessment for corn and switchgrass fermented by yeast or *Clostridium acetobutylicum*. Biomass Bioenergy. 2010;34: 515–524.

Poirier Y, Antonenkov VD, Glumoff T, Hiltunen JK. Peroxisomal β-oxidation - a metabolic pathway with multiple functions. Biochim. Biophys. Acta. 2006;1763: 1413–1426.

Postma E, Verduyn C. Enzymic analysis of the Crabtree effect in glucose-limited chemostat cultures of *Saccharomyces cerevisiae*. Appl Environ Microbiol. 1989;55: 468–477.

Ratnakumar S, Young ET. Snf1 dependence of peroxisomal gene expression is mediated by Adr1. J Biol Chem. 2010;285(14): 10703–10714.

Ratushny AV, Ramsey SA, Roda O, Wan Y, Smith JJ. Aitchison JD. Control of transcriptional variability by overlapping feed-forward regulatory motifs. Biophys J. 2008;95: 3715–3723.

Reder C. Metabolic control theory: a structured approach. J Theor Biol. 1988;135: 175-201.

Rohwer JM, Hofmeyr J. Identifying and characterising regulatory metabolites with generalised supply–demand analysis. J Theor Biol. 2008;252: 546-554.

Robertson LS, Fink GR. The three yeast A kinases have specific signaling functions in pseudohyphal growth. Proc Natl Acad Sci USA. 1998;95: 13783–13787.

Rocha SN, Abrahão-Neto J, Gombert AK. Physiological diversity within the *Kluyveromyces marxianus* species. Antonie van Leeuwenhoek. 2011;100: 619–630.

Rosillo-Calle. Food versus fuel: toward a new paradigm - the need for a holistic approach. ISRN Renewable Energy. 2012; 954180.

Runquist D, Hahn-Hägerdal B, Bettiga M. Increased expression of the oxidative pentose phosphate pathway and gluconeogenesis in anaerobically growing xylose utilising *Saccharomyces cerevisiae*. Microb Cell Fact. 2009;8(49). doi: 10.1186/1475-2859-8-49.

Rusk N. Torrents of sequence. Nature Methods. 2011;8: 44. doi: 10.1038/nmeth.f.330.

Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94(3): 441–448.

Schabort DWP. Integration of kinetic models with data from ^{13}C-metabolic flux experiments. MSc thesis. Stellenbosch University, South Africa. 2007.

Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995;270(5235): 467–70.

Schilling CH, Edwards JS, Palsson BO. Toward metabolic phenomics: analysis of genomic data using flux balances. Biotechnol Prog. 1999;15: 288-295.

Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. Mol Syst Biol. 2007;3(119).

Schuller HJ. Transcriptional control of nonfermentative metabolism in the yeast *Saccharomyces cerevisiae*. Curr Genet. 2003;43(3): 139-60.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11): 2498-2504.

Sheldon JG, Williams S, Fulton AM, Brindle KM. ^{31}P NMR magnetization transfer study of the control of ATP turnover in *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA. 1996;93: 6399-6404.

Shirra MK, McCartney RR, Zhang C, Shokat KM, Schmidt MC. A chemical genomics study identifies Snf1 as a repressor of GCN4 Translation. 2008;283(51): 35889–35898.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009;19: 1117-1123.

Simon M, Binder M, Adam G, Hartig A, Ruis H. Control of peroxisome proliferation in *Saccharomyces cerevisiae* by ADR1, SNF1 (CAT1, CCR1) and SNF4 (CAT3). Yeast. 1992;8(4): 303-309.

Sinha S, Tompa M. Discovery of novel transcription factor binding sites by statistical overrepresentation. Nucl Acids Res. 2000;30(24): 5549-5560.

Smith JA. Experimental supply demand analysis of yeast fermentative free energy metabolism: an *in vivo* and *in situ* investigation. MSc thesis. Stellenbosch University, South Africa. 2010.

Smith JJ, Ramsey SA, Marelli M, Marzolf B, Hwang D, Saleem RA. Transcriptional responses to fatty acid are coordinated by combinatorial control. Mol Syst Biol. 2007;3(115). doi:10.1038/msb4100157.

Smith JJ, Miller LR, Kreisberg R, Vazquez L, Wan Y, Aitchison JD. Environment-responsive transcription factors bind subtelomeric elements and regulate gene silencing. Mol Syst Biol. 2011;7(455): 1-15. doi:10.1038/msb.2010.110.

Smits A, Vermeulen M. Characterizing protein–protein interactions using mass spectrometry: challenges and opportunities. Trends Biotechnol. 2016;34(10): 825–834.

Soontorngun N, Larochelle M, Drouin S, Robert F, Turcotte B. Regulation of gluconeogenesis in *Saccharomyces cerevisiae* is mediated by activator and repressor functions of Rds2. Mol Cell Biol. 2007;27(22): 7895–7905.

Souciet J. Ten years of the Génolevures Consortium: A brief history. C R Biol. 2011;334: 580–584.

Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. Nucl Acids Res. 1982;10(9): 2997–3011.

The SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol. 2014;32: 903–914.

Ter Kuile BH, Westerhoff HV. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. FEBS lett. 2001;500. pii: S0014-5793(01)02613-8.

Teusink B, Passarge J, Reijenga CA, Esgalhado E, van der Weijden CC, Schepper M, Walsh MC, Bakker BM, van Dam K, Westerhoff HV, Snoep JL. Can yeast glycolysis be understood in terms of *in vitro* kinetics of the constituent enzymes? Testing biochemistry. Eur J Biochem. 2000;267(17): 5313-5329.

Thakur JK, Athanari H, Yang F, Chau KH, Wagner G, Näär AM. Mediator subunit Gal11p/MED15 is required for fatty acid-dependent gene activation by yeast transcription factor Oaf1p. J Biol Chem. 2009;284(7): 4422–4428.

Tomás-Cobos L, Sanz P. Active Snf1 protein kinase inhibits expression of the *Saccharomyces cerevisiae* HXT1 glucose transporter gene. Biochem J. 2002;368: 657–663.

Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9): 1105-1111.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimental H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2013;7(3): 562–578. doi: 10.1038/nprot.2012.016.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan GJ, van Baren M, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nat Biotechnol. 2010;28(5): 511–515.

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30(9): 418-426.

Verduyn C, Stouthamer AH, Scheffers WA, van Dijken JP. A theoretical evaluation of growth yields in yeasts. Antonie van Leeuwenhoek. 1990;59: 49-63.

Visser D, van Zuylen GA, van Dam JC, Oudtshoorn A, Eman MR, Ras C. Rapid sampling for analysis of in vivo kinetics using the BioScope: a system for continuous-pulse experiments. Biotechnol Bioeng. 2002;79(6): 674-81.

Voet D, Voet JG. Biochemistry, fourth edition. New York: J. Wiley & Sons; 2011.

Wan Y, Arens CE, Wang S, Zuo X, Zhou Y, Xing J, Liu H. Role of the repressor Oaf3p in the recruitment of transcription factors and chromatin dynamics during the oleate response. Biochem J. 2013;449: 507–517.

Wiechert W, ^{13}C-Metabolic flux analysis. Metab Eng. 2001;3: 159-206.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008;452: 872-876.

Wilson WA, Hawley SA, Hardie DG. Glucose repression/derepression in budding yeast: SNF1 protein kinase is activated by phosphorylation under derepressing conditions, and this correlates with a high AMP:ATP ratio. Curr Biol. 1996;6: 1426–1434.

Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. Nature. 1997;387: 708–713.

Woods A, Munday MR, Scott J, Yang X, Carlson M, Carling D. Yeast SNF1 is functionally related to mammalian AMP activated protein kinase and regulates acetyl-CoA carboxylase *in vivo*. J Biol Chem. 1994;269: 19509–19515.

Young ET, Dombek KM, Tachibana C, Ideker T. Multiple pathways are co-regulated by the protein kinase Snf1 and the transcription factors Adr1 and Cat8. J Biol Chem. 2003;278(28): 26146–26158.

Zaman S, Lippman SI, Schneper L, Slonim N, Broach JR. Glucose regulates transcription in yeast through a network of signaling pathways. Mol Syst Biol. 2009;5(245). doi:10.1038/msb.2009.2.

Zamboni N, Fischer E, Sauer U. FiatFlux – a software for metabolic flux analysis from ^{13}C-glucose experiments. BMC Bioinf. 2005;6(209). doi:10.1186/1471-2105-6-209.

Zeng H, Edwards MD, Liu G, Gifford DK. Convolutional neural network architectures for predicting DNA–protein binding. Bioinformatics. 2016;32: i121-i127.

Zerbino DR, Birney E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 2008;18(5): 821–829. doi:10.1101/gr.074492.107].

Chapter 2

A first draft genome for *Kluyveromyces marxianus* strain UFS-Y2791

Abstract

Next-generation sequencing of DNA has transformed the field of biotechnology into an information rich science. It is foreseeable that the initial steps in biotechnology projects involving the study or manipulation of metabolism in microorganisms should involve genome sequencing to serve as a blueprint for further downstream analyses. Using next-generation sequencing, a draft genome could rapidly be sequenced and assembled *de novo* at minimal cost, allowing a variety of genome-scale investigations of newly discovered species and other non-model species. Genome finishing, however, is more costly and time consuming. The question remains, however, whether a properly annotated draft genome would be sufficient for the purpose of genomics and systems biology. To this end, a draft genome of the yeast *Kluyveromyces marxianus* was assembled *de novo* and annotated. Metabolic pathway construction showed that 235 pathways could be constructed from this assembly, suggesting that this cost-effective protocol is also sufficiently rich in information content to serve as blueprint for genomics and systems biology.

Introduction

At the time of commencement of this investigation, the only genomic information on *Kluyveromyces marxianus* was a draft genome, covering only approximately 20% of the genome, and was not publicly available. This level of completion was not sufficient for in-depth comparative genomics, construction of metabolic pathways or gene regulatory networks, or for read-mapping of RNA-seq data. The cost of next-generation sequencing (NGS) has dramatically decreased during the last few years, and efficient algorithms have become available for genome assembly and annotation. As first objective, a nearly complete draft genome for *K. marxianus* strain UFS-Y2791 was sequenced and assembled. This assembly may have been the first for the species at this level of completeness, although other draft genomes and even complete genomes for different strains were assembled and published soon afterwards [Jeong et al. 2012, Lertwattanassakul et al. 2015]. The focus in this investigation was rather a more in-depth analysis of the genome-wide gene regulatory programming of the species, as opposed to providing the first annotated genome. In this chapter, a brief description of the resulting draft genome and annotation is given.

The de Bruijn graph approach to *de novo* genome assembly is performed by mapping identical or nearly identical words (k-mers) between NGS reads to construct a graph (network), and subsequently the graph is traversed to find the shortest path, corresponding to the assembled sequence [Compeau et al. 2011]. The k-mer length is an important parameter and needs to be optimised for the data to yield and optimal assembly. The N50 value is often used to estimate the quality of a *de novo* assembly. The Abyss assembler for short reads works on the de Bruijn graph principle [Simpson et al. 2009]. However, at the time of assembling the draft genome for *K. marxianus* UFS-Y2791, no automated optimisation algorithm was available. Part of the work was, therefore, to find the optimal assembly parameters. For this purpose, a convenient programme was coded in Python to facilitate optimisation of the procedure by viewing assembly statistics. One useful innovation was to annotate each assembly during the optimisation procedure as well, which gave insight into the value of each assembly as a basis for gene-based research. Although this initial draft genome contained many contigs, it provided a wealth of information and formed the basis for the RNA-seq read-mapping and method development presented in later chapters. Even the genome-scale gene regulatory network based on this draft genome compared well against that constructed from a complete genome of a different strain. This shows that the power of NGS could be utilised to rapidly gain deep insights into non-model organisms, for which there is not a large community, even when the genomic blueprint is only a draft genome that could be generated at a low cost.

Materials and Methods

Strain cultivation and DNA extraction

K. marxianus strain UFS-2791, obtained from the yeast culture collection of the University of the Free State, Bloemfontein, was grown in YPD medium consisting of 20 $g.l^{-1}$ glucose, 20 $g.l^{-1}$ peptone and 10 $g.l^{-1}$ yeast extract. Cells were cultivated in 500 ml aerobic shake flasks with a 50 ml working volume, at 180 rpm at 35^{o}C, and harvested in mid-exponential phase, which was determined by growth studies to occur at an OD_{600} value of 0.8. Genomic DNA was extracted from the washed pellet using a kit from Qiagen DNeasy (Hilden, Germany) and according to the manufacturer's instructions without modification. DNA in elution buffer was sent to the Onderstepoort Biotechnology Platform, Pretoria, South Africa for sequencing.

DNA library preparation and sequencing

Library preparation for DNA sequencing was performed by the Onderstepoort Biotechnology Platform, Pretoria, with the Nextera DNA Sample Preparation Kit from Illumina (San Diego, CA, USA) according to the manufacturer's instructions. The Nextera kit both fragments the DNA and ligates adapters containing the barcodes used for de-multiplexing. Paired-end sequencing was performed on the Illumina HiScanSQ platform providing 2×100 bp reads to approximately 100 X coverage of the 10 Mb genome.

De novo assembly and annotation

The data obtained from the Onderstepoort Biotechnology Platform were quality assessed using FastQC and trimmed using Trimmomatic [Bolger et al. 2014] in Galaxy [Afgan et al. 2015]. Initially, Illumina adapter dimers and Illumina adapter readthrough sequences were removed, followed by a sliding filter of four bases with a minimum average Phred score of 20. This resulted in 14 495 422 paired-end reads that could be used in assembly. The Abyss program [Simpson et al. 2009], which uses distributed computer memory, was run on a high-performance cluster at the University of the Free state, South Africa. Assemblies were performed using a k-mer length from 15 to 63. Statistics of all assemblies were inspected using a programme developed in Python, which uses the rpy2 wrapper for calling R code, for the purpose of visualisation. The code was implemented with a graphical user interface using the Tkinter package for Python. N50, N70 and N90 values were calculated in Python 3.0 and visualised. Each assembly was also annotated in terms of open reading frames (ORFs) using Augustus [Stanke et al. 2008]. After finding the optimal assembly, the ORFs were annotated separately

against the SwissProt and Trembl databases of UniProt, using only sequences from the fungal kingdom. The database search was performed on a local high-performance computing cluster. The threshold for a blastp match was set to 1E-10 (see Chapter 3). To increase the richness of the annotation to avoid near-perfect matches to uncharacterised database sequences of *K. marxianus* or other closely related species that were previously submitted to UniProt, a match against a SwissProt protein was taken over a match to a protein in Trembl, in which the vast majority of *K. marxianus* proteins are currently maintained, and have not been manually curated by UniProt.

Results

Agarose gel electrophoresis showed that the genomic DNA was mostly intact (Figure 1). NGS was performed to approximately 100-fold read depth of the genome, which was estimated at close to 11 Mb. Figures 2 to 4 show the quality assessment of genomic NGS reads before trimming, showing outstanding read quality. The median Phred score per read was 38.

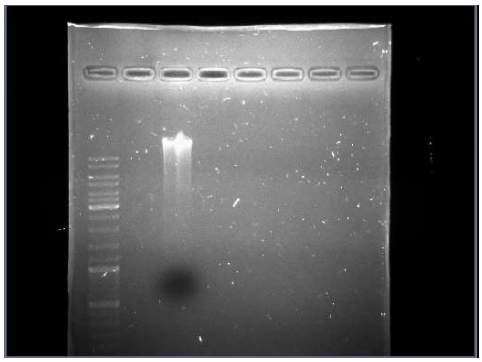


Figure 1. A 5% agarose gel showing genomic DNA from *K. marxianus* UFS-Y2791 to be mostly intact. The GeneRuler™ DNA ladder mix was used, with the upper band representing 10 Kb. Genomic DNA is visible in lane 3.

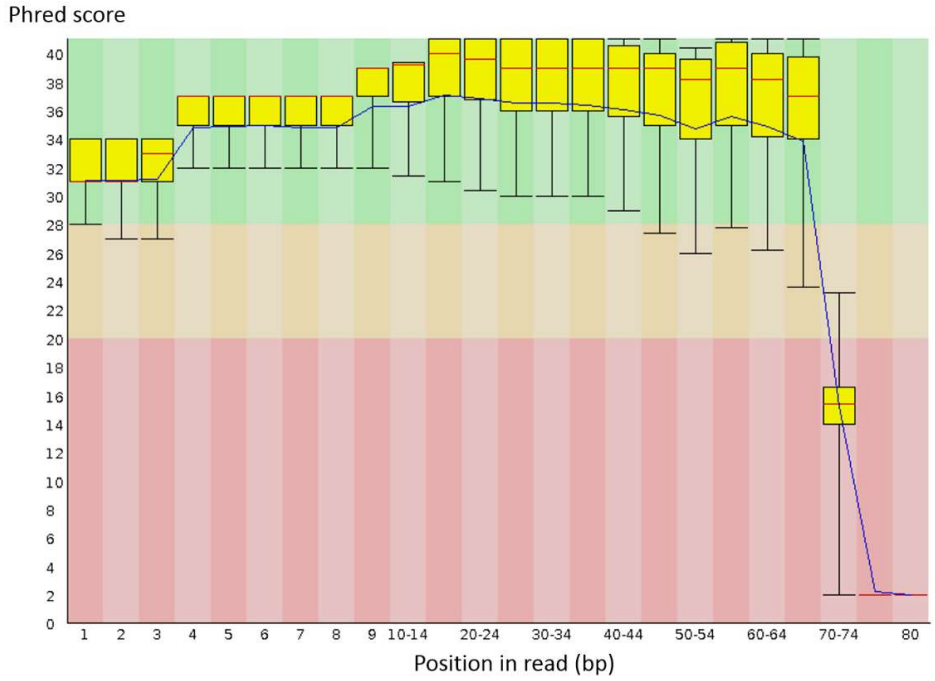


Figure 2. Box-and-whiskers plot of per base quality assessment of genomic NGS data before trimming, visualised using the FastQC tool. Orientation of reads is 5'-3', as sequenced on the Illumina instrument.

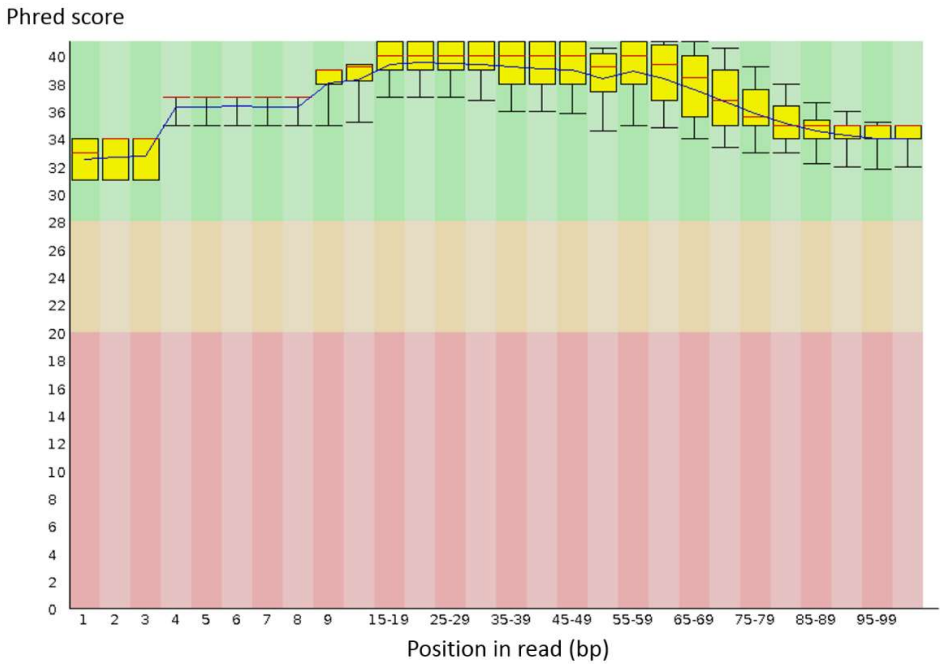


Figure 3. Box-and-whiskers plot of per base quality assessment of genomic NGS data after trimming using the Trimmomatic tool, and visualised using the FastQC tool, showing higher average base quality. Illumina adapters were removed, followed by a sliding filter of four bases (from the 3' end), testing each read for a local minimum average Phred score of 20. The number of bases removed differs for each read. Orientation of reads is 5'-3', as sequenced on the Illumina instrument.

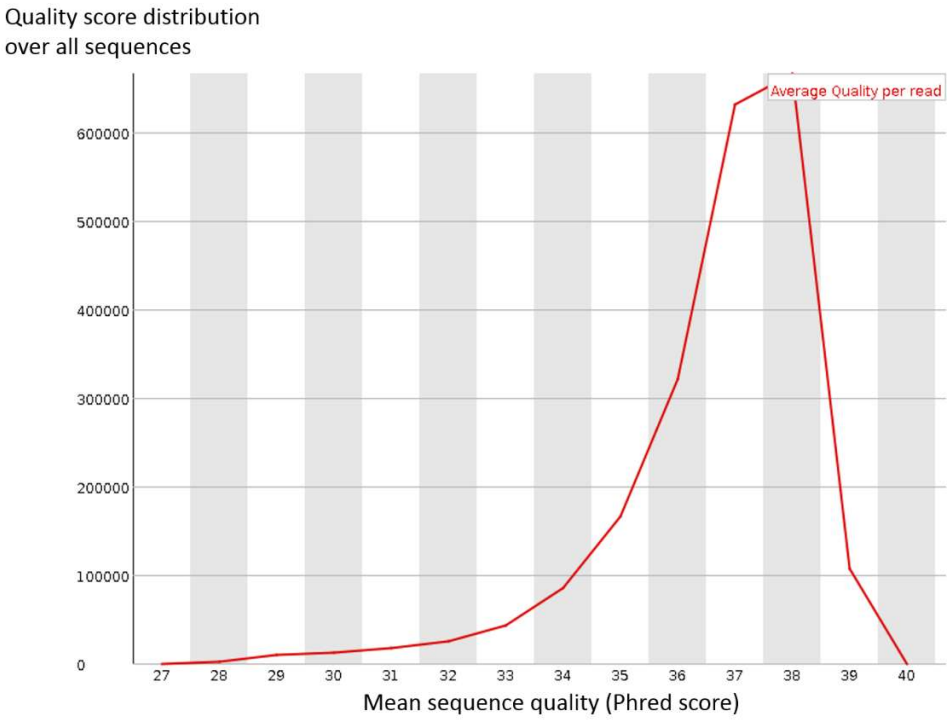


Figure 4. Mean quality score per read before trimming, as obtained from the FastQC tool. Phred scores were calculated over the length of each read. The median Phred score was 38, with a similar mean value.

Scripts were coded in Python to calculate the assembly statistics N50, N75 and N90. In addition, every assembly was annotated in terms of ORFs using Augustus. These, along with assembly statistics, were plotted by the script using the rpy2 wrapper to call algorithms in the R statistical language. A graphical user interface (GUI) was also developed using the Tkinter toolkit for Python. The GUI is shown in Figure 5. An optimal N50 value was found when using a k-mer length of 46 bp (Figure 6), while the longest contig was found when using a k-mer length of 48 bp (Figure 7). The number of contigs was lowest in with k-mer lengths between 34 and 50 bp (Figure 8).

Plotting annotation statistics as a function of k-mer length (Figure 9) revealed and interesting pattern: the number of complete ORFs, defined as having both translation start START and STOP codons, showed a broad plateau when using k-mer lengths of from 33 to 59 bp. This result suggested that the final outcome in functional annotation space would be rather robust to the assembly parameters. Using a k-mer length that is too short (below 33) may result in incorrect assemblies, while too long a k-mer length may result in very low coverage, also resulting in a fragmented genome. In the Augustus annotation, 267 putative introns were found in total, occurring in 201 putative genes.

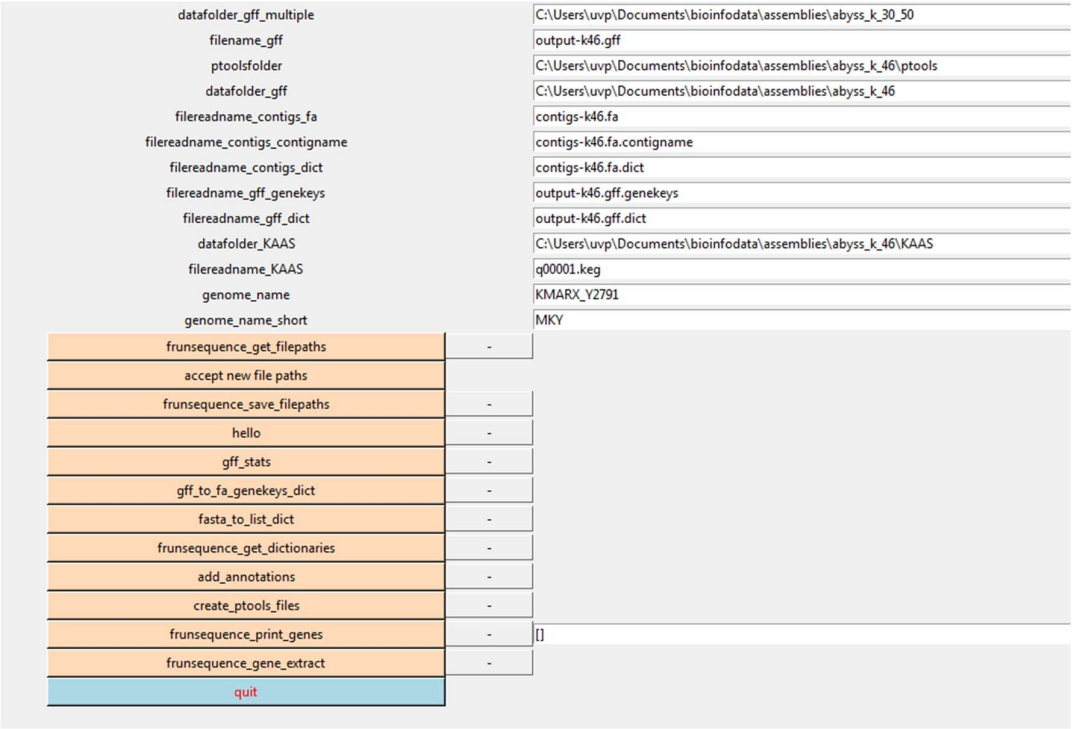


Figure 5. Graphical user interface for plotting assembly statistics using Abyss output. The programme was written in Python and the graphical user interface is based on the Tkinter package.

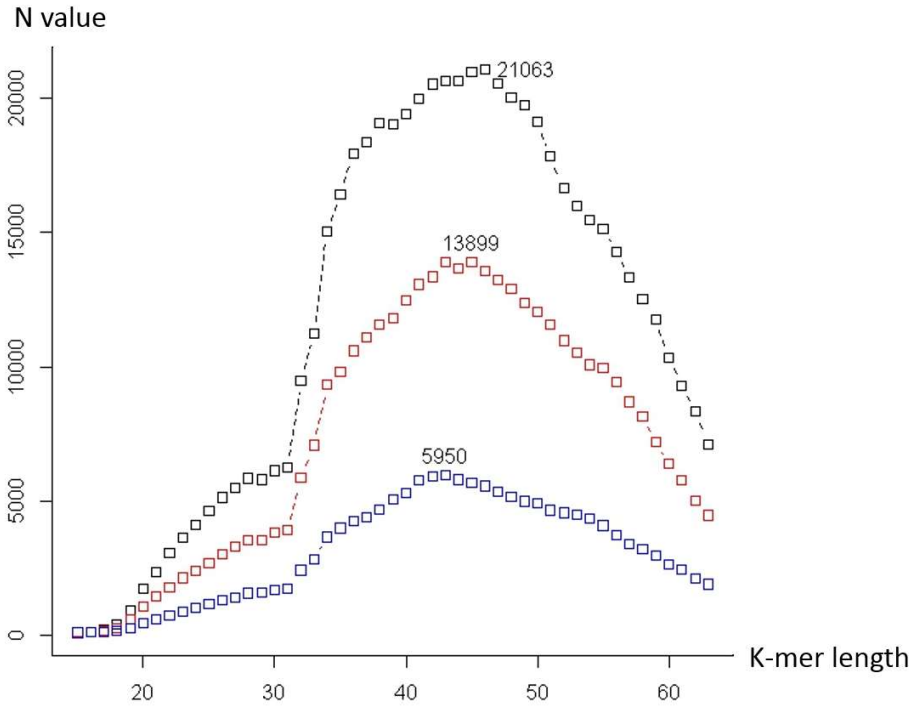


Figure 6. N-value assembly statistics of the UFS-Y2791 draft genome as a function of the k-mer length used in the Abyss assembly. Black indicates the N50, red indicates the N75 and blue indicates the N90. The N50 value of 21 063 was used as metric for choosing the best assembly, corresponding to a k-mer length of 46 bp.

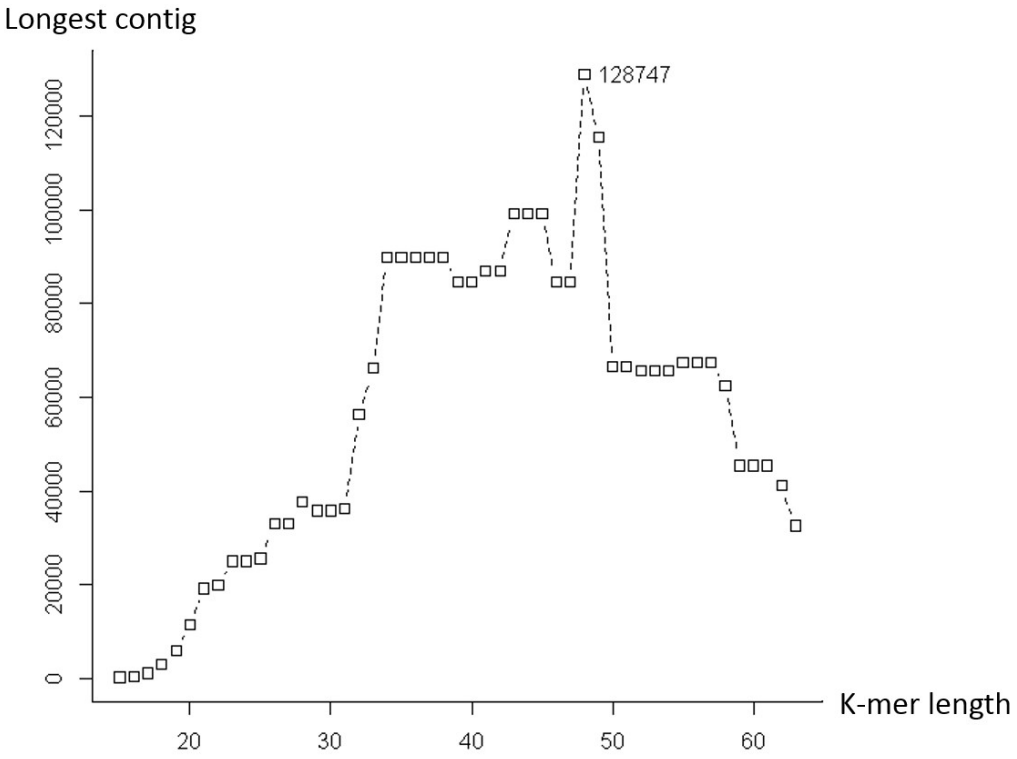


Figure 7. The longest contig in a given assembly, as a function of the k-mer length used in the Abyss assembly.

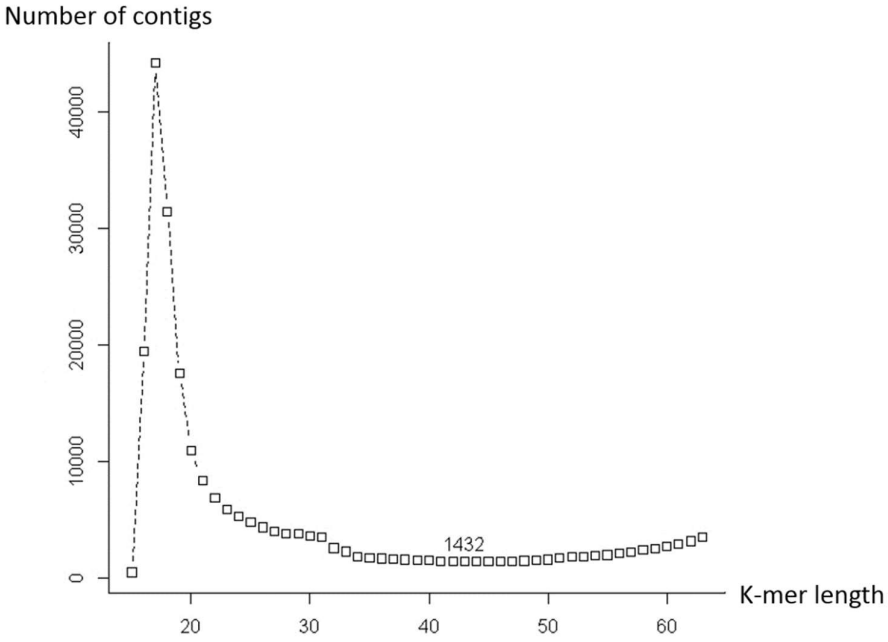


Figure 8. Number of contigs as a function of the k-mer length. The number of contigs reaches an inversed plateau using k-mer lengths of between 35 and 50 bp, in which the most optimal assembly was central.

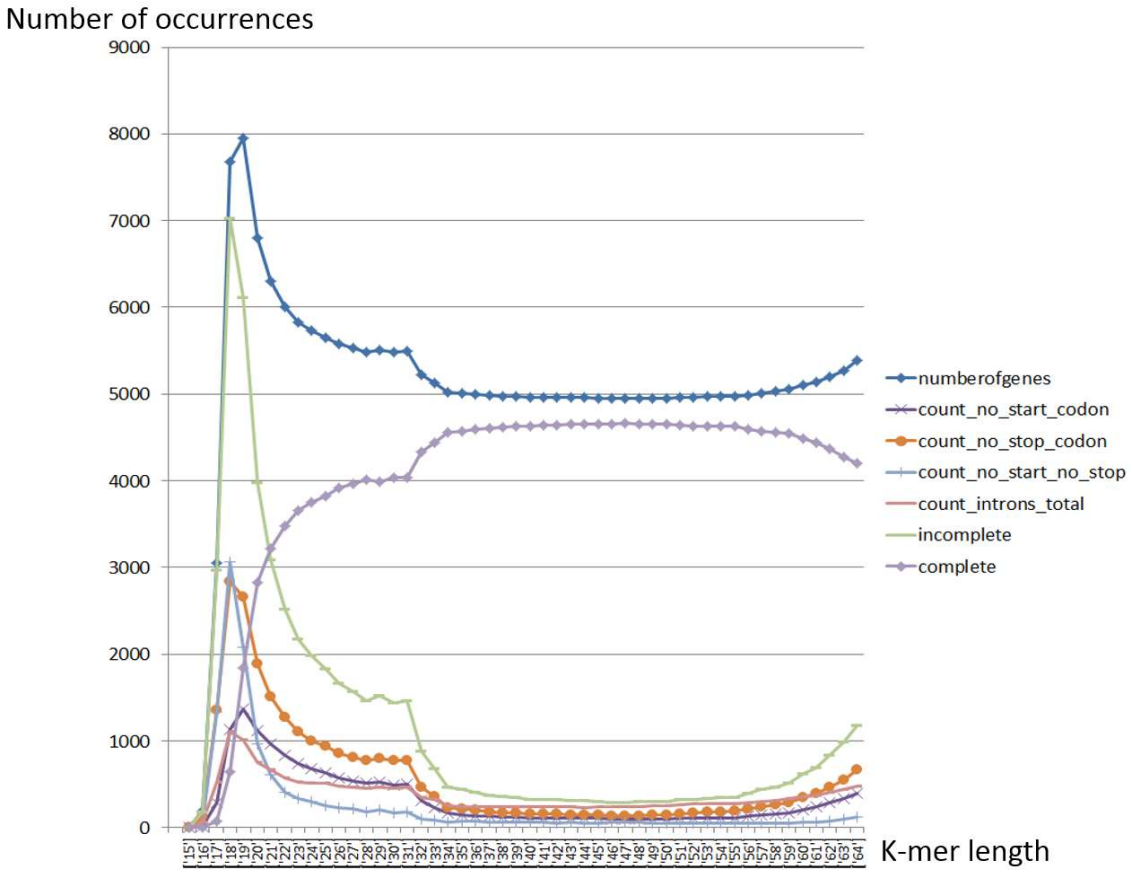


Figure 9. Annotation statistics for various genome assemblies as a function of assembly k-mer length.

Although the longest contig was found at a k-mer length of 48 bp, the best N50 value was chosen as the criterion for optimisation since it is commonly used. The optimal N50 value was 21 063 bp and found at a k-mer length of 46. The threshold length of a contig for submission to the WGS database of the NCBI is 200 bp. In this annotation, 13 genes were contained on contigs smaller than 200 bp. As might be expected, these were mostly putative membrane proteins known to have many internal repeats of hydrophobic alpha helices, which may be the reason for incomplete assembly using the *de novo* approach (Table 1). These ORFs would often be incomplete. Moreover, the upstream intergenic regions would, in most cases, be too short to extract meaningful regulatory features, therefore these contigs were removed from further analyses. The file uploaded to NCBI Genbank contained only contigs of at least 200 bp. This file was used to create mapping files between the contigs and gene features.

During the submission of the sequence data to Genbank, an NCBI BioProject was registered, which is reserved for this strain. All subsequent molecular data for this strain will subsequently be linked to this BioProject (ID PRJNA316809), BioSample (ID SAMN04590183), Submission ID SUB1551262,

accession number LYPD00000000 and locus tag A4A45. All contig names were accordingly adapted with locus tags. Contig names start with the locus tag 'A4A45', followed by '_', followed by a six-digit string ending with the number assigned by the assembly programme Abyss with zeros prepended. Previous contig names like '25' were thus adapted to 'A4A45_000025'. Eight contigs were found to be mitochondrial by the automated Genbank upload server and was annotated as such in the fasta file, before uploading the file a second time to Genbank (Table 2). The version of the draft genome described in the research article from Chapter 3 [Schabort et al. 2016] is version LYPD01000000. The length distribution of the 1 096 contigs is given in Figure 10.

Table 1. Putative genes on contigs shorter than 200 bp that have been discarded.

Reactomica ID	UniProt ID	Protein name	Other names	Gene systematic name
g6.t1	P32478	Cell wall mannoprotein HSP150 (150 kDa heat shock glycoprotein) (Covalently-linked cell wall protein 7) (Protein with internal repeats 2)	HSP150 CCW7 ORE1 PIR2 YJL159W J0558	HSP150
g503.t1	Q02785	ATP-dependent permease PDR12	PDR12 YPL058C LPE14C	PDR12
g1426.t1	W7A6M6	Uncharacterized protein	C922_00460	
g2014.t1	P46587	Heat shock protein SSA2	SSA2 SSA1 CaO19.1065 CaO19.8667	SSA2
g2114.t1	P39004	High-affinity hexose transporter HXT6	HXT7 YDR342C D9651.11	HXT7
g2803.t1	A0A090BNY7	Flo11 super family	KMAR_70233	
g2824.t1	P32901	Peptide transporter PTR2 (Peptide permease PTR2)	PTR2 YKR093W YKR413	PTR2
g3009.t1	P32478	Cell wall mannoprotein HSP150 (150 kDa heat shock glycoprotein) (Covalently-linked cell wall protein 7) (Protein with internal repeats 2)	HSP150 CCW7 ORE1 PIR2 YJL159W J0558	HSP150
g3173.t1	A0A090BDF9	Flocculation protein FLO9	KMAR_30001	
g3805.t1	P18631	Low-affinity glucose transporter (Hexose transporter 1)	RAG1 KHT1 KLLA0D13310g	RAG1

g3832.t1	W0TD05	Cell wall mannoprotein HSP150	KLMA_50254	
g4000.t1	P09435	Heat shock protein SSA3	SSA3 YBL075C YBL06.07 YBL0610	SSA3
g4780.t1	P49374	High-affinity glucose transporter	HGT1 KLLA0A11110g	HGT1

Table 2. Mitochondrial genes as annotated by the Genbank upload server.

Sequence name	Length	Apparent source
A4A45_000025	299	mitochondrion
A4A45_001165	338	mitochondrion
A4A45_001219	670	mitochondrion
A4A45_001222	623	mitochondrion
A4A45_001756	567	mitochondrion
A4A45_002411	261	mitochondrion
A4A45_002559	1585	mitochondrion
A4A45_004867	279	mitochondrion

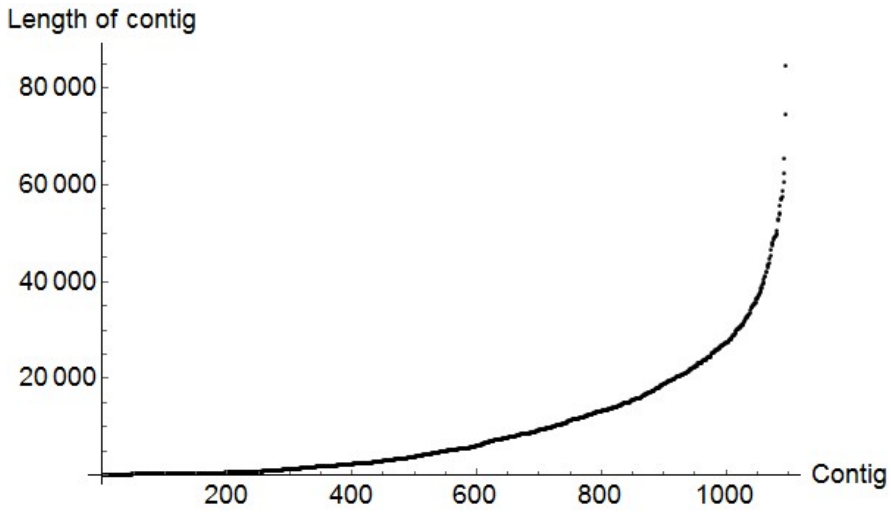


Figure 10. Length distribution of 1 096 contigs (of at least 200 bp) for *K. marxianus* UFS-Y2791. The total length of this draft genome was 10 695 463 bp.

The resulting 4953 protein encoding genes were subsequently functionally annotated on the Kegg-Kaas server, resulting in enzyme EC numbers and GO terms (reported in Chapter 3). The pathways discovered by the Pathway Tools Pathologic algorithm are shown in Table 3 below. A total of 235 pathways could be identified in this manner.

Table 3. A summary of pathways discovered in the draft genome of K. marxianus UFS-Y2791 using the PathoLogic algorithm of Pathway Tools.

MetaCYC ID	Pathway name	number of genes discovered
PWY-6126	adenosine nucleotides *de novo* biosynthesis	35
TRNA-CHARGING-PWY	tRNA charging pathway	34
FERMENTATION-PWY	mixed acid fermentation	21
GLUCONEO-PWY	gluconeogenesis I	17
ANAGLYCOLYSIS-PWY	glycolysis III	16
GLYCOLYSIS	glycolysis I	15
PWY66-21	ethanol degradation II (cytosol)	15
PWY-5690	TCA cycle variation III (eukaryotic)	15
PWY-561	superpathway of glyoxylate cycle	14
MANNOSYL-CHITO-DOLICHOL-BIOSYNTHESIS	dolichyl-diphosphooligosaccharide biosynthesis	14
ANARESP1-PWY	respiration (anaerobic)	12
PWY-6333	acetaldehyde biosynthesis I	11
PWY-5486	pyruvate fermentation to ethanol II	11
PWY-3841	folate transformations II (plants)	11
ARGSYNBSUB-PWY	arginine biosynthesis II (acetyl cycle)	10
GLYOXYLATE-BYPASS	glyoxylate cycle	10
PWY-2201	folate transformations I	9
PWY-5079	phenylalanine degradation III	8
PWY66-162	ethanol degradation IV (peroxisomal)	7
LYSINE-AMINOAD-PWY	lysine biosynthesis IV	7
PWY-2161	folate polyglutamylation	7
PWY-5076	leucine degradation III	7
PWY-5057	valine degradation II	7
PWY3O-4108	tyrosine degradation III	7
PWY-6168	flavin biosynthesis III (fungi)	7
PWY4FS-8	phosphatidylglycerol biosynthesis II (non-plastidic)	7
PWY-6352	3-phosphoinositide biosynthesis	7
TRIGLSYN-PWY	triacylglycerol biosynthesis	7
HISTSYN-PWY	histidine biosynthesis	6
PWY-5751	phenylethanol biosynthesis	6
PWY-5686	uridine-5'-phosphate biosynthesis	6
PWY-5082	methionine degradation III	6
PWY-5078	isoleucine degradation II	6
DETOX1-PWY	superoxide radicals degradation	6
BRANCHED-CHAIN-AA-SYN-PWY	superpathway of leucine, valine, and isoleucine biosynthesis	6
PWY-3001	isoleucine biosynthesis I	6
ILEUSYN-PWY	isoleucine biosynthesis I (from threonine)	6
PWY66-5	superpathway of cholesterol biosynthesis	6
PWY-922	mevalonate pathway I	6
PWY-5667	CDP-diacylglycerol biosynthesis I	6
PWY0-166	pyrimidine deoxyribonucleotides *de novo* biosynthesis I	6
PWY-5136	fatty acid β-oxidation II (core pathway)	5
FAO-PWY	fatty acid β-oxidation I	5
PWY-6629	superpathway of tryptophan biosynthesis	5
TRPSYN-PWY	tryptophan biosynthesis	5
PWY-5750	itaconate biosynthesis	5
MALATE-ASPARTATE-SHUTTLE-PWY	aspartate degradation II	5
VALSYN-PWY	valine biosynthesis	5
PWY-6074	zymosterol biosynthesis	5
PWY-6293	cysteine biosynthesis IV (fungi)	5
PWY-821	superpathway of sulfur amino acid biosynthesis (*Saccharomyces cerevisiae*)	5
PWY-801	homocysteine and cysteine interconversion	5
GLUCOSE1PMETAB-PWY	glucose and glucose-1-phosphate degradation	5
PWY0-163	salvage pathways of pyrimidine ribonucleotides	5
PWY-5083	NAD/NADH phosphorylation and dephosphorylation	5

PWY-6351	D-*myo*-inositol (1,4,5)-trisphosphate biosynthesis	5
PYRUVDEHYD-PWY	acetyl-CoA biosynthesis (from pyruvate)	5
LEUSYN-PWY	leucine biosynthesis	4
ARO-PWY	chorismate biosynthesis I	4
NONOXIPENT-PWY	pentose phosphate pathway (non-oxidative branch)	4
PWY0-1182	trehalose degradation II (trehalase)	4
PANTO-PWY	phosphopantothenate biosynthesis I	4
PWY-5938	*R*-acetoin biosynthesis I	4
PWY0-662	PRPP biosynthesis I	4
PWY-4041	γ-glutamyl cycle	4
PWY-5920	heme biosynthesis II	4
HEME-BIOSYNTHESIS-II	heme biosynthesis from uroporphyrinogen-III I	4
PWY-5767	glycogen degradation III	4
PWY-5941	glycogen degradation II	4
PWY-5659	GDP-mannose biosynthesis	4
PWY-6277	superpathway of 5-aminoimidazole ribonucleotide biosynthesis	4
PWY-6121	5-aminoimidazole ribonucleotide biosynthesis I	4
PWY-6122	5-aminoimidazole ribonucleotide biosynthesis II	4
PYRIDNUCSAL-PWY	NAD salvage pathway I	4
PWY-6357	phosphate utilization in cell wall regeneration	4
PWY-6348	phosphate acquisition	4
PWY-5189	tetrapyrrole biosynthesis II	4
SPHINGOLIPID-SYN-PWY	sphingolipid metabolism	4
PENTOSE-P-PWY	pentose phosphate pathway	4
OXIDATIVEPENT-PWY	pentose phosphate pathway (oxidative branch)	4
PWY-6075	ergosterol biosynthesis	4
PWY-3781	aerobic respiration -- electron donor II	4
PWY-841	purine nucleotides *de novo* biosynthesis II	4
PWY-6124	inosine-5'-phosphate biosynthesis II	4
COA-PWY	coenzyme A biosynthesis	4
UDPNACETYLGALSYN-PWY	UDP-*N*-acetyl-D-glucosamine biosynthesis II	4
PWY-5760	β-alanine biosynthesis IV	3
PWY-2724	fatty acid ω-oxidation	3
VALDEG-PWY	valine degradation I	3
PWY-6317	galactose degradation I (Leloir pathway)	3
PWY-4981	proline biosynthesis II (from arginine)	3
TRESYN-PWY	trehalose biosynthesis I	3
SERSYN-PWY	serine biosynthesis	3
SER-GLYSYN-PWY	superpathway of serine and glycine biosynthesis I	3
GLYSYN-PWY	glycine biosynthesis I	3
PWY0-381	glycerol degradation I	3
PLPSAL-PWY	pyridoxal 5'-phosphate salvage pathway	3
FOLSYN-PWY	superpathway of tetrahydrofolate biosynthesis and salvage	3
PWY-6613	tetrahydrofolate salvage from 5,10-methenyltetrahydrofolate	3
ARG-PRO-PWY	arginine degradation VI (arginase 2 pathway)	3
GLYCLEAV-PWY	glycine cleavage complex	3
PWY-6164	3-dehydroquinate biosynthesis I	3
PROSYN-PWY	proline biosynthesis I	3
PWY-5067	glycogen biosynthesis II (from UDP-D-Glucose)	3
PWY-6614	tetrahydrofolate biosynthesis	3
PWY-5669	phosphatidylethanolamine biosynthesis I	3
ARGASEDEG-PWY	arginine degradation I (arginase pathway)	3
PWY-5344	homocysteine biosynthesis	3
PWY3O-450	phosphatidylcholine biosynthesis I	3
PWY-5041	*S*-adenosyl-L-methionine cycle II	3
ARGDEG-V-PWY	arginine degradation X (arginine monooxygenase pathway)	3
PWY0-1507	biotin biosynthesis from 7-keto-8-aminopelargonate	3
PWY-3561	choline biosynthesis III	3
LIPAS-PWY	triacylglycerol degradation	3
PWY-5084	2-ketoglutarate dehydrogenase complex	3
PWY-6129	dolichol and dolichyl phosphate biosynthesis	3

PWY-4061	glutathione-mediated detoxification	3
LIPASYN-PWY	phospholipases	3
PWY-5123	*trans, trans*-farnesyl diphosphate biosynthesis	3
PWY0-162	pyrimidine ribonucleotides *de novo*biosynthesis	3
PWY-6628	superpathway of phenylalanine biosynthesis	3
PWY-5687	pyrimidine ribonucleotides interconversion	3
PHESYN	phenylalanine biosynthesis I	3
PWY-4261	glycerol degradation IV	2
PWY-1722	formaldehyde oxidation V (tetrahydrofolate pathway)	2
HOMOSER-THRESYN-PWY	threonine biosynthesis from homoserine	2
HOMOSERSYN-PWY	homoserine biosynthesis	2
PWY-2541	plant sterol biosynthesis	2
PWY-5670	epoxysqualene biosynthesis	2
PWY-6147	6-hydroxymethyl-dihydropterin diphosphate biosynthesis	2
PWY-6118	glycerol-3-phosphate shuttle	2
SO4ASSIM-PWY	sulfate reduction I (assimilatory)	2
PWY-5340	sulfate activation for sulfonation	2
PWY-6392	*meso*-butanediol biosynthesis II	2
PWY-6391	*meso*-butanediol biosynthesis I	2
PWY-5951	(*R*,*R*)-butanediol biosynthesis	2
PWY3O-246	(*R*,*R*)-butanediol degradation	2
PWY-4441	DIMBOA-glucoside degradation	2
PWY-5143	fatty acid activation	2
PWY-5080	very long chain fatty acid biosynthesis	2
PWY-6000	γ-linolenate biosynthesis II (animals)	2
PWY0-1313	acetate conversion to acetyl-CoA	2
PWY-6536	4-aminobutyrate degradation III	2
PWY-6693	galactose degradation IV	2
PWY-5194	siroheme biosynthesis	2
PWY-5081	tryptophan degradation VIII (to tryptophol)	2
ILEUDEG-PWY	isoleucine degradation I	2
PWY0-1296	purine ribonucleosides degradation to ribose-1-phosphate	2
METHIONINE-DEG1-PWY	methionine degradation I (to homocysteine)	2
P21-PWY	pentose phosphate pathway (partial)	2
PWY-5691	urate degradation to allantoin	2
PWY-5386	methylglyoxal degradation I	2
PWY3DJ-11470	sphingosine and sphingosine-1-phosphate metabolism	2
PWY-6151	*S*-adenosyl-L-methionine cycle I	2
GLUTATHIONESYN-PWY	glutathione biosynthesis	2
PWY-5970	fatty acids biosynthesis (yeast)	2
PWY-5122	geranyl diphosphate biosynthesis	2
PWY0-501	lipoate biosynthesis and incorporation I	2
PWY-6556	pyrimidine ribonucleosides degradation II	2
THRESYN-PWY	threonine biosynthesis	2
ASPARTATESYN-PWY	aspartate biosynthesis	2
NAD-BIOSYNTHESIS-II	NAD salvage pathway II	2
PWY0-1264	biotin-carboxyl carrier protein assembly	2
PWY4FS-6	phosphatidylethanolamine biosynthesis II	2
PWY-5921	L-glutamine biosynthesis II (tRNA-dependent)	2
PWY-6281	selenocysteine biosynthesis II (archaea and eukaryotes)	2
PWY-4081	glutathione redox reactions I	2
PWY3O-4106	NAD salvage pathway III	2
PWY-5697	allantoin degradation to ureidoglycolate I (urea producing)	2
PWY-5458	methylglyoxal degradation V	1
PWY-5148	acyl-CoA hydrolysis	1
GLYSYN-ALA-PWY	glycine biosynthesis III	1
PWY0-42	2-methylcitrate cycle I	1
BGALACT-PWY	lactose degradation III	1
XYLCAT-PWY	xylose degradation I	1
PWY-4101	sorbitol degradation I	1
PWY3O-19	ubiquinol-6 biosynthesis (eukaryotic)	1

PWY-5324	lysine degradation IX	1
GLYSYN-THR-PWY	glycine biosynthesis IV	1
CITRULLINE-DEG-PWY	citrulline degradation	1
PWY-6330	acetaldehyde biosynthesis II	1
GLUTAMATE-DEG1-PWY	glutamate degradation I	1
ASPARAGINE-DEG1-PWY	asparagine degradation I	1
ALLANTOINDEG-PWY	superpathway of allantoin degradation in yeast	1
PWY-5703	urea degradation I	1
GLUTSYNIII-PWY	glutamate biosynthesis III	1
PWY-6619	adenine and adenosine salvage VI	1
ASPARAGINE-BIOSYNTHESIS	asparagine biosynthesis I	1
ALANINE-SYN2-PWY	alanine biosynthesis II	1
LCYSDEG-PWY	L-cysteine degradation II	1
AMMASSIM-PWY	superpathway of glutamate biosynthesis	1
GLUTSYN-PWY	glutamate biosynthesis I	1
GLUGLNSYN-PWY	glutamate biosynthesis IV	1
GLUTAMINEFUM-PWY	glutamine degradation II	1
SAM-PWY	S-adenosyl-L-methionine biosynthesis	1
ERGOSTEROL-SYN-PWY	superpathway of ergosterol biosynthesis	1
PWY66-341	cholesterol biosynthesis I	1
PWY66-3	cholesterol biosynthesis II (via 24,25-dihydrolanosterol)	1
PWY66-4	cholesterol biosynthesis III (via desmosterol)	1
PWY-6132	lanosterol biosynthesis	1
PWY-6361	1D-*myo*-inositol hexakisphosphate biosynthesis I	1
PWY-46	putrescine biosynthesis III	1
THIOREDOX-PWY	thioredoxin pathway	1
PWY0-1021	alanine biosynthesis III	1
PWY-5910	superpathway of geranylgeranyldiphosphate biosynthesis I (via mevalonate)	1
PWY-5120	geranylgeranyldiphosphate biosynthesis	1
PWY-5389	methylthiopropionate biosynthesis	1
PWY-5350	thiosulfate disproportionation III (rhodanese)	1
GLNSYN-PWY	glutamine biosynthesis I	1
PWY3O-210	glutamate degradation IX (via 4-aminobutyrate)	1
PWY-6612	superpathway of tetrahydrofolate biosynthesis	1
PWY-6358	superpathway of D-*myo*-inositol (1,4,5)-trisphosphate metabolism	1
PWY0-1305	glutamate dependent acid resistance	1
PWY-6364	D-*myo*-inositol (1,3,4)-trisphosphate biosynthesis	1
PWY-6363	D-*myo*-inositol (1,4,5)-trisphosphate degradation	1
SALVPURINE2-PWY	xanthine and xanthosine salvage	1
PWY-6605	adenine and adenosine salvage II	1
P121-PWY	adenine and adenosine salvage I	1
PWY-5269	cardiolipin biosynthesis II	1
PWY-6424	sitosterol biosynthesis	1
PWY-5966	fatty acid biosynthesis initiation II	1
PWY-5996	oleate biosynthesis II (animals)	1
PWY-6012	acyl carrier protein metabolism	1
PWY-6543	p-aminobenzoate biosynthesis	1
GLUT-REDOX-PWY	glutathione redox reactions II	1
PWY-5870	ubiquinol-8 biosynthesis (eukaryotic)	1
HEXPPSYN-PWY	hexaprenyl diphosphate biosynthesis	1
ARGSPECAT-PWY	spermine biosynthesis	1
BSUBPOLYAMSYN-PWY	spermidine biosynthesis I	1
PWY-5886	4-hydroxyphenylpyruvate biosynthesis	1
GLUCONSUPER-PWY	D-gluconate degradation	1
PWY-6019	pseudouridine degradation	1
PWY-6606	guanosine nucleotides degradation II	1
ACETOACETATE-DEG-PWY	acetoacetate degradation (to acetyl CoA)	1
MANNCAT-PWY	D-mannose degradation	1

Discussion

NGS has become a cost-effective method of genome sequencing. Genome finishing is still more laborious and expensive, however. It is now necessary to explore the information content obtainable with only NGS and *de novo* assembly, for the purpose of projects studying a species for the first time. Most likely, the short read length of 100 bp obtained on the Illumina instrument was the reason for incomplete assembly. The fact that the genome is likely diploid might have been an additional cause of a rather low N50 value. Nevertheless, the outstanding read quality resulted in a functionally annotated draft genome that is ready for various bioinformatics and computational biology approaches to study the metabolism and gene regulation of this species. A total of 235 pathways could be assigned. For many of these pathways, large and possibly complete gene sets were assigned, including adenosine nucleotide *de novo* biosynthesis (35 genes), glycolysis (15 genes), TCA cycle variation III (15 genes) and fatty acid β-oxidation I (5 genes). For other pathways, only a few genes or even a single gene were mapped. The Pathologic algorithm [Karp et al. 2009] variably defines a pathway as either a complete set of reactions, or a uniquely identifying reaction or set of reactions, distinguishing a pathway from others, hence some pathways are represented by a single reaction and gene. Even with partial gene information, this nearly complete blueprint of the cellular machinery should be sufficient to enable the analysis of differential gene expression of metabolic pathways. In addition, gene set enrichment statistics [Ideker et al. 2002, Patil and Nielsen 2005] is potentially robust to missing data.

Conclusions

A first draft genome was assembled for *K. marxianus* UFS-Y2791 using *de novo* assembly. Pathway annotation resulted in 235 pathways, indicating that the cost and time effective protocol of NGS and *de novo* genome assembly could lead to a wealth of information in the context of systems biology. For non-model organisms such as *K. marxianus*, the goal of discovering differentially regulated pathways could thus be reached, even before the more time consuming and expensive process of genome finishing. Hence, the procedure should gain popularity with the increasing read length and cost effectivity of NGS.

References

Afgan E, Sloggett C, Goonasekera N, Makunin I, Benson D, Crowe M, Gladman S, Kowsar Y, Pheasant M, Horst R, Lonie A. Genomics Virtual Laboratory: A Practical Bioinformatics Workbench for the Cloud. PLoS One. 2015;10(10): e0140829. doi: 10.1371/journal.pone.0140829.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15): 2114-2120.

Compeau PEC, Pevsner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol. 2011;29(11): 987-911.

Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signaling circuits in molecular interaction networks. Bioinformatics. 2002;18(1): 233-240.

Jeong H, Lee DH, Kim SH, Kim HJ, Lee K, Song JY, Kim BK, Sung BH, Park JC, Sohn JH, Koo HM, Kim JF. Genome sequence of the thermotolerant yeast *Kluyveromyces marxianus* var. marxianus KCTC 17555. Eukaryot Cell. 2012; 12: 1584-1585. doi: 10.1128/EC.00260-12.

Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, Kaipa P, Gilham F, Spaulding A, Popescu L, Altman T, Paulsen I, Keseler IM, Caspi R. PathwayTools version 13.0: integrated software for pathway/genome informatics and systems biology. Briefings Bioinf. 2009;11(1): 40-79. doi:10.1093/bib/bbp043.

Lertwattanasakul N, Kosaka T, Hosoyama A, Suzuki Y, Rodrussamee N, Matsutani M, Murata M, Fujimoto N, Suprayogi S, Tsuchikane K, Limtong S, Fujita N, Yamada M. Genetic basis of the highly efficient yeast *Kluyveromyces marxianus*: complete genome sequence and transcriptome analyses. Biotechnol Biofuels. 2015;8(47). doi:10.1186/s13068-015-0227-x.

Patil KR, Nielsen J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc Natl Acad Sci USA. 2005;102(8): 2685-2689.

Schabort DWP, Letebele PK, Steyn L, Kilian SG, du Preez JC. Differential RNA-seq, multi-network analysis and metabolic regulation analysis of *Kluyveromyces marxianus* reveals a compartmentalised response to xylose. PLoS ONE. 2016; 11(6): e0156242. doi:10.1371/journal.pone.0156242.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABySS: A parallel assembler for short read sequence data. Genome Res. 2009;19: 1117-1123.

Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics. 2008;24(5): 637-644. doi: 10.1093/bioinformatics/btn013.

Chapter 3

Differential RNA-seq, multi-network analysis and metabolic regulation analysis of *Kluyveromyces marxianus* reveals a compartmentalised response to xylose

This paper was published online in PLoS ONE on 17 June 2016. See supporting information online at doi:10.1371/journal.pone.0156242**.**
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0156242

Schabort DWP, Letebele PK, Steyn L, Kilian SG, du Preez JC. Differential RNA-seq, multi-network analysis and metabolic regulation analysis of *Kluyveromyces marxianus* reveals a compartmentalised response to xylose. PLoS ONE. 2016;11(6): e0156242. doi:10.1371/journal.pone.0156242.

Author contributions

Schabort DWP: Conceptualisation of investigation, design of experiments, writing of programmes, bioinformatics and analysis of RNA-seq data, writing of the manuscript.

Letebele PK: Yeast cultivation for RNA-seq analysis, read-mapping of RNA-seq data and differential expression (BSc Hons student of Schabort).

Steyn L: Provision of laboratory equipment and materials, contribution to literature study.

Kilian SG: Provision of laboratory equipment and materials, review of manuscript.

du Preez JC: Design of experiments, provision of laboratory equipment and materials, editing and review of manuscript.

Abstract

We investigated the transcriptomic response of a new strain of the yeast *Kluyveromyces marxianus*, in glucose and xylose media using RNA-seq. The data were explored in a number of innovative ways using a variety of networks types, pathway maps, enrichment statistics, reporter metabolites and a flux simulation model, revealing different aspects of the genome-scale response in an integrative systems biology manner. The importance of the subcellular localisation in the transcriptomic response is emphasised here, revealing new insights. As was previously reported by others using a rich medium, we show that peroxisomal fatty acid catabolism was dramatically up-regulated in a defined xylose mineral medium without fatty acids, along with mechanisms to activate fatty acids and transfer products of β-oxidation to the mitochondria. Notably, we observed a strong up-regulation of the 2-methylcitrate pathway, supporting capacity for odd-chain fatty acid catabolism. Next we asked which pathways would respond to the additional requirement for NADPH for xylose utilisation, and rationalised the unexpected results using simulations with Flux Balance Analysis. On a fundamental level, we investigated the contribution of the hierarchical and metabolic regulation levels to the regulation of metabolic fluxes. Metabolic regulation analysis suggested that genetic level regulation plays a major role in regulating metabolic fluxes in adaptation to xylose, even for the high capacity reactions, which is unexpected. In addition, isozyme switching may play an important role in re-routing of metabolic fluxes in subcellular compartments in *K. marxianus*.

Introduction

The yeast *Kluyveromyces marxianus* is emerging as a host for metabolic engineering and recombinant protein production, having a number of advantages over *Saccharomyces cerevisiae*. These characteristics include thermotolerance and the ability to utilise a wide variety of sugars, including the pentose xylose, that are abundant in lignocellulosic biomass. Moreover, it is probably the fastest growing eukaryote on Earth [1]. Previous studies of this yeast highlighted the large physiological variation among *K. marxianus* strains, in terms of their proneness to fermentation and ability to utilize various substrates, suggesting genetic diversity within the species [2]. Our strain UFS-Y2791 produces a significant amount of ethanol even under aerobic conditions, suggesting that this strain is not typically Crabtree negative. We recently assembled a draft genome of strain UFS-Y2791 from our University of the Free State MIRCEN yeast culture collection. The strain was originally isolated from juice prepared from the arid region succulent *Agave americana*. To enable metabolic engineering of *K. marxianus* and other yeasts, it is important to understand genetic responses to environmental factors and different substrates. Such an understanding requires both advanced high-throughput omics methods as well as various integrative computational methods. Thus far, only two studies of genome-scale transcript levels in *K. marxianus* have been published. Work on strain DMKU 3-1042 suggested that the yeast up-regulated β-oxidation in the absence of glucose repression in a complex xylose medium due to the use of lipids as additional carbon source in the absence of glucose, implying possible utilization of the small amounts of lipids present in the rich medium [3]. A question that may also arise here is that in case lipids were indeed present in the rich medium, whether the response in the β-oxidation was due to glucose de-repression or due to stimulation by lipids. Another study focussed on the response of strain Y179 to anaerobic versus micro-aerobic conditions, as well as to the concentration of inulin in the medium. This study focussed on highlighting differential expression in various stress-response genes, those involved in autophagy, and a number of individual key enzymes. Both of these recent studies used a complex medium containing peptone and yeast extract [4].

Next-generation sequencing (NGS) has become a popular method not only for sequencing genomes but also for various other experiments [5,6]. RNA-seq is a method in which RNA transcripts are reverse-transcribed and quantitatively sequenced [7]. This method can accurately quantify differential expression and the improved sensitivity and dynamic range, as well as the ability to elucidate splice variants, makes it superior to microarray technology. The combination of genome sequencing and RNA-seq of a new

species or strain is a powerful approach to rapidly gain both a blueprint (genome) and a response (transcriptome) to some perturbations [3] or when comparing different species [8]. Innovative methods are now needed to effectively use both the blueprint and the response in a concise manner that is attractive to scientists, since a massive amount of data is generated in a single experiment. It is a major challenge to investigate and represent omics results at the genome scale. Gene set enrichment using Gene Ontology (GO) is an established method for microarrays and is now becoming established for RNA-seq (reviewed in [9]).

Analysis of metabolic pathways could be a sensible additional approach as it gives a sense of directionality to the response, and the scientist often associates a certain endpoint metabolite with a pathway, providing a means to simplify the understanding of the dataset. Painted renderings could be made of individual pathways, but often the number of pathways excludes a concise representation, and the integrated nature of metabolism is lost. Further, in metabolic pathways the interactions (reactions) consist of hypergraphs containing more than one substrate or product, complicating the rendering. Cellular overviews, as is available with software such as Pathway Tools, summarises metabolism into a single image [10]. However, the cellular overview method usually assumes a single master framework and inevitably ignores selected highly connected nodes. Cellular compartmentalisation is also often neglected in omics analysis due to the added complexity.

An approach that is complementary to the pathway-based understanding of metabolism is that of the study of metabolite levels. Metabolomics, which entails the characterisation and quantification of small compounds in the cell, is still technologically demanding and labour intensive, however (reviewed in [11,12]). A new approach in systems biology is to derive compounds that are likely differentially expressed or extensively homeostatically regulated, from a differential transcriptomics dataset. These are the reporter metabolites [13] and the method could be generalised and applied in many scenarios such as described elsewhere [14].

Models of biochemical pathways have much potential to reveal new insights that are not obvious from the exploration of datasets. The understanding of complex biochemical datasets using computational models is called systems biology. Some flux modelling approaches such as Flux Balance Analysis (FBA) simulations could be done even at the genome scale [15,16], as was recently reported for *K. lactis* [17]. A long-standing fundamental question is how the flux through a metabolic pathway is regulated. Is the

change in a given flux achieved by changes in the concentrations of metabolites that affect that enzyme (metabolic level), or through changes in gene expression or post-translational modifications (hierarchical level)? By combining flux models with measured metabolite exchange fluxes, or through more complex ^{13}C-Metabolic Flux Analysis, estimation of fluxes at different physiological states can be very informative. Metabolic Regulation Analysis (MRA) combines differential expression levels with differential flux estimates to reveal for each flux the contribution of the metabolic and hierarchical levels of regulation [18].

Here we report on a detailed RNA-seq transcriptomics study to explore the response of *K. marxianus* UFS-2791 to glucose and xylose in a chemically defined medium under aerobic conditions, including a number of different systemic analyses. Although a major potential future application of *K. marxianus* may be ethanol production from lignocellulosic biomass, which is an anaerobic or oxygen limited process where both glucose and xylose may be present, we chose to perform aerobic cultivations with glucose or xylose separately to remove any confounding effects. The strain also utilizes these sugars sequentially. The high cultivation temperature of 35^{o}C was also chosen as this strain was determined to have a growth optimum close to 35^{o}C. Our samples were also taken during mid-exponential phase to eliminate the effect of any possible ethanol stress that may occur later during the fermentation. To address the need for effective integrative exploration of omics data, including RNA-seq, and to combine these data with modelling and simulation, we developed the Reactomica software. We combined gene set enrichment of Gene Ontology (GO), reporter metabolites, metabolic pathway maps with a strong focus on subcellular compartmentalisation, and two new approaches of representation, namely pathway-to-pathway networks and reporter metabolite-enzyme networks.

The aims were to first effectively explore the key features of the differential response to xylose in a defined medium under aerobic conditions and determine whether the peroxisomal lipid catabolic response previously observed [3] was limited to a rich medium. Subsequently, the central carbon metabolism was investigated in detail, separating the response into subcellular compartments. It is known that in most yeasts that are able to utilize xylose, the two-step conversion via NADPH dependent xylose reductase and the NAD^+ dependent xylitol dehydrogenase is present and that the co-factor independent isomerase reaction is absent, as in *K. marxianus* [3]. Under aerobic conditions, the additional NADH produced by xylitol dehydrogenase is easily oxidised by the electron transport chain. However, the yeast would require additional NADPH for xylose reductase. We investigated which of the key NADPH producing

reactions would be up-regulated to support this proposed additional requirement for NADPH during xylose utilization. The somewhat unexpected results were rationalised by estimating fluxes in central metabolism for both conditions. Finally, MRA was used to answer the fundamental question of how changes in gene expression and in metabolite concentrations, respectively, contributed to the regulation of fluxes.

Materials and methods

Genome sequencing and annotation

The genome of *K. marxianus* UFS-Y-2791 was sequenced on the Illumina HiScanSQ platform to 100-fold coverage at the Onderstepoort Biotechnology Platform, Pretoria, South Africa. Assembly was performed *de novo* using Abyss. Open reading frames were found by Augustus. Putative protein sequences were annotated using one of two methods. For annotation of enzymes (778 genes), creation of a pathway genome database (PGDB) and all subsequent analysis involving metabolic pathways, protein sequences were annotated against the Kegg-Kaas server with default settings on the server, resulting in a list of EC number annotations. Output was subsequently parsed and converted to input for the PathoLogic algorithm of Pathway Tools. For gene set enrichment using Gene Ontology (GO), sequences were additionally annotated against the UniProtKB database on a high-performance computing cluster, using BLASTP. An E-value cut-off of 1E-10 was used for gene set enrichment as was done by others [3]. An additional 73 genes with E-values between 1E-5 and 1E-10 were included in the list of annotated genes and flagged for further annotation. We preferred the rich manually curated SwissProt annotations over automated Trembl annotations, resulting in 68.3% of annotations from *S. cerevisiae*, 17.4% from *K. lactis*, 3.2% and 2.8% from two strains of *K. marxianus*, and the rest from other species. The draft genome, predicted open reading frames, predicted protein sequences and functional annotations are made available in the Supplementary materials (S1 Draftgenome, S1 ORF, S1 Proteins, S1 Table).

Strains and cultivation

All chemicals and fermentation media used in cultivations were obtained from Sigma Aldrich, Seelze, Germany. *K. marxianus* UFS-Y2791 from our University of the Free State MIRCEN yeast culture collection was maintained on YPD agar slants at 4^{o}C. Cultivation was carried out under aerobic conditions in 500 ml shake flasks with 30 ml working volume at 180 rpm. We chose the shake flask format to allow expensive ^{13}C-isotopic tracer studies which would be cost prohibitive in bioreactors. All cultivations were carried out

at 35 °C. The pre-inoculum was incubated in YPD medium for 8 h. The inoculum was prepared by dilution of the pre-inoculum to an OD_{690} of 0.05 in a chemically defined medium and grown at 35 °C for 16 h. The defined medium contained (g l^{-1}): glucose or xylose, 5; $(NH_4)_2SO_4$, 2.5; $MgSO_4 \cdot 7H_2O$, 0.5; $CaCl_2 \cdot H_2O$, 0.03; NaCl, 0.1; citric acid, 0.25 and KH_2PO_4, 10. Filter-sterilised vitamins were added to the autoclaved medium at the following concentrations (mg l^{-1}): biotin, 0.025; calcium pantothenate, 0.5; nicotinic acid, 0.5; *p*-aminobenzoic acid, 0.1; pyridoxine HCl, 0.5; thiamine HCl, 0.5 and *myo*-inositol, 12.5. Trace elements were added according to du Preez and van der Walt [19]. The pH of the medium was adjusted to 5.5. Glucose and xylose were quantified using HPLC. Acetate, ethanol and glycerol were the only fermentation metabolites secreted in significant amounts and were quantified using HPLC. All samples were taken before an OD_{600} of 0.8 was reached.

RNA-seq data generation

Cells from duplicate cultivations were harvested in mid-exponential phase at an OD_{690} of 0.8 by centrifugation at 1 000 × g at 4 °C for 5 min. RNA extraction was carried out according to the RNeasy yeast RNA kit (Qiagen) protocol, including a DNAase step. Ribosomal RNA was removed by the Ribo-Zero rRNA Removal Kit (Illumina) and remaining RNA was sequenced at the Onderstepoort Biotechnology platform in Pretoria, South Africa. Paired-end reads were generated using Illumina HiSeq Next Generation Sequencing. For each of duplicate samples, 3.75 million trimmed paired-end reads were mapped to the UFS-Y2791 genome. Processing of sequencing data was carried out in Galaxy. Reads were trimmed using Trimmomatic [20] and mapped to the genome using TopHat, while CuffDiff [20] was used to test for differential expression. CuffDiff reports p-values as the statistical significance and well as the q-value, which is the p-value after accounting for multiple comparisons [21]. Genes were only considered to be significantly differentially expressed when q-values were below 0.05.

Gene set enrichment and reporter metabolites

For gene set enrichment and all other network constructions and renderings, programs were developed as part of a new software suite for integrative systems biology that we call Reactomica, implemented in the Wolfram Mathematica language. Gene set enrichment scores for GO terms and pathways were calculated similar to that described by Ideker [22] as follows: GO ontologies goslim_yeast.obo and go-basic.obo.txt.m were obtained from the Gene Ontology database [23]**.** The ontologies were converted to graphs using the 'is_a' child-parent mappings. For obtaining the gene set, a depth-first scan was

performed from each GO term and all additional GO terms found from the node of interest were collected to ensure that highly specialised terms would find utility in GO_slim. Mappings from the annotated genes to the GO-terms in the GO 'biological_process', 'molecular_function' and 'cellular_component' attributes of a UniProt entry were used to map from GO terms to genes. Significance q-values could be converted to Z-scores as the negative of the inverse cumulative distribution function and then summed over all genes in the gene set to give the representative statistic for a group of genes as the total Z-score [22]. Random gene sets were generated with bootstrapping (1000 iterations) and total Z-scores calculated. The mean and standard deviation at a variety of gene set sizes were calculated and used to calculate the enrichment score S:

$$S = \frac{Z(\text{total,Test}) - \text{Me}\quad(Z,\text{Background})}{\text{Standard deviation}(Z,\text{Background})} \qquad (1)$$

For pathway gene set analysis, MetaCyc pathways were used from the BioCyc pathway genome database constructed for this strain using Pathway Tools, which is based on the MetaCyc database. Reporter metabolite enrichment scores were calculated in the same manner as described above and according to Patil and Nielsen [13]. For reporter metabolites, the background mean and standard deviation of random genes sets were rather generated by sampling enzyme-encoding genes only, as opposed to the complete gene set, since a large fraction of the differentially expressed genes were metabolic, generating a higher random background enrichment.

Pathway maps

To explore the metabolic response, metabolic pathway maps were created from MetaCyc pathways and RNA-seq data were mapped using various colouring schemes with Reactomica, harnessing automated hypergraph map layout and manual override. The initial linkage between genes and reactions were made by the Pathway Tools algorithm PathoLogic, with genomic annotations from the Kegg-Kaas annotation server. For the Log_2(fold change) colouring scheme, in the case of more than one enzyme that could perform the same function, the largest fold change in expression was used for the colour rendering. Additional gene-reaction linkages were made from a UniProt BLASTP annotation. Subsequent compartmentalisation made use of the GO 'cellular_component' ontology terms. For purpose, a mapping was built into Reactomica that maps GO 'cellular_component' ontology terms to subscellular compartments.

Pathway-to-pathway networks

Pathways were clustered by the number of metabolites in common to generate a scoring matrix. The number of metabolites in common was normalised by the smaller of the two metabolite sets of a pathway and a threshold was selected for including a mapping that resulted in optimal rendering. Orphaned pathways were clustered together.

Molecular networks

Molecules were clustered by the simple similarity criterion of string matching of SMILES strings of each compound in the PGDB using the edit distance. Only the closest match was included as a mapping. The method is similar to that done by Barupal et al. [24].

Reporter metabolite-enzyme networks

The metabolic network was converted from a hypergraph into a graph in which only the interactions between differentially expressed enzyme genes and enriched reported metabolites were retained.

Differential Flux Analysis

The aim was to approximate the fluxes on glucose and xylose, sufficient for approximation of MRA values. FBA was used which included optimisation of biomass formation, and measured specific sugar consumption rate and specific ethanol, acetate and were included to constrain the flux solution. The model was constructed in Reactomica using reactions from MetaCyc for which representative genes were found in the genome annotation of *K. marxianus* UFS-Y2791. Some additional reactions were defined such as the combined electron transport and oxidative phosphorylation reaction. The biomass reaction was adapted from Fischer et al. [25]. The flux model and parameters is provided in S5 Table. FBA was implemented in Reactomica and the method of FBA was reviewed elsewhere [15]. FBA uses linear optimisation, maximising the biomass formation reaction, constrained by the reaction stoichiometry, reversibility constraints, uptake rate of nutrients (glucose or xylose) and production rates (ethanol and acetate).

Metabolic Regulation Analysis

The metabolic regulation of a flux can be separated into a metabolic component ρm and a hierarchical component ρh. The two levels of regulation are combined in the relation below [18].

$$1 = \rho h + \rho m \quad (2)$$

The metabolic component ρm models the contribution to regulation that changes in metabolite concentrations make and is described by

$$\rho m = \sum_X \frac{d\ ln(v)}{d\ ln(X)} \cdot \frac{d\ ln(X)}{d\ ln(J)} \quad (3)$$

where *J* is the flux though that enzyme, modelled by the rate equation *v*, and affected by changes in the concentration of metabolite *X*. The hierarchical component ρh models the effect of changes in maximal activity of the enzyme, which is usually linearly dependent on the enzyme concentration *e* and described by

$$\rho h = \frac{d\ ln(e)}{d\ ln(J)} \quad (4)$$

The hierarchical component could thus be experimentally determined by measuring a difference in maximal enzyme activity. By using equation (2) and substituting with the experimentally determined *ρh, ρm* could also be calculated. Though maximal activity is affected not only by the protein concentration and post-translational modifications, it is reasonable and practical to use changes in transcript levels obtained from RNA-seq instead of maximal activities or protein levels for an estimated MRA. For MRA, gene expression changes were considered significant if at least one gene (among paralogs) was considered significant from q-values. Transcript fold changes were calculated from total transcript abundances for all genes mapping to a reaction, in case of isozymes resulting from paralogs or multi-functional proteins. In this manner, the method is robust to potential errors in annotation of paralogs since central metabolic genes are known to have on average higher transcript abundance in comparison with the vast majority of other genes. The fold changes of individual minor isozymes with very low expression levels also cannot dominate the analysis.

Results

In glucose medium, respiro-fermentative growth was observed, even though fully aerobic conditions were ensured by the small working volume in a large flask, vigorous shaking, and sampling at low OD_{600} values, with ethanol, glycerol and acetate as fermentation products. In xylose medium, fermentation products were absent, with an apparently purely respiratory metabolism. The maximum specific growth rate in

glucose medium was approximately 0.8 h^{-1}, while the maximum specific growth rate in xylose medium was approximately 0.35 h^{-1}. All graphs and other renderings were generated using a new software suite for bioinformatics and integrated systems biology, termed Reactomica, which was developed in-house. Interactive datasets are provided in Computable Document Format (.CDF) as supplementary materials and are viewable using the free Wolfram CDF player, which can be downloaded from the Wolfram website [26], or by using Mathematica. High quality differential RNA-seq datasets were generated on the Illumina HiSeq platform to a high read depth. Figs 1 and 2 show the distribution of the data. Throughout, "up-regulated" refers to genes statistically up-regulated in a xylose medium compared to the condition with glucose as the carbon source, with a q-value below 0.05, as reported by CuffDiff (see 'Materials and methods'). Out of the total of 4 953 putative genes analysed, 329 were up-regulated and 251 down-regulated. Supplementary file S1 Table provides all expression values, differential expression statistics and UniProt annotations of all genes.

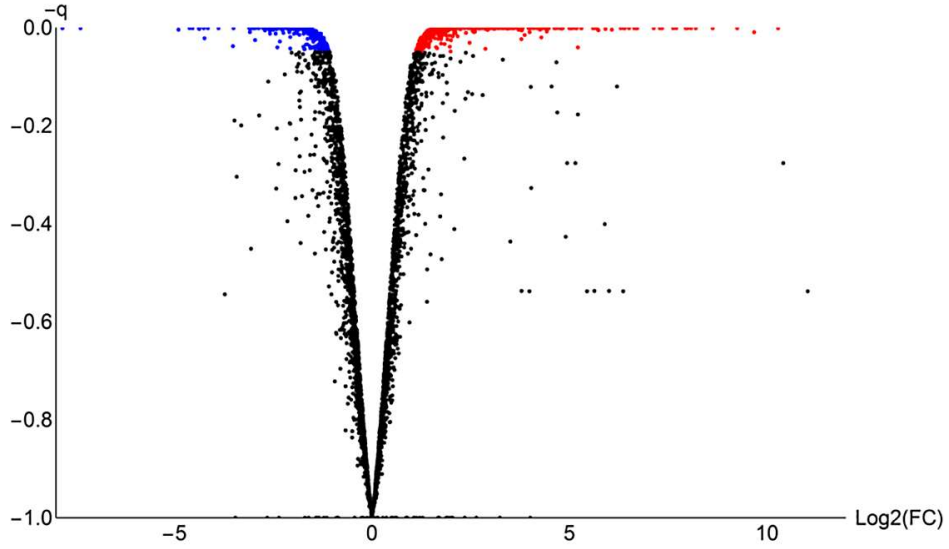


Figure 1. Volcano plot of RNA-seq data. FC: fold change. –q: the corrected p- values after taking multiple comparisons into account, as performed by CuffDiff. Red: up-regulated on xylose. Blue: down-regulated on xylose. Black: constitutively expressed. Statistical procedures were performed in CuffDiff.

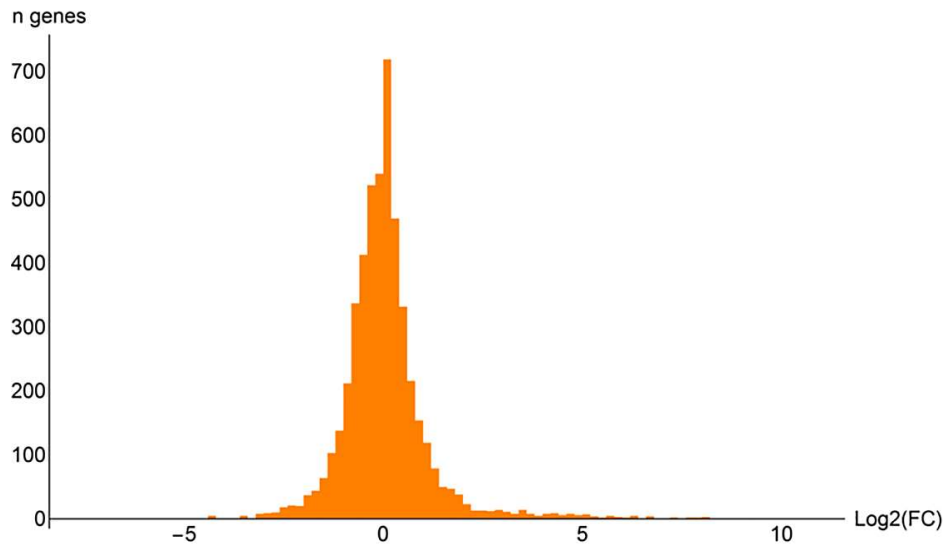


Figure 2. Histogram of RNA-seq log_2(fold change) values. Statistical procedures were performed in CuffDiff.

The role of rich medium versus defined medium in the xylose response

We compared results from reference [3], which reported 36 genes in strain DMKU 3-1042 that were up-regulated only in the aerobic xylose-containing rich medium and not in any of the other three conditions tested (see S1 Table) (we refer to this gene set as the xylose-present/glucose absent unique gene set). Differential expression of 13 genes were correlated with the reference data (up-regulated in UFS-2791) and should be specific to the xylose response (or the absence of glucose), independent from the background medium, and present across strains. Notably, POT1, POX1 and FOX2 of peroxisomal β-oxidation were dramatically up-regulated in both datasets. Carnitine O-acetyltransferase (CAT2), associated with inter-compartmental transport of fatty acids, was also strongly up-regulated in both sets, supporting increased capacity for lipid catabolism (see a detailed discussion later). ICL2 and CIT3 of the 2-methylcitrate cycle was also strongly up-regulated in both experiments (see a detailed discussion later). Aldehyde dehydrogenase (ALD4), glucokinase 1 (GLK1) and dicarboxylic amino acid permease (DIP5) are the other enzymes functioning in central metabolism. Two regulatory proteins Ty transcription activator (TEC1) and the G2/mitotic-specific cyclin-4 (CLB4) were also among these. The rest of the genes in this list are not well-characterised. These include stationary phase protein 4 (SPG4), putative metabolite transport protein ywtG and uncharacterized membrane protein YMR155W.

Thirteen genes were uncorrelated with the reference data (constitutive in UFS-2791) (YPR011C, HSP12, YHL008C, POP6, NCE103, YLL032C, NCS2, GAS1, TOS1, PDR5, MYO1, EIS1, YDR134C) and are thus specific to either the strain or the rich medium. Most of these are uncharacterised proteins.

Finally, two genes were anti-correlated with the reference data (down-regulated in UFS-2791). These are the ammonia transport outward protein 3 (ATO3) and dihydroxyacetone kinase 1 (DAK1). The exact functions of ATO1, ATO2 and ATO3 are not currently known, apart from their possible involvement in mitochondrial retrograde signalling and ammonia production during starvation. It seems that ATO3 requires rich medium containing amino acids to be up-regulated, as was also observed in *S. cerevisiae* [27, 28]. In summary, several genes were highlighted as up-regulated both in our data and the xylose-present/glucose absent unique gene set from reference [3], in particular, the peroxisomal beta-oxidation and the supporting fatty acyl transporters. The roles of a number of other genes are unclear from this preliminary analysis, however.

Gene Ontology

A total of 1 611 GO terms were included in the UniProt annotations of all proteins in the genome (see Materials and methods). To obtain an initial overview, the concise 'GO_slim' yeast from the Gene Ontology Consortium was used. GO term enrichment was performed separately for 'cellular_component', 'molecular_function' and 'biological_process' components of 'GO_slim' and the enrichment scores were used to render maps (see 'Materials and methods'). There is currently no generally accepted cutoff for interpreting significance of an enrichment value [14]. For the 'cellular_component' ontology, assuming a score cut-off at 1.64 ($p = 0.05$), five terms were considered to be significantly enriched (Fig 3 and S2 Table). These are 'extracellular region', 'peroxisome', 'plasma membrane', 'membrane', 'mitochondrion' and 'cell wall'. 'Plasma membrane' is a subset of 'membrane' in the GO-slim ontology. It is striking that the peroxisomal genes were up-regulated (14 out of 32 genes) and only one down-regulated. The 'extracellular region' exhibited mostly up-regulation (22 out of 80 genes) with seven down-regulated genes. Of the 269 plasma membrane genes, 31 were up-regulated and 22 were down-regulated. In the mitochondrion the number of up and down-regulated genes were approximately equal (29 and 23, respectively out of 308 genes). The cell wall genes showed also mostly up-regulation, with 17 out of 80 genes up-regulated and four down-regulated.

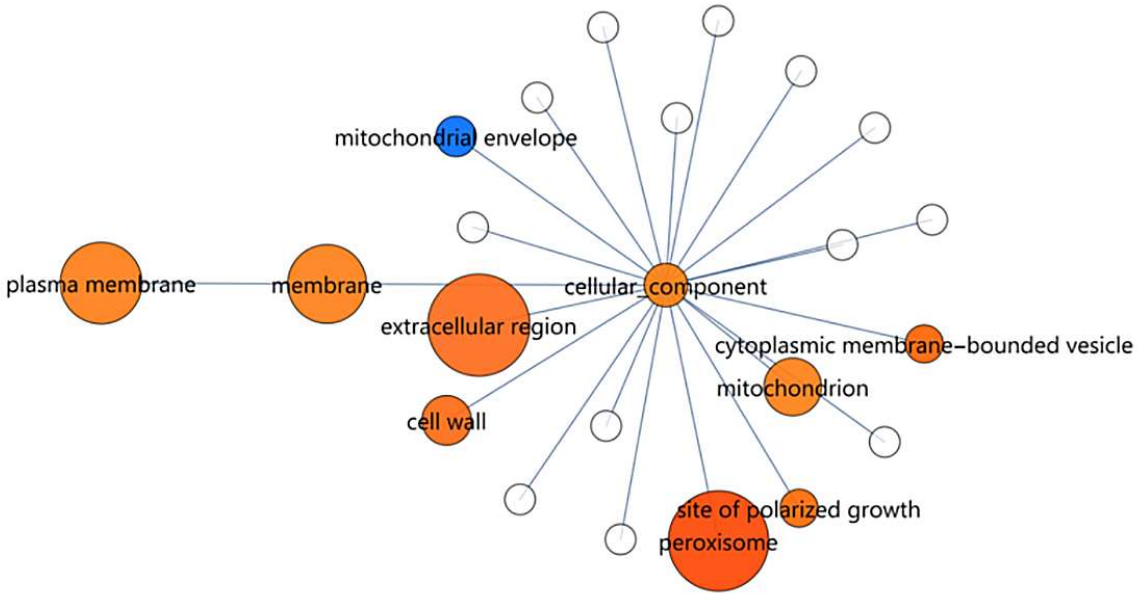


Figure 3. Gene set enrichment map of RNA-seq data using the 'GO_slim' yeast 'cellular_component' ontology. Size indicates the enrichment score of a gene set. Colour indicates the up/down direction of regulation: Red, up; blue, down. Brightness rises with the fraction of genes that are regulated in the dominant direction (up or down). Only nodes larger or equal than 'cell wall' are significant.

In the 'molecular_function' ontology, only three terms are regarded as significant. 'Oxidoreductase activity', 'kinase activity' and 'peptidase activity' (Fig 4 and S3 Table). 'Oxidoreductase activity' is the most highly enriched term in the 'GO_slim' gene sets (enrichment score = 4.49) and also in the complete GO enrichment (score = 13.2). It is notable that the redox balance, which is regulated by oxidoreductases, is one of the key considerations in the ability to utilize pentoses by yeasts, as it usually involves a requirement for NADPH and the additional generation of NADH, as should also be the case with *K. marxianus* since it does not possess a xylose isomerase.

The 'biological_process' ontology is rich in providing a framework for exploring the regulation of cellular processes in response to different carbon sources. The 'carbohydrate transport' gene set was the most significantly altered, with 21 out of 65 transporter genes up-regulated and 7 down-regulated (Fig 5 and S4 Table). The 'transmembrane transport' gene set displayed a more equal up vs down-regulation, whereas 'peroxisomal organization' genes had exclusively up-regulated genes. The latter feature is shared with the 'peroxisome' gene set in the 'cellular_component' ontology, indicative of a general up-regulation of peroxisomal components and activities.

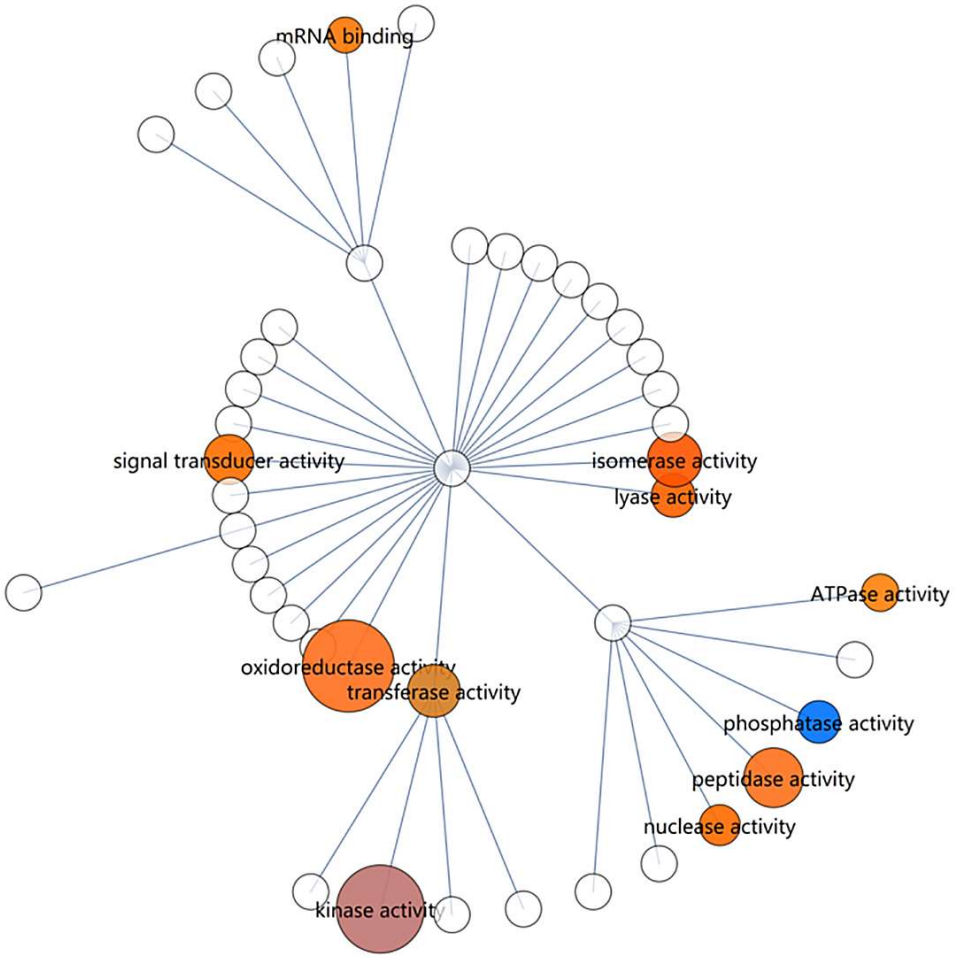


Figure 4. Gene set enrichment map of RNA-seq data using the 'GO_slim' yeast 'molecular_function' ontology. Size indicates the enrichment score of a gene set. Colour indicates the up/down direction of regulation: Red, up; blue, down. Brightness rises with the fraction of genes that are regulated in the dominant direction (up or down). Only nodes larger or equal than 'peptidase activity' are significant.

The 'amino acid transport' gene set also indicated differential expression. The gene set of 'ion transport' was also enriched, but is closely related to 'amino acid transport'. Also significant is 'carbohydrate metabolic process', which is better interpreted in terms of metabolic pathways. Although 'histone modification' and 'proteolysis involved in cellular protein catabolic process' both have significant scores, they only have one and four genes, respectively, in the gene sets, and they were therefore not interpreted for further investigation.

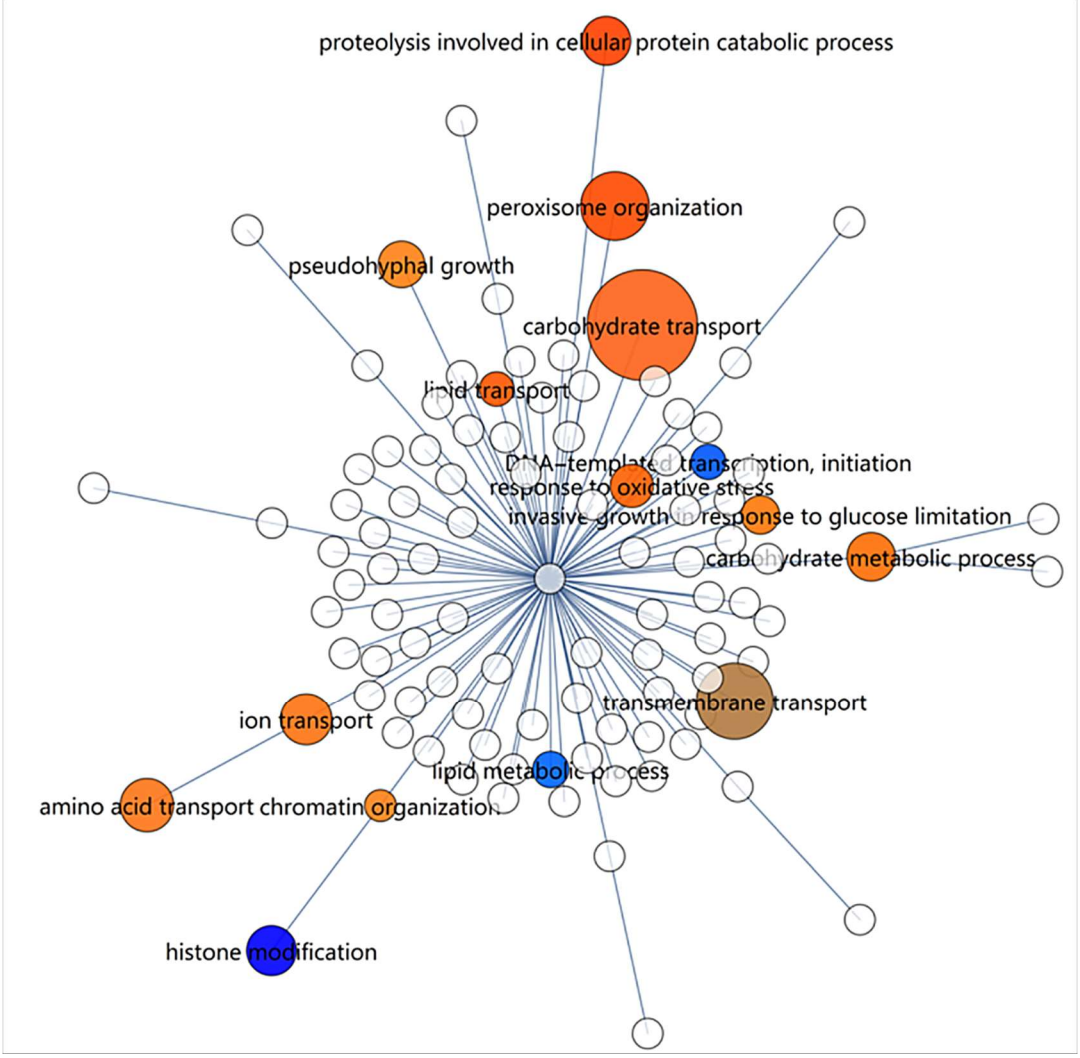


Figure 5. Gene set enrichment map of RNA-seq data using the 'GO_slim' yeast 'biological_process' ontology. Size indicates the enrichment score of a gene set. Colour indicates the up/down direction of regulation: Red, up; blue, down. Brightness rises with the fraction of genes that are regulated in the dominant direction (up or down). Only nodes larger or equal than 'proteolysis involved in cellular protein catabolic process' are significant.

In summary, peroxisomal organisation and metabolism were clearly and strongly up-regulated on xylose, consistent with a previous observation [3] that at least fatty acid β-oxidation was up-regulated in a complex medium that likely contained small amounts of lipids, which were absent in the present study where a chemically defined medium was used. Secondly, a strong effect was seen on carbohydrate transporters, which was consistent with the experimental setting. Some of these differentially expressed putative sugar transporter genes, such as the various *HGT1* homologs of *K. lactis*, were up-regulated as high as 1242-fold (see S4 Table). Some of these may encode for xylose transporters. As sugar transporters

are highly similar, additional annotation of this group is required before more conclusions regarding sugar transport may be drawn. Thirdly, the extracellular region was affected. These are proteins that are secreted into the medium, such as lytic enzymes or receptors that sense a new environment, as well as the mating genes. Some of these gene sets are explored in more detail in the supplementary text S1 Text.

It is interesting to note that although many genes in the mitochondrion were differentially regulated, these genes are not the main enzymes of the TCA cycle, electron transport or oxidative phosphorylation. Thus, there is no response suggesting adaptation to a change in the internal energy charge. As the maximum specific growth rate in xylose medium was less than half of that in glucose medium, whereas the cell sizes were comparable, one might expect differential regulation of cell cycle progression genes, or even biomass formation-specific genes like those encoding for ribosomal proteins. There was a notable absence of such gene sets in the enrichment analysis. This is also reasonable since cell cycle control signalling proteins are mostly kinases, which are controlled by post-translational modifications. Most genetic responses observed in the data thus deal with utilisation of alternative substrates. Further, there was only a weak response in the oxidative stress genes, as was expected from aerobic cultivation. From the genes annotated with 'response to oxidative stress' (PRDX5, CTT1, CCP1, YFH1, DCS1, HMX1, SVF1, YDR222W), only peroxiredoxin-5 (PRDX5, mitochondrion, cytoplasm, peroxisome, 3.9-fold up-regulated) and catalase T (CTT1, cytoplasm, 2.4-fold up-regulated) were moderately differentially regulated, resulting in an enrichment score of 1.19, which was insignificant against the background enrichment.

The data indicated several interesting aspects regarding sugar utilisation, including up-regulation of pathways for utilisation of galactose, xylose and arabinose (see S5). Several enzymes with β-glucosidase activity were found in the annotation (see S1 Text), whereas only one enzyme was both extracellular and significantly up-regulated (48.8-fold), namely cellobiase (EC 3.2.1.21). Cellobiase is not a typical cellulase that can depolymerise cellulose, as it functions to hydrolyse the disaccharide cellobiose. *K. marxianus* UFS-Y2791 also does not possess typical secreted xylanases, proteases, peptidases or lipases, which would allow a microorganism to thrive on plant matter or even attack a live plant. Instead, the strain possesses the inulinase gene *INU1* which hydrolyses the fructan inulin or the disaccharide sucrose. The *INU1* gene was dramatically up-regulated 91-fold and abundantly expressed in xylose medium.

The pheromone signalling system involved in sexual reproduction, as well as several other genes involved in the conjugation process as well as in invasive growth, were also up-regulated (see 'extracellular region'

in S2 Table). In the presence of xylose (or absence of glucose) there is thus a response to physiologically adapt to an invasive lifestyle and utilise other sources of nutrients (xylose, arabinose, inulin, cellobiose and amino acids), as would be found in the natural plant environment. A long-term survival strategy (sexual reproduction) is also activated in the more nutrient-poor condition.

Most metabolic pathways for amino acid synthesis were down-regulated, as was expected at the lower growth rate; μ_{max} values of 0.8 and 0.35 were recorded on glucose and xylose, respectively. The well-known importers of ammonia were constitutive (MEP3) or down-regulated (MEP2, 2.8-fold).

Global metabolic response elucidated by pathway-to-pathway networks

To capture the global metabolic response to growth on xylose in a single view, an innovative pathway-to-pathway network was constructed by clustering pathways together by their common metabolites (Fig 6, S1 Network and S1 Table). The wide down-regulatory profile of amino acid metabolic pathways is visible, where all amino acid biosynthetic pathways were shut down, with only glycine biosynthesis III having some element of up-regulation. Similarly, most amino acid catabolic pathways were shut down with only glutamate, valine, tyrosine, phenylalanine, tryptophan and methionine degradation III having some up-regulated genes.

Fatty acid oxidation, peroxisomal ethanol degradation and mixed acid fermentation, which cluster together, were mostly up-regulated. Close by in the network is glycolysis, which was down-regulated. A moderately increased capacity in the glycerol-3-phosphate shuttle was evident here, with the mitochondrial glycerol-3-phosphate dehydrogenase *GUT2* gene 4.4-fold up-regulated. Note that the up-regulated glycerol-3-phosphate shuttle functions independently from down-regulated glycerol production.

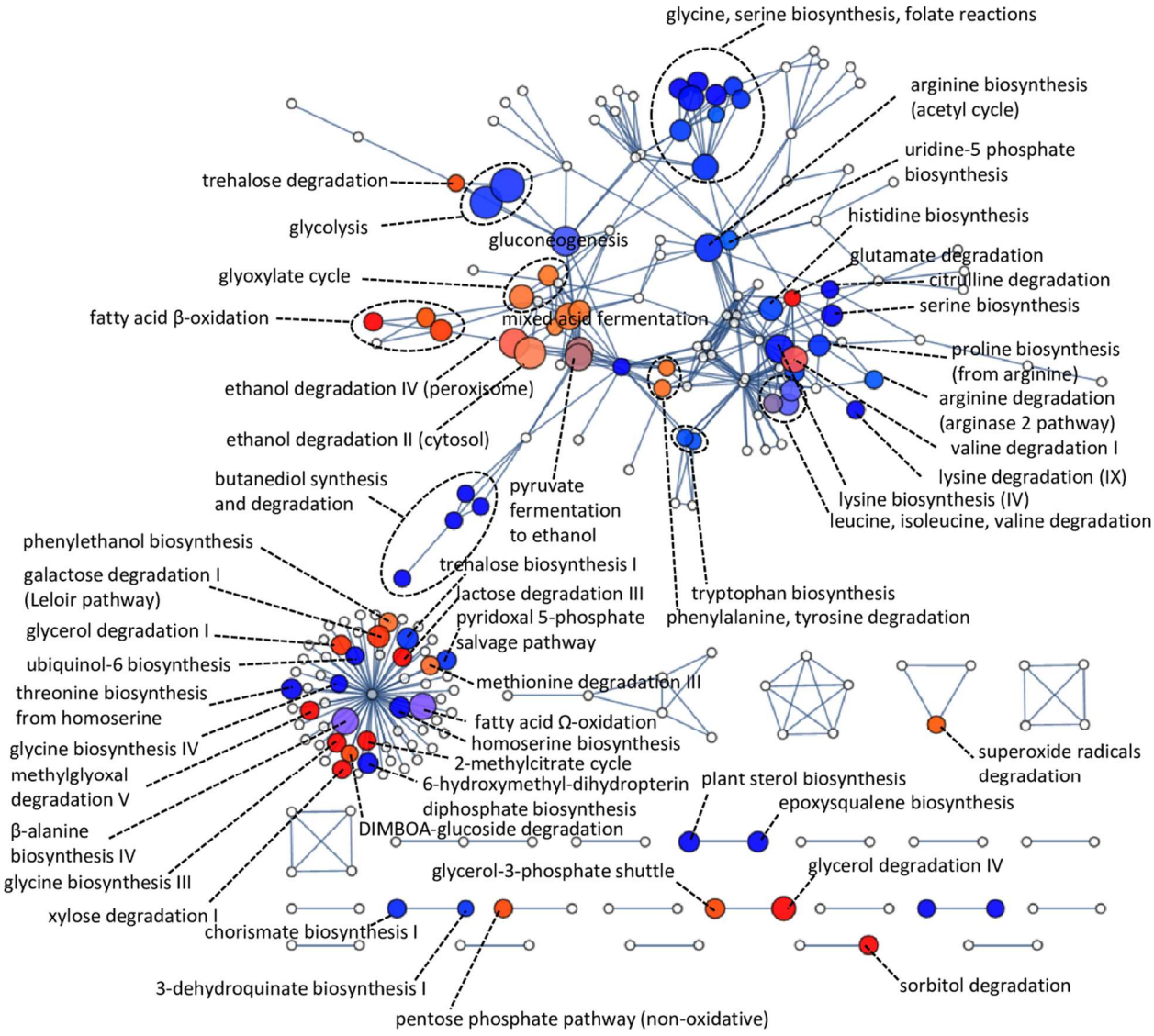


Figure 6. Gene set enrichment map of RNA-seq data using the pathway-to-pathway network. Size indicates the enrichment score of the gene set representing each pathway. Colour indicates the up/down direction of regulation: Red, up; blue, down. Brightness indicates uni-directionality of regulation. The 'Self' cluster on bottom-left includes pathways not mapped to other pathways since their intersections scores were below a threshold.

Reporter metabolites

Reporter metabolite enrichment was performed to reveal those metabolites around which significant differential expression of enzymes took place. These compounds could be interpreted as those from which there was *(a)* a marked change in capacity for their utilisation or production between conditions, *(b)* to have required a significant degree of regulation by their neighbouring enzymes to establish homeostasis in the different environment, or *(c)* to have a different concentration predicted between conditions where it may be a candidate as a signalling molecule. The top eight reporter metabolites of highest enrichment were 3-phosphoglycerate, acetaldehyde, NADH, glyceraldehyde-3-phosphate, β-alanine, NAD^+, glycerate and ethanol, in this order (S1 Table). NADH, NAD^+, acetaldehyde and ethanol are involved in redox metabolism and thus support the high gene set enrichment score of the GO term 'oxidoreductase activity' (GO:0016491). It was interesting to note that the enrichment score of ATP was among the lowest against the background and considered insignificant. Also, glutathione and its reduced form, which are known to be involved in redox metabolism, were not significantly enriched, suggesting no severe oxidative stress. Oxidative stress in *K. marxianus* is seemingly more important under oxygen limiting conditions, as imposed by static and high-temperature conditions [3].

Molecular networks

To determine whether the reporter metabolites represented any distinct molecular structural groups, a molecular network of the reporter metabolites was reconstructed using a simple molecular structure similarity matching protocol. This approach could identify some co-regulated groups of structurally related molecules that were not evident from pathways-based analyses. Fig 7 shows a number of clusters of enriched reporter metabolites grouped by their molecular structures (see interactive file S2 Network for molecular structures and annotations). Groupings of CoA-conjugates, long-chain and short-chain fatty acids are representative of increased catabolic activity of β-oxidation and activation steps. Up-regulated sugar clusters corresponded to those found in reporter metabolite-enzyme networks, but were represented in a clearer fashion. Noteworthy is that effectors of trehalose containing sugars and sugar lipids were down-regulated. The three-carbon molecules in central carbon metabolism, which are close together in the network including 3-phosphoglycerate, 2-phosphoglycerate, glyceraldehyde-3-phosphate and glyceraldehyde, were strongly enriched and their effectors down-regulated. This mapping is useful as a concise representation of potentially all metabolites in a cell, grouped by their structures. The method would especially be useful for visualising metabolomics datasets of which the link to reactions and

pathways is not yet clear. A better separation of clusters is obtained in mapping larger metabolites as found in secondary metabolism.

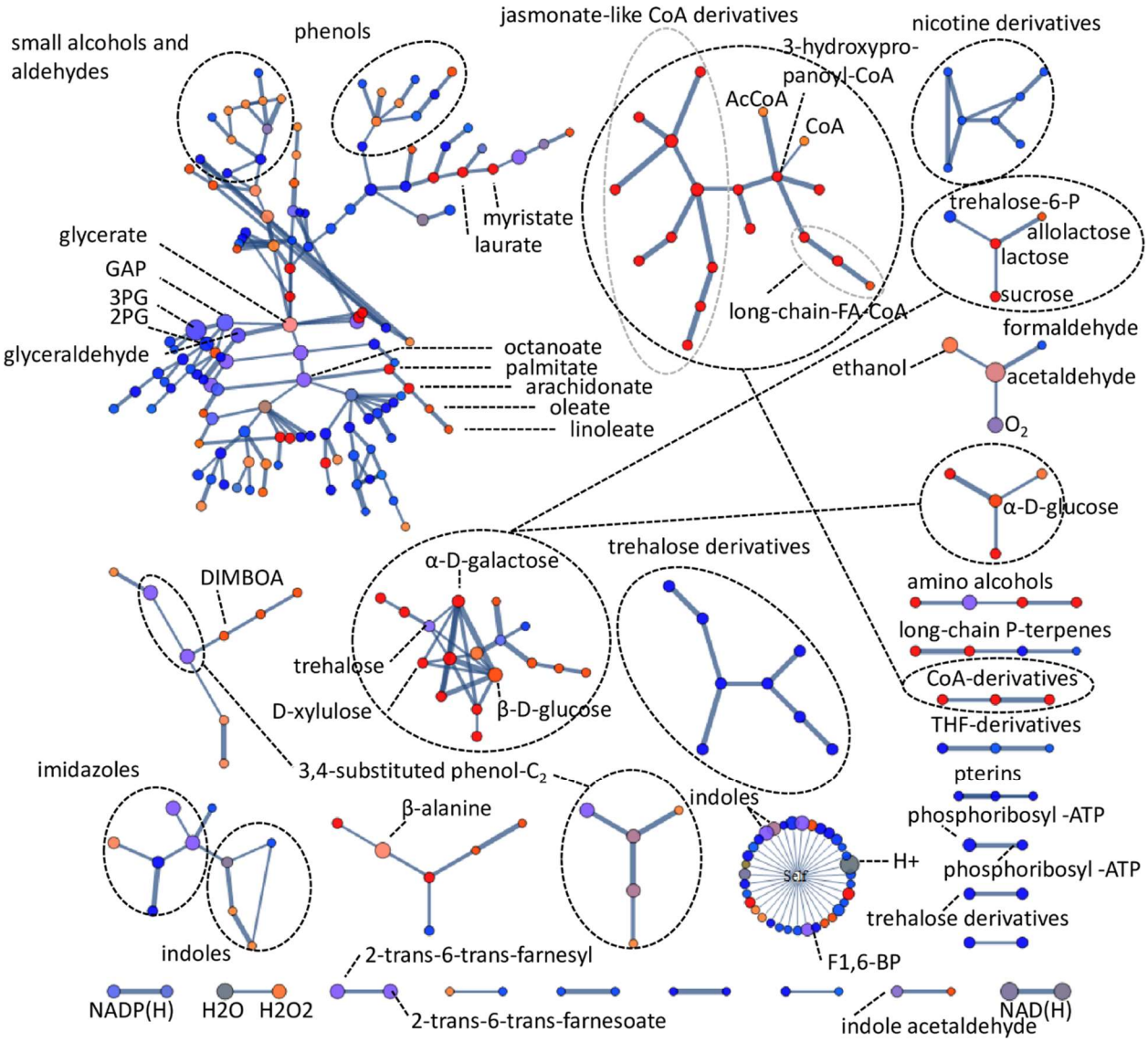


Figure 7. A molecular network of all reporter metabolites enriched at or above enrichment score >=1.64 (p = 0.05). Mapping between compounds was performed with a local string matching procedure based on the SMILES string of each compound. Scores were normalised to the string size for the longest of the two molecules in a pairwise comparison. Edge weights represent normalised similarity scores. The "Self" node maps all compounds with an insufficient normalised similarity score to other compounds (<0.4). Size indicates the enrichment score of a gene set. Colour indicates the up/down direction of regulation: Red, up; blue, down. Brightness indicates uni-directionality of regulation.

Key enzymes that may affect metabolite pools

Combining both the reporter metabolites and the enzymes by which they are regulated into an enzyme-reporter metabolite network, effectively reveals the hotspots of metabolic regulation as well as the key players in regulation. Including all interactions with enriched compounds (enrichment score > 1.64, q < 0.05) was not feasible for a detailed investigation (1352 interactions). Instead, we extracted the interactions containing nodes with enrichment scores above 3.0 (Fig 8, see interactive file S3 Network).

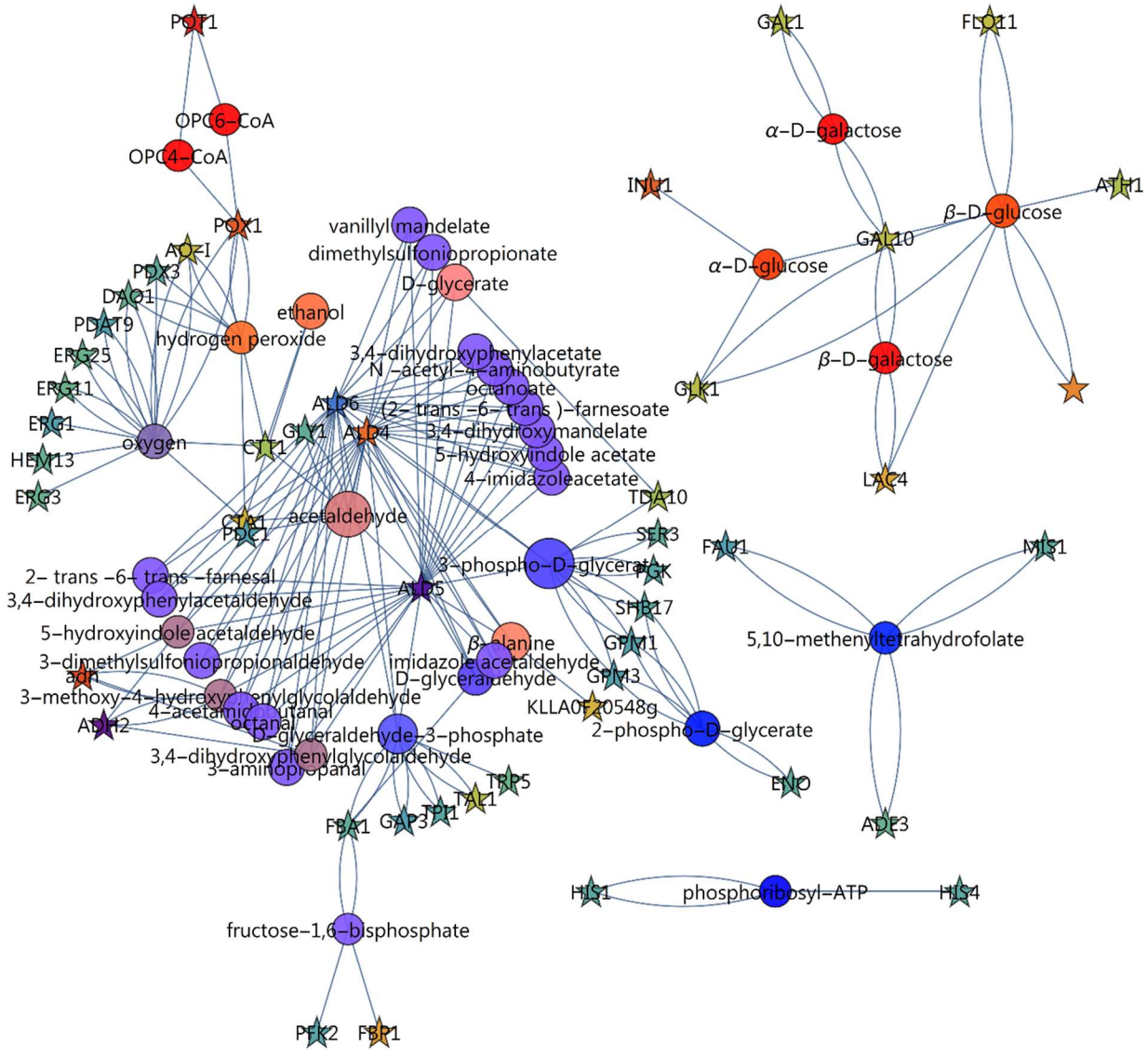


Figure 8. Reporter metabolite-enzyme network, capturing enriched reporter metabolites and the enzymes that affect them. Only reporter metabolites with enrichment values above 3.0 were included and only the differentially expressed genes from RNA-seq with q-values (corrected p-values) below 0.05. For reporter metabolites (circles), size indicates the enrichment score of a gene set and colour indicates the up/down direction of regulation. Brightness

indicates uni-directionality of regulation. For enzymes (stars), colour indicates the up/down direction of regulation based on the $\log_2$(fold change) scheme. Reporter metabolite enrichment values are represented in S1 Table. For full information on gene names, the interactive file S3 Network or the corresponding annotations in S1 Table, using the "Gene names (primary)" column, may be consulted.

A major network is evident in which NAD^+/NADH, oxygen and the aldehyde dehydrogenase have a strong involvement. A second subnetwork involves sugar metabolism and glycosylation. A third revolves around one-carbon metabolism. The subnetwork of NAD^+ (enrichment score = 4.2) was extracted (Fig 9, left, see interactive file S4 Network), which reveals a number of genes involved with biosynthesis being down-regulated.

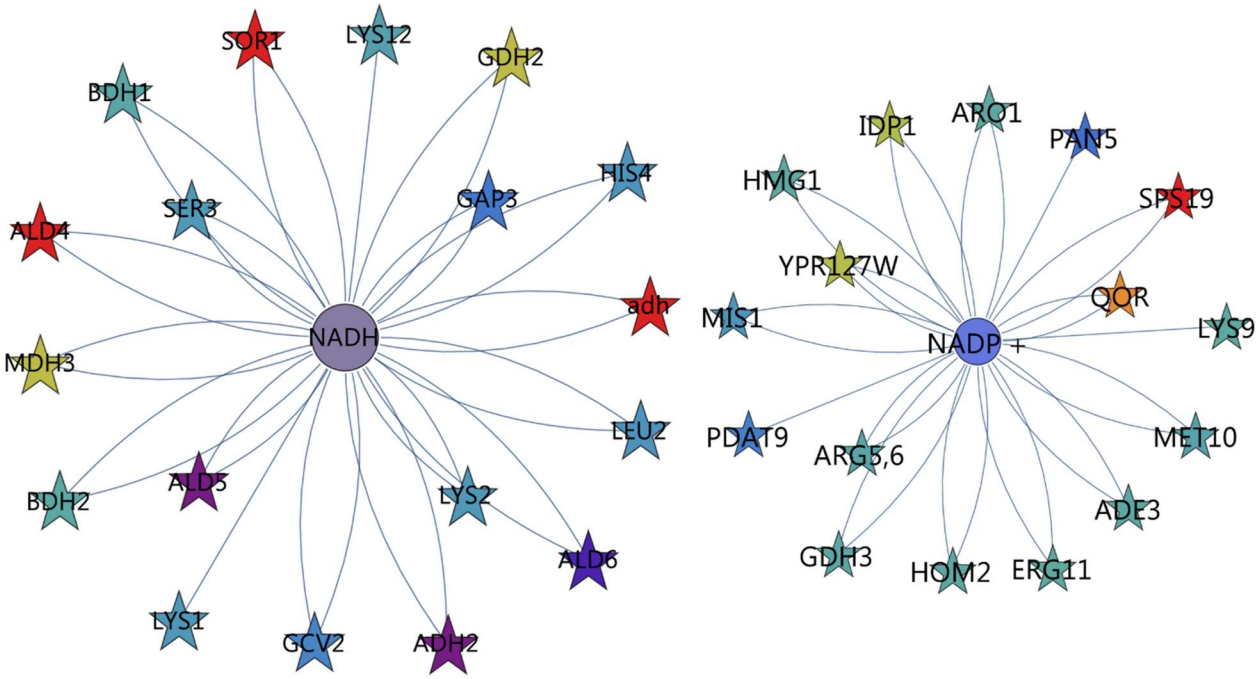


Figure 9. Enzyme-metabolite interaction network around redox cofactors. Left: NADH; Right: $NADP^+$. For reporter metabolites (circles), size indicates the enrichment score of a gene set and colour indicates the up/down direction of regulation. Brightness indicates uni-directionality of regulation. For enzymes (stars), colour indicates the up/down direction of regulation based on the $\log_2$(fold change) scheme. For full information on gene names, the interactive file S4 Network or the corresponding annotations in S1 Table, using the "Gene names (primary)" column, may be consulted.

In addition, a large contribution to the enrichment score is made by the aldehyde dehydrogenases *ALD4*, *ALD5* and *ALD6*, by the alcohol dehydrogenases annotated as *ADH2* and *ald*, and by sorbitol dehydrogenase *SOR1*.

Since *K. marxianus* relies on an NADPH dependent xylose reductase, it was expected that NADPH would be enriched (enrichment score = 2.6). The enzymes directly affecting NADPH are explored in Fig 9 (right, see interactive file S5 Network). We anticipated up-regulation of major NADPH producing enzymes to supply reducing power for xylose utilisation by xylose reductase. Instead, only three enzymes directly involved with NADP(H) were up-regulated. *IDP1* (mitochondrial NADP-specific isocitrate dehydrogenase) was only moderately up-regulated (2.2-fold). A more significantly up-regulated enzyme, mitochondrial quinone oxidoreductase (*QOR,* 9.9-fold up-regulated) reduces 1,4-benzoquinone, as was shown by cloning this gene from *K. marxianus* in *E. coli* [29]. It is not part of the major catabolic processes in central metabolism, however. Another is *YPR127W*, a putative pyridoxal reductase that functions to degrade vitamin B6 and involved with multidrug resistance and not central metabolism [30]. In the oxidative pentose phosphate pathway, which is assumed to be the main generator of NADPH in most species, *GND1* (6-phosphogluconate dehydrogenase) and *SOL1* (6-phosphogluconolactonase) were, however, constitutively expressed. Another enzyme, *SPS19* (peroxisomal 2,4-dienoyl-CoA reductase), was strongly up-regulated at 269-fold. It functions to reduce double bonds to facilitate β-oxidation of unsaturated fatty acids in the peroxisome [31], another indication that peroxisomal metabolism was strongly differentially regulated.

Central carbon metabolism

To explore the metabolic response of the central carbon metabolism, metabolic pathway maps were created from MetaCyc pathways and RNA-seq data were mapped using various colouring schemes. Figs 10 and 11 show total transcript abundances mapped to reactions (see also S1 Pathway). It is evident that on glucose, the central glycolytic route from glucose to ethanol is highly expressed. In xylose medium, transcript abundance is less pronounced in glycolysis and ethanol production, whereas PPP, the pyruvate dehydrogenase bypass and β-oxidation enzymes increase in transcript abundance.

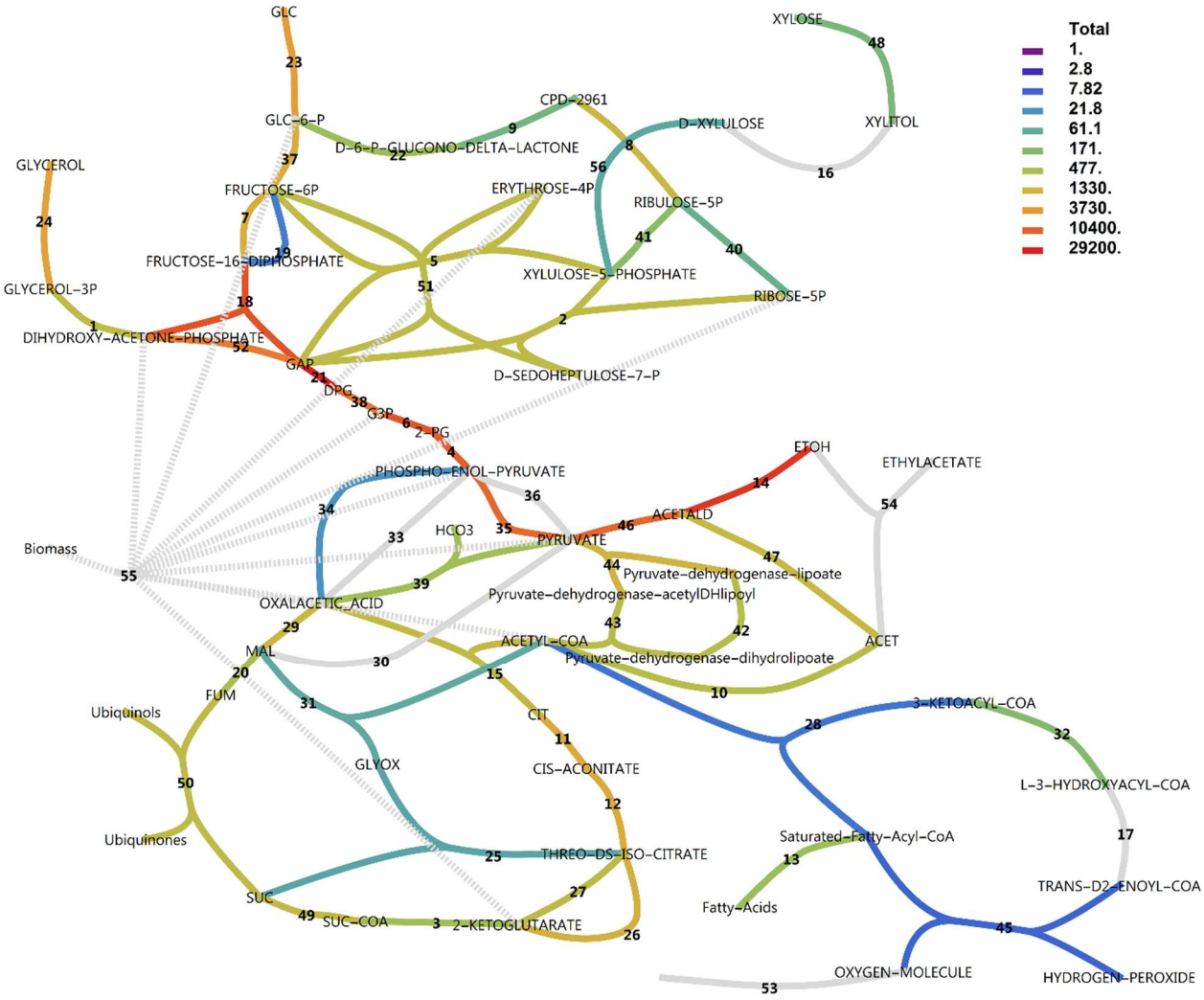


Figure 10. Total transcript levels in central metabolic pathways with glucose as the carbon source. Transcript levels for all genes catalysing a reaction were summed. Note the logarithmic scale. Dark grey indicates genes present and constitutively expressed. Light grey indicates genes not found in annotation or combined reactions. For reaction names, see S1 Pathway.

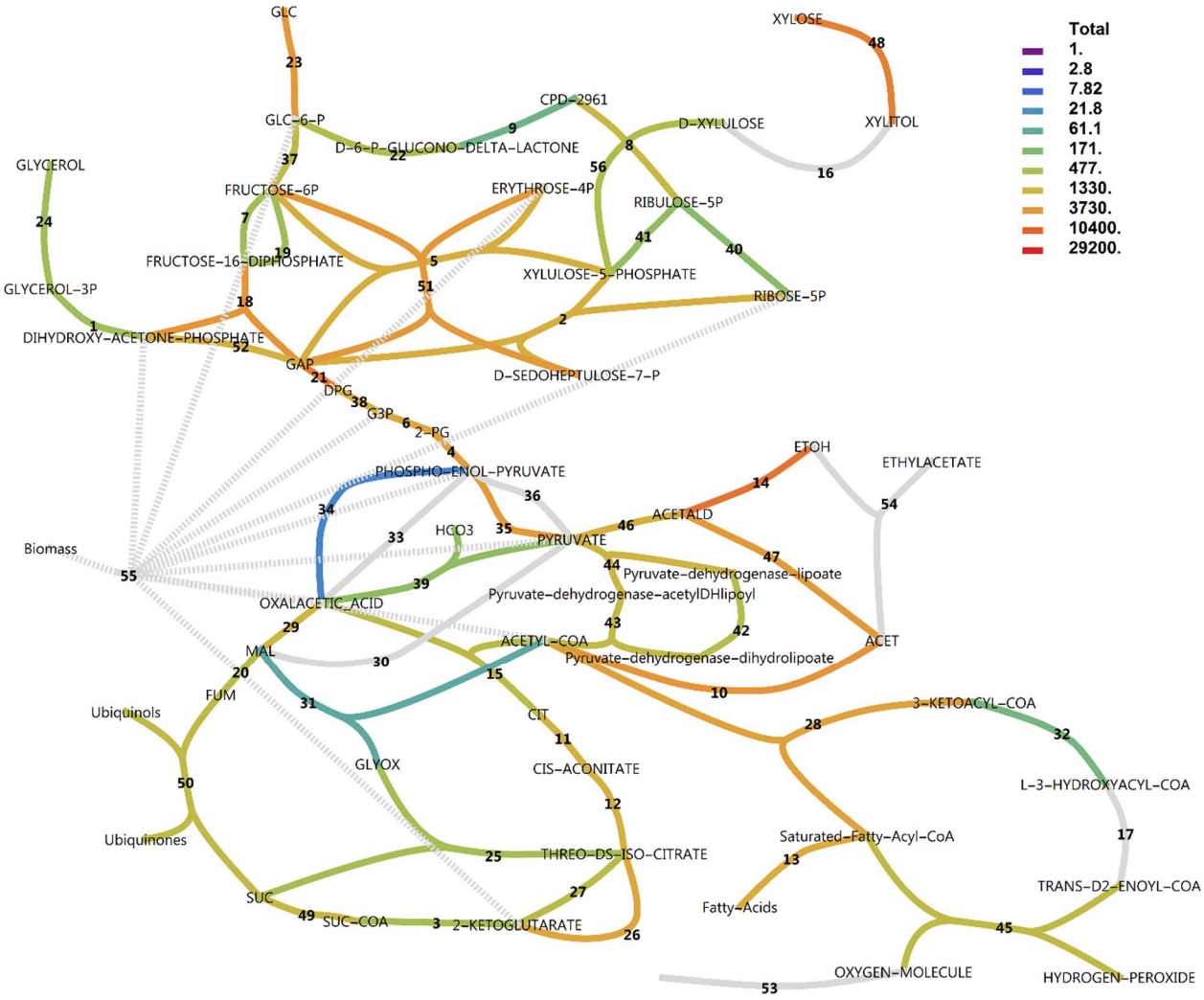


Figure 11. Total transcript levels in central metabolic pathways with xylose as the carbon source. For reaction names, see S1 Pathway**.**

An interesting differential expression pattern is evident in Fig 12. Consistent with the experimental setting, the NADPH-dependent D-xylose reductase gene *XYL1* was drastically up-regulated on xylose (48.8-fold) with very high gene expression levels on xylose. Xylitol dehydrogenase was absent from the annotation. However, it has been demonstrated that sorbitol dehydrogenase *SOR1* can act as a xylitol dehydrogenase in *S. cerevisiae* [32]. Significant up-regulation of *SOR1* by 208-fold supports this function. Xylulokinase *XKS1* was also 11.3-fold up-regulated. Transaldolase *TAL1* of the non-oxidative pentose phosphate pathway (PPP) and ribose-5-phosphate isomerase *RKI1* were moderately up-regulated (4.8 and 2.6-fold). We did, however, not observe up-regulation of any enzymes in the oxidative branch to support additional

NAPDH production for xylose utilisation, or to combat oxidative stress by charging of the glutathione system. We found a remarkably clear down-regulation of glycolytic genes on xylose with a moderate fold change. The gluconeogenesis-associated gene fructose-1,6-bisphosphatase *FBP1* was also sharply up-regulated by 27-fold, which was also observed in a microarray experiment of a xylose utilising recombinant *S. cerevisiae* strain [33]. β-Oxidation reactions and fatty acid activation reactions were clearly up-regulated.

Further, there was also a strong apparent down-regulatory effect on alcohol dehydrogenases and aldehyde dehydrogenases, and both the citrate synthase of the TCA cycle and the isocitrate lyase of the glyoxylate cycle were seemingly dramatically up-regulated. The latter two observations in this analysis, however, are misguided by the lack of compartmentalisation of the response. It should be noted that many metabolic reactions could be catalysed by more than one enzyme, where for some reactions like those catalysed by alcohol dehydrogenase (ADH), there are at least five gene products that could potentially catalyse the same reaction. Fig 12 used only the most extremely altered differentially expressed gene for rendering. Fig 13 provides a different perspective, however, where reactions that have more than one associated gene and which were regulated in opposite directions are identified. These are isocitrate lyase, aldehyde dehydrogenase and alcohol dehydrogenase.

Subsequently, the GO 'cellular_component' terms from the UniProt_SwissProt_fungi were used to render compartmentalised maps reconstructed from MetaCyc pathways. Fig 14 shows that in the cytosol xylose utilisation reactions were up-regulated, whereas glycolysis, together with pyruvate decarboxylase, NAD^+-specific acetaldehyde dehydrogenase (ALD) and glycerol production were down-regulated. Further, a number of reactions usually associated with the TCA cycle were also present in the cytosol and were constitutively expressed. More than one type of ADH with opposite regulatory direction is present in the cytosol. In the peroxisomal compartment (see S1 Pathway), β-oxidation of lipids, which in *S. cerevisiae* is performed exclusively in the peroxisomes [34], is clearly visible with only the 3-hydroxyacyl-CoA dehydrogenase (OHACYL-DEHYDROG-RXN) gene missing from the annotation. In mitochondria (see S1 Pathway) the reactions catalysed by the ALDs (*ALD4*, *ALD5*, *ALD6*) and ADHs were still isozyme switching reactions. Both *ALD4* (up-regulated 83-fold) and *ALD5* (down-regulated 167-fold) were strongly differentially expressed. Several ADH genes were found in *K. marxianus*. Two of these were annotated as *ADH3* and *ADH4*, both mitochondrial, whereas five (*ADH1*, *ADH2, ADH6*, *SFA1* and *adh*) were taken to be cytoplasmic. As previously reported [35], ADH2 was only expressed significantly in the presence of

glucose, and was in fact the most significantly down-regulated gene in our dataset (229-fold down-regulated). This corresponds to the regulation of the ADH2 orthologs in *S. cerevisiae* and *K. lactis* [35]. Conversely, ADH1 was constitutively expressed, again as previously reported [35], which differs from the regulation in *S. cerevisiae* and *K. lactis* where glucose stimulated ADH1 expression. Transcriptional rewiring of ADH isozymes is thus present in *K. marxianus* compared to its relatives.

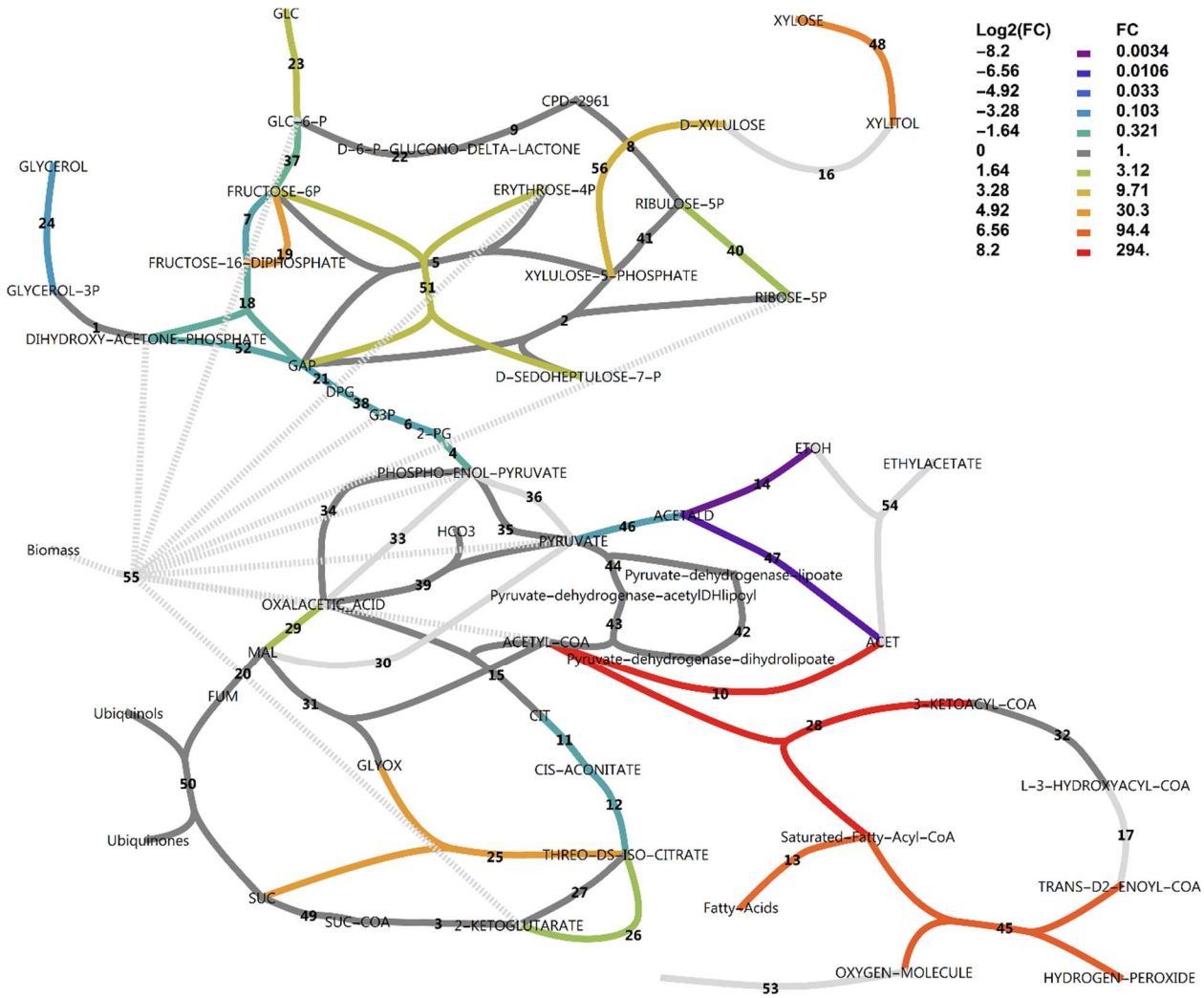


Figure 12. Uncompartmentalised response to xylose in central carbon metabolism in a $\log_2$(fold change) scheme. A $\log_2$(fold change) is defined as the $\log_2$ ratio of transcripts on xylose divided by that on glucose, as reported by CuffDiff. Reactions vETC, vEthylAcetate and vGrowth were manually added to the model. In the case of more than one enzyme that could perform the same function, the largest fold change in expression was used for the colour rendering. For reaction names, see S1 Pathway.

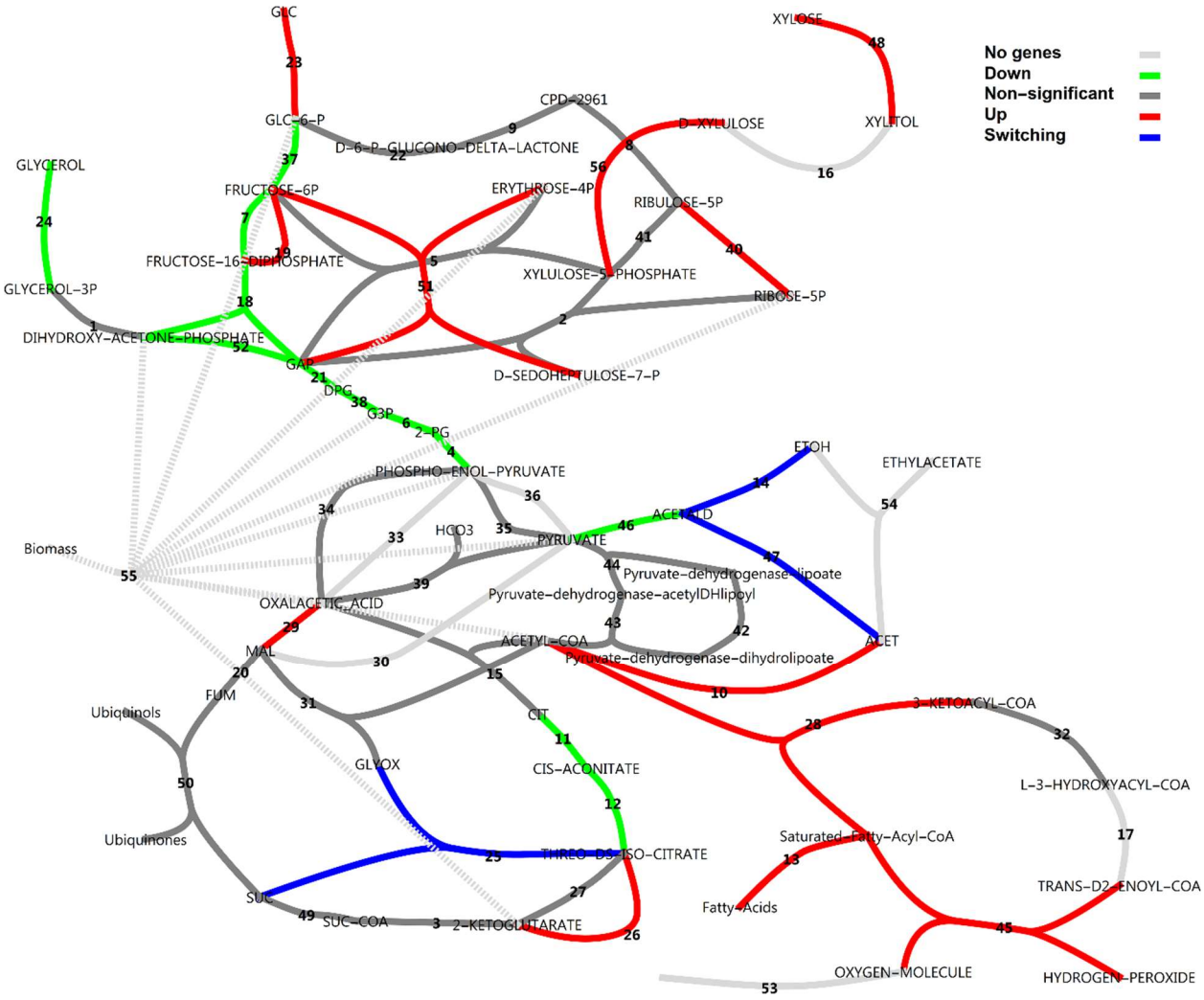


Figure 13. Uncompartmentalised response to xylose in central carbon metabolism as a classification scheme. Blue reactions represent those for which more than one enzyme gene has been assigned and for which some were up-regulated and some down-regulated, referred to as isozyme switching. For reaction names, see S1 Pathway.

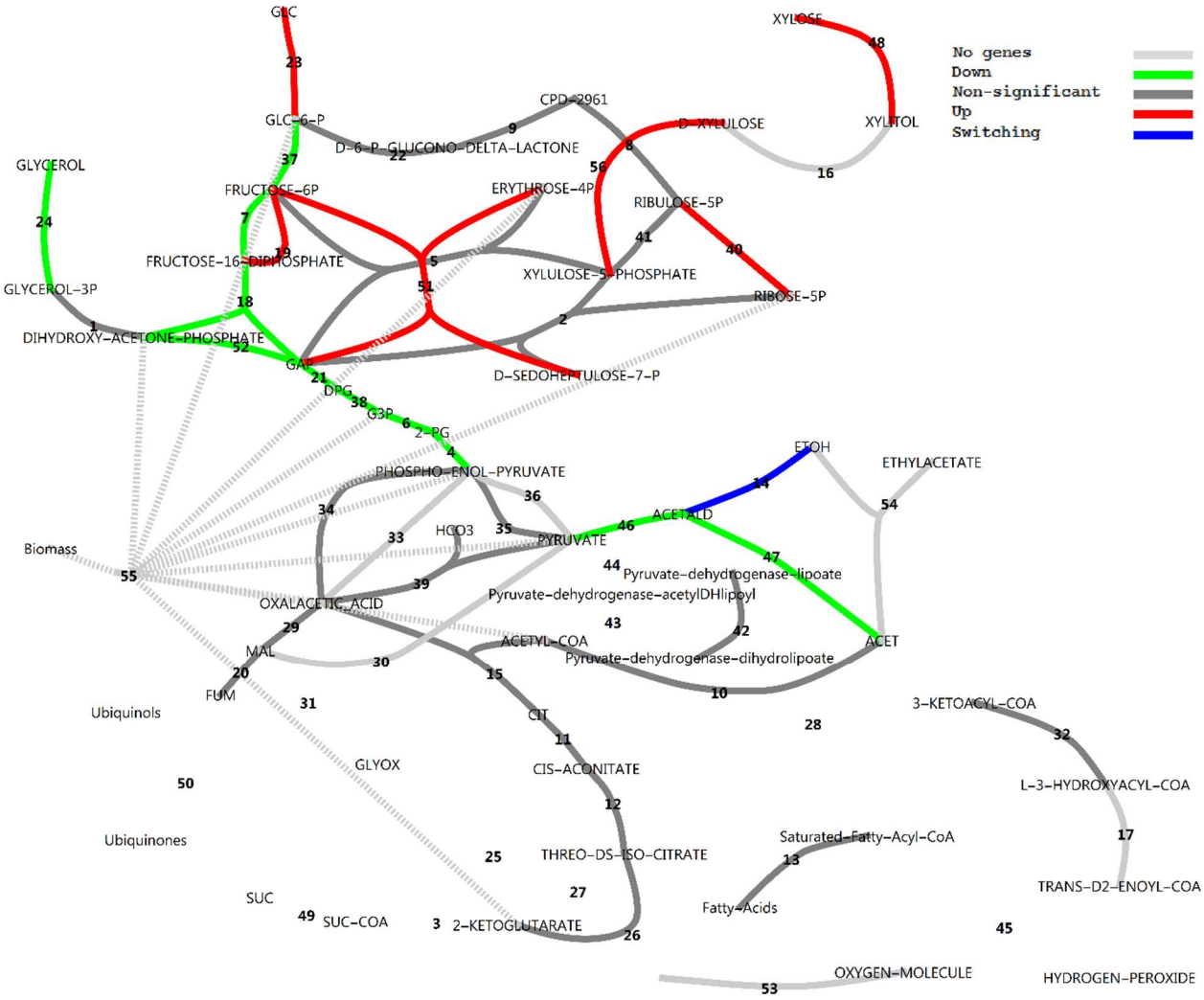


Figure 14. Compartmentalised response to xylose in central metabolism in the cytoplasm using the classification scheme. Blue reactions represent those for which more than one enzyme gene has been assigned and for which some were up-regulated and some down-regulated, referred to as isozyme switching. For reaction names, see S1 Pathway.

In the mitochondrial map, the NADPH specific isocitrate dehydrogenase gene was up-regulated, whereas citrate synthase of the TCA cycle and isocitrate lyase were seemingly up-regulated. Surprisingly, the glyoxysome map revealed that the glyoxylate cycle specific isocitrate lyase (*ICL1*) was down-regulated. Upon further investigation, the isozyme present in the mitochondrion, which was mapped to the TCA cycle by the PathoLogic algorithm using Kegg-Kaas annotations, was in fact *ICL2*, the 2-methylisocitrate lyase, perhaps confusingly termed *ICL2*. Although *ICL1* and *ICL2* share a high sequence similarity, the latter does not use isocitrate as a substrate and does not produce glyoxylate but uses 2-methylisocitrate instead

to produce succinate and pyruvate [36]. The gene product from *ICL2* is part of the propionyl-coezyme A pathway, otherwise known as the 2-methylcitrate pathway [37]. The pathway starts with *CIT3* (*YNR001C*), the mitochondrial enzyme that condenses oxaloacetate with propionyl coenzyme A to form 2-methylcitrate. Indeed, the citrate synthase that was up-regulated (Fig 12) was not the *CIT1* gene (*YPR001W*) that is both cytosolic and mitochondrial and associated with the TCA cycle, but instead the mitochondrial *CIT3*. Further inspection of the full GO ontology enrichment set revealed that all three key genes in the 2-methylcitrate pathway were significantly up-regulated. The strength of the response in the 2-methylcitrate pathway is given in Table 1, where *CIT3*, *PDH1* and *ICL2* were up-regulated 294, 32 and 28-fold, respectively, and Fig 15 shows the connection of this pathway with the TCA cycle. The GO term 'propionate catabolic process' (GO:0019543) was found to be significantly enriched with an enrichment score of 5.2. Apart from the last cycle of β-oxidation of odd-chain fatty acids in the peroxisomes as the source of propionyl-CoA, the latter could be derived from propionate or even from threonine breakdown [36]. In the gene set for 'threonine metabolic process' (GO:0006566), only one gene was up-regulated, namely the low specificity L-threonine aldolase (*GLY1*). Another gene was also annotated as *GLY1* and down-regulated. Thus, there is no conclusive evidence that threonine catabolism was up-regulated. The up-regulation of the 2-methylcitrate pathway is, therefore, more likely for the catabolism of short-chain fatty acids and not threonine. Painted differential expression pathway maps were generated for all of 235 metabolic pathways found in the annotation. They can be explored in see S2 Pathway.

Table 1. Differential expression of the constituent genes mapped to the GO term 'propionate catabolic process' (GO:0019543) on glucose and xylose, respectively.

Id	Value (Glc)	Value (Xyl)	log2(FC)	qvalue	signt	Entry	Protein names	Gene names
GK5S-1542	5.2	1512.1	8.2	0.001	yes	P43635	Citrate synthase 3 (EC 2.3.3.16)	CIT3 YPR001W YP9723.01
GK5S-1543	27.5	901.5	5.0	0.001	yes	Q12428	Probable 2-methylcitrate dehydratase (EC 4.2.1.79)	PDH1 YPR002W LPZ2W YP9723.02
GK5S-2306	15.5	436.9	4.8	0.001	yes	Q12031	Mitochondrial 2-methylisocitrate lyase (EC 4.1.3.30)	ICL2 YPR006C LPZ6C YP9723.06C

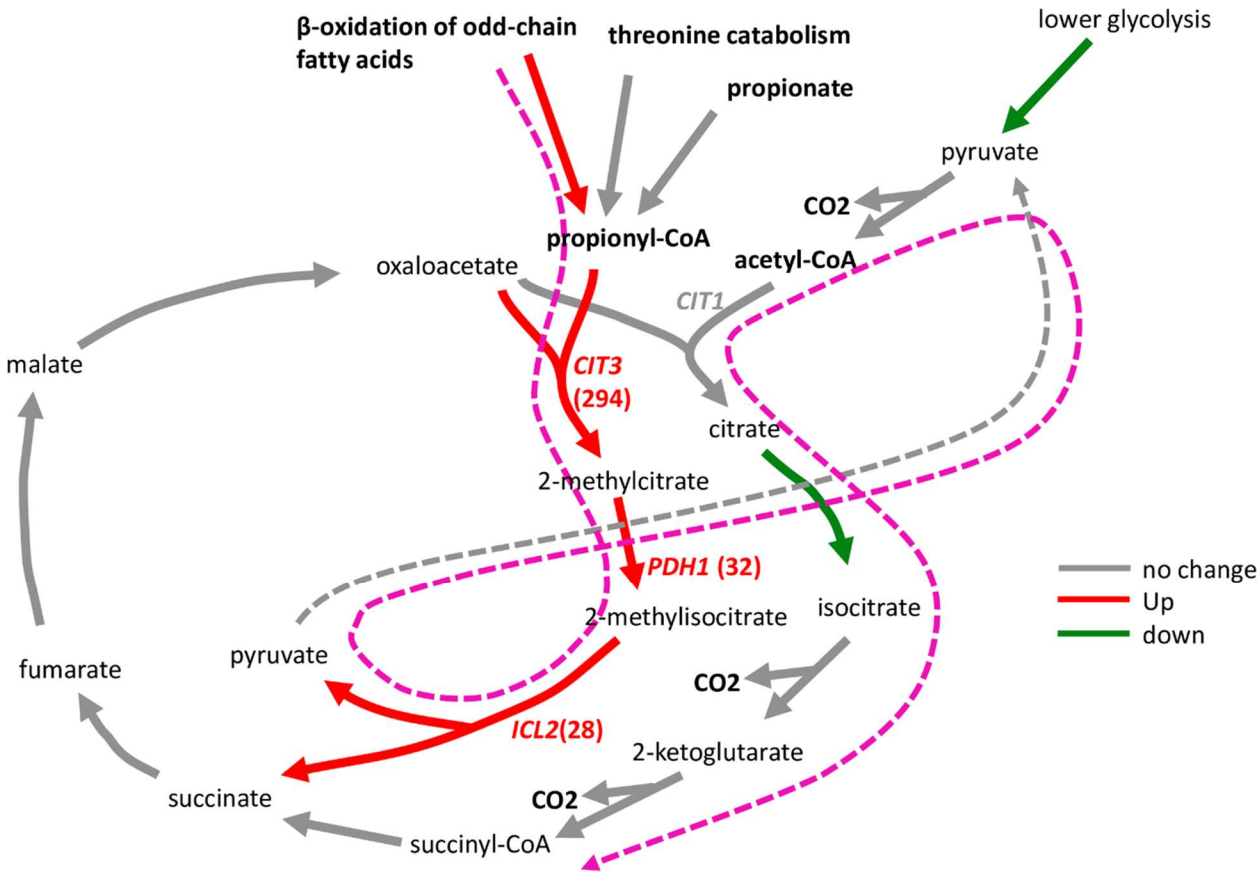


Figure 15. 2-Methylcitrate pathway. The suggested route of three-carbon units is indicated in magenta.

Metabolic Regulation Analysis to dissect hierarchical and metabolic levels of regulation

Differential metabolic flux analysis can be combined with differential gene expression data in the framework of MRA. From Figs 10 and 11 it is evident that in anaplerosis, phosphoenolpyruvate carboxykinase had very low transcript levels, while phosphoenolpyruvate carboxylase was absent from the annotation. Only the pyruvate carboxylase showed substantial transcript levels and was constitutively expressed. Hence, only pyruvate carboxylase was used in the model for flux estimations. In glucose medium, respiro-fermentative metabolism was observed. The specific uptake rate of glucose (on a dry cell weight basis) was 9.4 mmol h^{-1} mg^{-1} with acetate and ethanol produced at 2.0 and 2.7 mmol h^{-1} mg^{-1}, respectively with no glycerol formation. In xylose medium, acetate, ethanol and glycerol were absent and the specific xylose uptake rate was estimated at approximately 5 mmol h^{-1} mg^{-1}. The ethanol flux calculated could be considered as a conservative estimate, since some evaporation of ethanol could be expected from the aerobic condition. Fig 16 shows the differential flux outputs, as approximated by

combining FBA with the consumption rates of sugars and production rates of ethanol and acetate (see S5 Table and S3 Pathway for the FBA model, parameters and MRA outputs). The xylose utilisation pathways as well as the transaldolases and transketolase fluxes of the non-oxidative PPP were up-regulated, while the direction of the glucose-6-phosphate isomerase was changed towards the direction of glucose-6-phosphate. Notably, the difference in oxidative pentose phosphate pathway flux was only 5% between glucose and xylose utilisation modes and was thus considered constitutive, as was found in the transcript levels. The rest of central metabolic fluxes were down-regulated. (See S3 Pathway for a more detailed $\log_2$(fold change) scheme). Although transcript levels do not generally correspond well to protein levels and hence to fluxes according to reports using model organisms [38], the overall patterns of expression in central carbon metabolism in our RNA-seq analysis corresponded well with the predicted flux patterns.

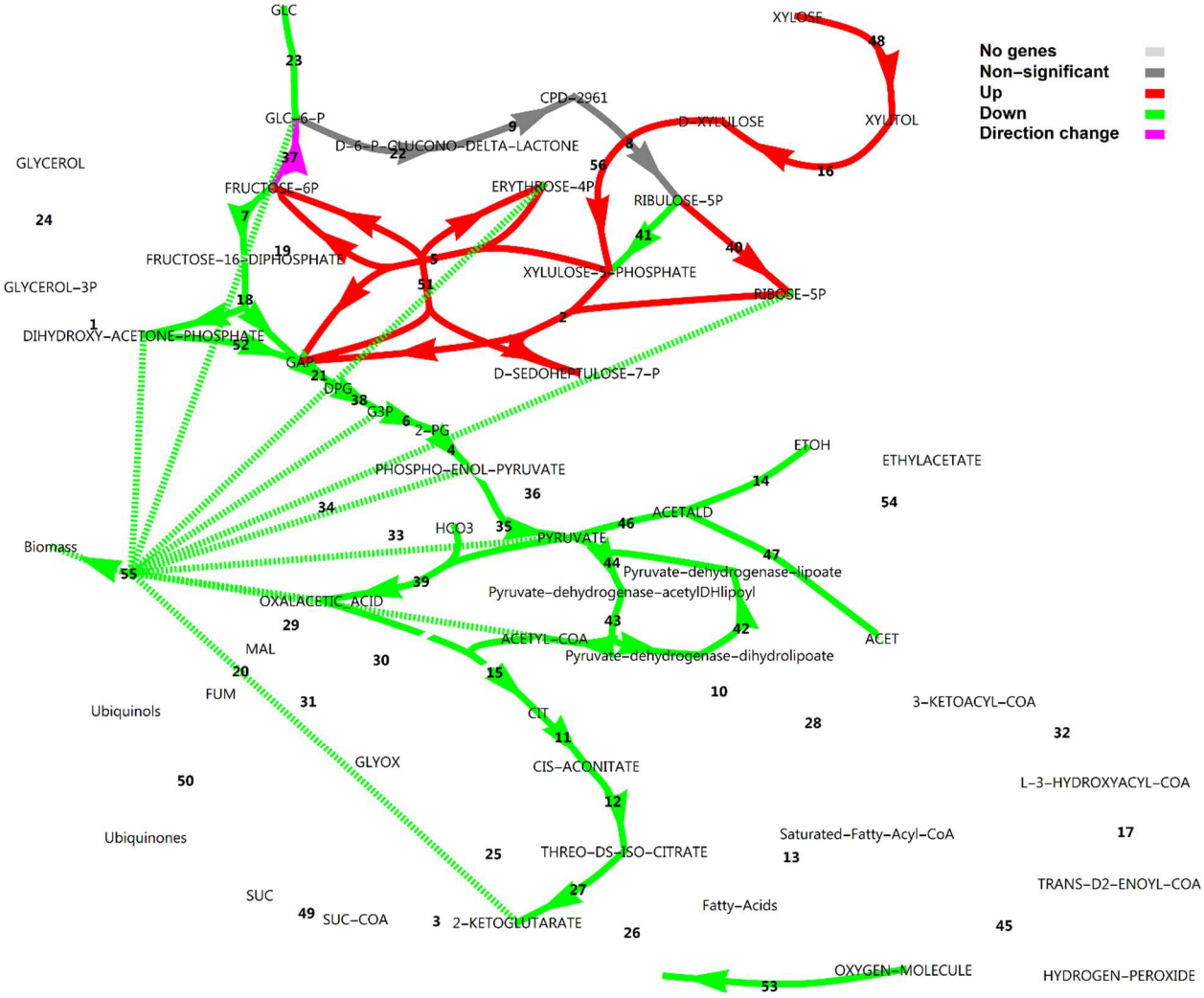


Figure 16. Differential flux analysis of central carbon metabolism. Fluxes which differed by less than 10% between conditions were considered constitutive. For reaction names, see S3 Pathway.

Although there is not always a linear correlation between transcript level, protein concentrations and maximal activities of an enzyme, we assume differential transcript expression to provide a sufficient approximation to hierarchical regulation, as was done by others [39]. Fig 17 shows the separation of metabolic regulation into hierarchical and metabolic level regulation. It is evident that the genetic level up-regulation of xylose reductase, sorbitol dehydrogenase, xylulose kinase, transaldolase and ribose-5-phosphate isomerase rendered the regulation as purely hierarchical. Of these, sorbitol dehydrogenase acting as a xylitol dehydrogenase would be the most extreme example, as it showed a 208-fold genetic up-regulation. These four reactions interact with neighbouring reactions via the metabolites xylulose-5-phosphate, erythrose-4-phosphate, fructose-6-phosphate and ribose-5-phosphate to stimulate fluxes through the transketolases, and to reverse the flux through glucose-6-phosphate isomerase, whereas ribose-5-phsophate isomerase may lower the flux through ribulose-phosphate 3-epimerase by lowering the concentration of ribulose-5-phosphate. Notably, regulation of fluxes in lower glycolysis is dominated by the genetic component, whereas upper glycolysis is regulated approximately equally by the metabolic and genetic regulation levels. Downward from pyruvate kinase, regulation is dominated by changes in metabolite levels, while some contribution of the genetic regulation level is evident for the fluxes through the aconitate hydratases in the TCA cycle. Due to the absence of measured ethanol and acetate in the xylose medium, the disappearance of fluxes through pyruvate decarboxylase, alcohol dehydrogenase and aldehyde dehydrogenase rendered classification of regulation in these reactions as purely metabolic.

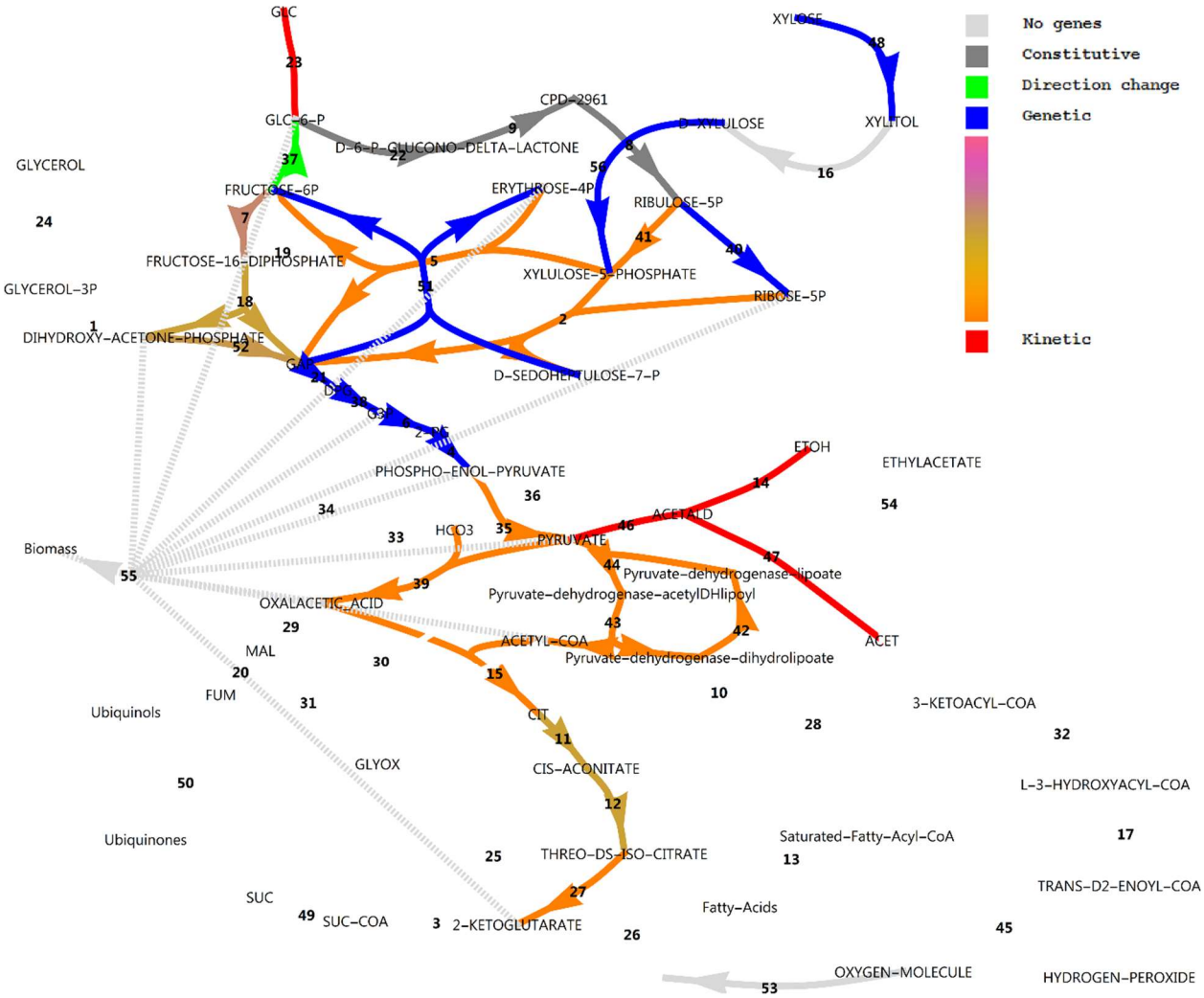


Figure 17. Metabolic Regulation Analysis of central carbon metabolism. For reaction names, see S3 Pathway.

Discussion

Modelling of bioprocesses has been attempted for a long time, in many different forms. It was realised in the second half of the twentieth century that several confounding factors had to be eliminated in the study of metabolism and its gene regulation, arguably with the growth rate as the most important confounding factor, since the expression level of many genes are correlated with the growth rate. Chemostat cultivation is a useful tool, since the growth rate can be easily controlled. However, for the vast majority of industrial applications, chemostats are unrealistic; industrial applications typically involve batch cultivation, or fed-batch cultivation to avoid catabolite repression, improve volumetric productivity and allow sufficient aeration by controlling the growth rate. Moreover, systems biology has become

increasingly data-driven with the availability of high-throughput methods to measure large numbers of intracellular metabolites (metabolomics), proteins (proteomics) and especially RNAs (RNA-seq and microarray) from a single experiment. It is proposed here that much more could be learned in terms of genetic regulation in microorganisms by testing a variety of different cultivation conditions, especially substrates, in small batch experiments as compared to the same amount of work in a more sophisticated, labour intensive chemostat setup, or at least, complementary information. We showed that in a simple and cost-effective batch setup, a large amount of rich RNA-seq data could be generated and fruitfully explored. This small working volume also makes expensive isotope labelling studies feasible. In batch experiments, however, special care needs to be taken in terms of the timing of sampling, as, for instance, fermentation products would accumulate over time and a transcriptome sample late in the fermentation may be more reflective of chemical stress than of the metabolic mode. This may especially be a confounding factor when comparing the effects of the concentration of substrate. Our experimental design alleviated this effect and allowed us to focus on the effects of alternative substrates only.

Previously, exploration of omics datasets mostly focused on only the lists of the most significantly up-regulated or down-regulated genes separately, or on one or two more advanced types of analyses such as GO enrichment, and occasionally on the extraction of active networks. Here we presented one of the first examples that explore high-quality RNA-seq data from various perspectives, using different types of enrichments and networks, and rationalisation of the response with FBA and MRA as theoretical frameworks. Furthermore, combining information from databases such as GO and MetaCyc in the same problem and with subcellular compartmentalisation revealed several important features that would have been missed using either on its own.

The vastness of the xylose response under aerobic conditions indicates a different, opportunistic lifestyle that this yeast apparently adapts to when cultivated on xylose, which may be reminiscent of its natural environment where a variety of plant-derived substrates may be utilised. Down-regulation of many biosynthetic pathways is concordant with a lower growth rate, although there is a notable absence of growth rate specific gene sets in the enrichment statistics. Although many mitochondrial genes were differentially expressed, these are not the major enzymes in energy production in mitochondria. Oxidative stress also seems minimal in the aerobic xylose medium. Instead, a strong response is seen in the up-regulation of alternative sugar utilisation machinery, including inulinase, sugar transporters and catabolic

routes for alternative sugars. Inulin is a fructan stored in large amounts in some plants, including *A. americana* [40]. Our strain was indeed isolated from an *A. americana* sample.

K. marxianus lacks enzymes such as secreted proteases, lipases and carbohydrate hydrolases. This yeast may thus be dependent on other fungi in the environment for these functions, or its natural habitat may be some commensal niche where these monomers are supplied by the plant. The transcriptional regulatory basis for this response is likely to be glucose repression by transcription factors such as *MIG1* binding to carbon source response elements, as was suggested to be the case for the inulinase gene *KmINU1* [4, 41].

It is interesting to find that the majority of genes for a complete organelle like the peroxisome, which is dedicated to lipid oxidation, was dramatically up-regulated in the xylose medium, yet without apparent function in the experimental setting. Thus, glucose de-repression is likely sufficient to activate most of the response to enable lipid catabolism, and stimulation by lipids may play a smaller role. The mitochondrial 2-methylcitrate pathway for the degradation of three-carbon molecules was also strongly up-regulated, suggesting that a variety of odd-chain fatty acids originating from peroxisomes may be oxidised.

FBA simulations predicted no significant up-regulation of the oxidative PPP flux, but a more intense up-regulation of the non-oxidative PPP flux and a down-regulation of glycolysis, which were consistent with RNA-seq data. Since biomass formation, the major sink for NADPH, is down-regulated in xylose medium, and another NADPH sink, xylose reductase, is up-regulated, it thus makes sense that the major source flux of NADPH (oxidative PPP) may be similar in both conditions (considering absolute fluxes normalised by the biomass concentration). Normalising fluxes to the uptake rate of a carbon source or to the biomass formation rate could thus be misleading when interpreting expression data. FBA simulations here thus shed light on what could be expected in terms of gene regulation.

MRA was used to separate metabolic regulation into hierarchical and metabolic levels. Reactions in lower glycolysis are known to a have high flux capacity due to high enzyme concentrations, and our transcript data also indicated this feature on glucose. Chemostat studies of *S. cerevisiae* classified these high-capacity reactions as well as the non-oxidative PPP reactions as pseudo-equilibrium or near-equilibrium reactions, suggesting that they can be sufficiently described by simple equations making use of thermodynamics and empirical studies at various dilution rates [42]; hence both detailed enzyme kinetic

expressions and genetic regulation could be ignored. Reversible high capacity reactions like these should have low metabolic control coefficients over the flux and are not likely to be regulated at the genetic level - at least over small changes in the flux, close to a reference steady state. However, our transcript level data for *K. marxianus* showed that the transcript levels in lower glycolysis were substantially lower on xylose, and MRA showed a dominating hierarchical (genetic) level regulation. Flux through transaldolase was also dominated by hierarchical regulation. The contribution of hierarchical and metabolic level regulation of a reaction would differ between substrates utilised, and may also differ between species.

Reporter metabolite networking suggested that NAD(H), acetaldehyde, ethanol, glyceraldehyde-3-phosphate, 3-phosphoglycerate and hydrogen peroxide formed a strongly interconnected redox active system, with aldehyde dehydrogenases and alcohol dehydrogenases being the main players. This system may work across membranes, since acetaldehyde, ethanol and hydrogen peroxide can cross membranes, shuttles connect NAD(H) across compartments, and glyceraldehyde-3-phosphate and 3-phosphoglycerate are closely connected to NADH. At this stage it is not possible to resolve the fluxes and MRA through the acetaldehyde-dependent pyruvate dehydrogenase bypass. It was, however, evident that dramatic changes occurred in the transcript levels of the alcohol dehydrogenases and aldehyde dehydrogenases, including compartment-specific isozyme switching. As also suggested by Lertwattanasakul et al. [3], ethanol may be catabolised in the mitochondria as fast as it is produced. However, it has to be emphasised that care needs to be taken in extrapolating gene expression data between studies carried out under aerobic and anaerobic conditions. Oxygen limitation results in differential expression of many genes in yeasts.

Conclusions

We believe to have captured in a unique manner, and from a number of perspectives, a complex transcriptional pattern telling an interesting story about how the cell 'explores its options' when the nutrient availability changes under aerobic conditions. Strong up-regulation of transporters and pathways for utilisation of alternative carbohydrates was evident. In addition, the more opportunistic lifestyle was supported by invasive growth, and sexual reproduction was activated as a long-term survival strategy. The strong peroxisomal fatty acid catabolic response accompanied by the mitochondrial 2-methylcitrate pathway is likely explained by glucose de-repression, similar to that seen for carbohydrate utilisation. As

K. marxianus seemingly lacks the secreted enzymes required for depolymerisation of biopolymers, the species is probably dependent on other species for supply of monomers such as sugars, amino acids and free fatty acids, whereas inulinase is a specialist feature enabling this species to utilise this plant storage oligosaccharide. It would be interesting to see whether xylose may have a stimulatory role as suggested recently for *Saccharomyces cerevisiae* [43] and through which signaling pathways this may take place.

MRA was demonstrated here as an informative method to dissect the regulation of fluxes into the metabolic and hierarchical levels. It is evident that the genetic level plays a dominating role in the regulation of fluxes in central carbon metabolism, not only in the early enabling steps of utilisation of xylose as the carbon source, but also in the high capacity reactions of lower glycolysis. In kinetic modelling of metabolism, emphasis should thus be placed on genetic regulation, which is currently very challenging. Multiple omics would need to be combined to predict such regulatory networks and build realistic models. However, in order to resolve fluxes originating from pyruvate, isotopic tracer studies would be required, which currently is under investigation. In addition, the isozyme switching observed with alcohol dehydrogenase, aldehyde dehydrogenase and acetate-CoA ligase calls for a detailed investigation into the compartmentalisation and kinetics of these enzymes, and detailed bottom-up kinetic modelling to understand the role of this interesting behaviour.

Acknowledgements

The authors would like to thank Dr. Jonathan Featherston, Dr. Dirk Swanevelder, Prof. Jasper Rees and Ms Minique De Castro at the Onderstepoort Biotechnology Platform, Pretoria, South Africa for sequencing work and fruitful discussions, and to Ms Precious Letebele, the Honours student, who performed the RNA-seq experimental work.

References

1. Groeneveld P, Stouthamer AH, Westerhoff HV. Super life – how and why 'cell selection' leads to the fastest-growing eukaryote. FEBS J. 2009;276: 254-270.

2. Rocha SN, Abrahão-Neto J, Gombert AK. Physiological diversity within the *Kluyveromyces marxianus* species. Antonie van Leeuwenhoek. 2011;100: 619–630. doi 10.1007/s10482-011-9617-7

3. Lertwattanasakul N, Kosaka T, Hosoyama A, Suzuki Y, Rodrussamee N, Matsutani M, et al. Genetic basis of the highly efficient yeast *Kluyveromyces marxianus*: complete genome sequence and transcriptome analyses. Biotechnol. Biofuels. 2015;8(47). doi:10.1186/s13068-015-0227-x.

4. Gao J, Yuan W, Li Y, Xiang R, Hou, S, Zhong S, Bai F. Transcriptional analysis of *Kluyveromyces marxianus* for ethanol production from inulin using consolidated bioprocessing technology. Biotechnol. Biofuels. 2015;8:(115). doi:10.1186/s13068-015-0295-y.

5. van Dijk EL, Auger H, Yan Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30(9): 418-426. doi:10.1016/j.tig.2014.07.001.

6. Buermans HPJ, den Dunnen JT. Next generation sequencing technology: Advances and applications. Biochim. Biophys. Acta. 2014;1842 :1932–1941.

7. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan GJ, van Baren M, et al. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nat. Biotechnol. 2010;28(5): 511–515. doi:10.1038/nbt.1621.

8. Gao QG, Jin K, Ying S, Zhang Y, Xiao G, Shang Y, et al. Genome Sequencing and Comparative Transcriptomics of the Model Entomopathogenic Fungi *Metarhizium anisopliae* and *M. acridum*. PLOS Genet. 2011;1(7). doi: 10.1371/journal.pgen.1001264.

9. Maciejewski, H. Gene set analysis methods: statistical models and methodological differences. Briefings Bioinf. 2013;15: 504-518. doi:10.1093/bib/bbt002.

10. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, et al. PathwayTools version 13.0: integrated software for pathway/genome informatics and systems biology. Briefings Bioinf. 2009;11(1): 40-79. doi:10.1093/bib/bbp043.

11. Tang J. Microbial metabolomics. Curr. Genomics. 2011;12: 391-403. doi:10.2174/138920211797248619.

12. Fiehn O. Metabolomics – the link between genotypes and phenotypes. Plant Mol. Biol. 2002;48: 155–171.

13. Patil KR, Nielsen J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc. Natl. Acad. Sci. U. S. A. 2005;102(8): 2685-2689.

14. Oliveira AP, Patil KR, Nielsen J. Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. BMC Syst. Biol. 2008;2(17). doi:10.1186/1752-0509-2-17.

15. Schilling CH, Edwards JS, Palsson BO. Toward metabolic phenomics: analysis of genomic data using flux balances. Biotechnol. Prog. 1999;15: 288-295.

16. Covert MW, Schilling CH, Famili I, Edwards JS, Goryanin II, Selkov E, et al. Metabolic modeling of microbial strains *in silico*. Trends Biochem. Sci. 2001;26(3): 179-186. pii: S0968-0004(00)01754-0. doi:10.1016/S0968-0004(00)01754-0.

17. Dias O, Pereira R, Gombert AK Ferreira EC, Rocha I. iOD907, the first genome-scale metabolic model for the milk yeast *Kluyveromyces lactis*. Biotechnol. J. 2014;9(6): 776–790. doi: 10.1002/biot.201300242.

18. ter Kuile BH, Westerhoff HV. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. FEBS lett. 2001;500. pii: S0014-5793(01)02613-8.

19. du Preez JC, van der Walt JP. Fermentation of D-xylose to ethanol by a strain of *Candida shehatae*. Biotechnol. Lett. 1983;5(3): 357-362.

20. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics. 2014;30(15): 2114-2120. doi: 10.1093/bioinformatics/btu170.

21. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 2013;7(3): 562–578. doi:10.1038/nprot.2012.016.

22. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signaling circuits in molecular interaction networks. Bioinformatics. 2002;18(1): 233-240.

23. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2014;43: 1049-1056. doi: 10.1093/nar/gku1179.

24. Barupal DK, Haldiya PK, Wohlgemuth G, Kind T, Kothari SL, Pinkerton PE, et al. MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. BMC Bioinf. 2012;13(99). doi:10.1186/1471-2105-13-99.

25. Fischer E, Zamboni N, Sauer U. High-throughput metabolic flux analysis based on gas chromatography–mass spectrometry derived ^{13}C constraints. Anal. Biochem. 2004;325: 308-316. doi:10.1016/j.ab.2003.10.036.

26. Wolfram CDF Player. https://www.wolfram.com/cdf-player/.

27. Guaragnella N, Butow RA. 2003. ATO3 encoding a putative outward ammonium transporter is an RTG-independent retrograde responsive gene regulated by GCN4 and the Ssy1-Ptr3-Ssy5 amino acid sensor system. J. Biol. Chem. 278: 45882–87

28. Liu Z, Ronald A. Butow RA. Mitochondrial retrograde signaling. Annu. Rev. Genet. 2006;40: 159–185.

29. Kim WH, Chung JH, Back JH, Choi J, Koh J et al. Molecular cloning and characterization of an NADPH quinone oxidoreductase from *Kluyveromyces marxianus*. J. Biochem. Mol. Biol. 2003;36(5): 442-449.

30. Lucau-Danila A, Delaveau T, Lelandais G, Devaux F, Jacq C. Competitive promoter occupancy by two yeast paralogous transcription factors controlling the multidrug resistance phenomenon. J. Biol. Chem. 2003;278(52): 52641–52650.

31. Gurvitz A, Rottensteiner H, Kilpeläinen SH, Hartig A, Hiltunen JK, Binder M, et al. The *Saccharomyces cerevisiae* peroxisomal 2,4-dienoyl-CoA reductase is encoded by the oleate-inducible gene SPS19. J. Biol. Chem. 1997;272(35): 22140–22147.

32. Harcus D, Dignard D, Lépine G, Askew C, Raymond M, Whiteway, M et al. 2013. Comparative xylose metabolism among the ascomycetes *C. albicans*, *S. stipitis* and *S. cerevisiae*. PLoS One. 2013; 8(11): e80733.

33. Runquist D, Hahn-Hägerdal B, Bettiga M. Increased expression of the oxidative pentose phosphate pathway and gluconeogenesis in anaerobically growing xylose utilizing *Saccharomyces cerevisiae*. Microb. Cell Fact. 2009;8: 49. doi:10.1186/1475-2859-8-49.

34. Poirier Y, Antonenkov VD, Glumoff T, Hiltunen JK. Peroxisomal β-oxidation - A metabolic pathway with multiple functions. Biochim. Biophys. Acta. 2006;1763: 1413–1426. doi:10.1016/j.bbamcr.2006.08.034.

35. Lertwattanasakul N, Sootsuwan, Limtong S, Thanonkeo P, Yamada M. Comparison of the gene expression patterns of alcohol dehydrogenase isozymes in the thermotolerant yeast *Kluyveromyces marxianus* and their physiological Functions. Biosci Biotechnol Biochem. 2007;71(5): 1170-1182.

36. Luttik MAH, Kötter P, Salomons FA, van der Klei IJ, Dijken JP, Pronk JT. The *Saccharomyces cerevisiae* ICL2 Gene Encodes a Mitochondrial 2-Methylisocitrate Lyase Involved in Propionyl-Coenzyme A Metabolism. J. Bacteriol. 2000;182(24): 7007–7013.

37. Pronk JT, van der Linden-Beuman A, Verduyn C, Scheffers WA, van Dijken JP. Propionate metabolism in *Saccharomyces cerevisiae*: implications for the metabolon hypothesis. Microbiology. 1994;140: 717-722. doi: 10.1099/00221287-140-4-717.

38. de Sousa Abreu R, Penalva OP, Marcotte EM, Vogel C. Global signatures of protein and mRNA expression levels. Mol. Biosyst. 2009;5(12): 1512–1526. doi:10.1039/b908315d.

39. Orencio-Trejo M, Flores N, Escalante A, Hernandez-Chavez G, Bolivar F, Gosset G, et al. Metabolic regulation analysis of an ethanologenic *Escherichia coli* strain based on RT-PCR and enzymatic activities. Biotechnol. Biofuels. 2008;1(8). doi:10.1186/1754-6834-1-8.

40. Chi Z, Chi Z, Zhang T, Liu G, Yue L. Inulinase-expressing microorganisms and applications of inulinases. Appl. Microbiol. Biotechnol. 2009;82: 211–220. doi:10.1007/s00253-008-1827-1.

41. Lertwattanasakul N, Rodrussamee N, Suprayogi, LS, Thanonkeo P, Kosaka T, Yamada M. Utilization capability of sucrose, raffinose and inulin and its less-sensitiveness to glucose repression in thermotolerant yeast *Kluyveromyces marxianus* DMKU 3-1042. AMB Express. 2011;1(20): 1-11. doi: 10.1186/2191-0855-1-20.

42. Canelas AB, Ras C, Pierick A, van Gulik WM, Heijnen JJ. An *in vivo* data-driven framework for classification and quantification of enzyme kinetics and determination of apparent thermodynamic data. Metab. Eng. 2011;13: 294-306.

43. Alff-Tuomala S, Salusjärvi L, Barth D, Oja M, Penttilä M, et al. Xylose-induced dynamic effects on metabolism and gene expression in engineered *Saccharomyces cerevisiae* in anaerobic glucose-xylose cultures. Appl. Microbiol. Biotechnol. 2016;100(2): 969-985. doi:10.1007/s00253-015-7038-7.

Supporting Information

This paper was published in PLoS One on 17 June 2016. See supporting information online.

Schabort DWP, Letebele PK, Steyn L, Kilian SG, du Preez JC. Differential RNA-seq, multi-network analysis and metabolic regulation analysis of *Kluyveromyces marxianus* reveals a compartmentalised response to xylose. PLoS ONE. 2016;11(6): e0156242. doi:10.1371/journal.pone.0156242.

S1 Text. Supplementary text on gene set enrichment and extracellular enzymes.

S1 Table. Main data tables.

S2 Table. Enrichment statistics of GO cellular_component.

S3 Table. Enrichment statistics of GO molecular_function.

S4 Table. Enrichment statistics of GO biochemical_process.

S5 Table. Metabolic flux model and Metabolic Regulation Analysis

S1 Network. Pathway-to-pathway Network.

S2 Network. Molecule-to-molecule network.

S3 Network. Enzyme-Reporter metabolite network.

S4 Network. Enzyme-reporter metabolite network (NAD).

S5 Network. Enzyme-reporter metabolite network (NADP).

S1 Pathway. RNA-seq data mapping to central metabolism.

S2 Pathway. Pathway multi-map.

S3 Pathway. Estimated metabolic flux maps.

S1 Draftgenome. Draft genome for *K. marxianus* UFS-Y2791 reconstructed *de novo*.

S1 ORF. Putative open reading frames found in the genome.

S1 Proteins. Amino acid sequences.

Supplementary Text

This supplementary materials accompanied the paper in PLoS One on 17 June 2016. See other supporting information online.

Schabort DWP, Letebele PK, Steyn L, Kilian SG, du Preez JC. Differential RNA-seq, multi-network analysis and metabolic regulation analysis of *Kluyveromyces marxianus* reveals a compartmentalised response to xylose. PLoS ONE. 2016;11(6): e0156242. doi:10.1371/journal.pone.0156242.
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0156242

Sugar utilization

Some interesting aspects arose from the data. Firstly, a variety of carbohydrate utilization genes were differentially expressed on glucose vs xylose, mostly for utilization of the hexose galactose, and the pentoses xylose and arabinose. Even though these alternative sugars were not present in the medium, their utilization pathways were up-regulated, a phenomenon typical of alleviation of glucose repression. Secondly, the capacity for hexose utilization by hexokinase was constitutive and high. Thirdly, the initial steps in galactose, xylose and arabinose utilization were more strongly up-regulated as compared to later steps in each pathway.

Extracellular enzymes

Since *K. marxianus* seems to adopt an opportunistic metabolic gene set profile in response to the presence of xylose (or absence of glucose), the question arises whether extracellular hydrolases possessing β-glucosidase activity, such as cellulases, would be activated, allowing *K. marxianus* to thrive on decaying plant matter. Cellulases are prevalent in fungi but not as prevalent in yeasts. The strain possesses fourteen genes annotated as having some form of β-glucosidase activity, most of which are secreted (supplementary Table 1 below). Beta-glucosidase (EC 3.2.1.21, cellobiase, gentiobiase) is a likely candidate that could function to hydrolyze cellobiose, the disaccharide found after hydrolysis by cellulose. It is 48.8-fold up-regulated and transcribed at a high level on xylose. It would be interesting to see whether this strain can grow on cellobiose, which is a trait uncommon to yeasts. Cellulose degradation is a complex process and requires the initial degradation of cellulose to cellobiose by cellulases. No cellulases were found in the genome, however. No genes were found to have xylanase activity or related terms either. Additional searches were performed for proteases, peptidases and lipases, which would be typical of

opportunistic yeasts, especially pathogens. A number of proteases, peptidases and lipases were found, but none of them are clearly secreted proteins.

Table 1. Enzymes with glucosidase and related activities.

ID	val(Glc)	val(Xyl)	log2(FC)	q	signt	Entry	Protein names	Gene names	EC number
g3519.t1	33.1	1601.0	5.6	0.001	yes	P07337	Beta-glucosidase (Cellobiase) (Gentiobiase)		3.2.1.21
g3265.t1	135.8	347.4	1.4	0.016	yes	Q04951	Probable family 17 glucosidase SCW10	SCW10 YMR305C YM9952.07C	3.2.1.-
g3867.t1	262.7	582.8	1.1	0.047	yes	Q12628	Glucan 1,3-beta-glucosidase (EC 3.2.1.58)	KLLA0C05324g	3.2.1.58
g4768.t1	841.6	1765.9	1.1	0.062	no	P53334	Probable family 17 glucosidase SCW4	SCW4 YGR279C	3.2.1.-
g3788.t1	922.9	1826.6	1.0	0.094	no	P15703	Glucan 1,3-beta-glucosidase	BGL2 SCW9 YGR282C	3.2.1.58
g4569.t1	146.2	285.3	1.0	0.090	no	Q06625	Glycogen debranching enzyme	GDB1 YPR184W	2.4.1.25; 3.2.1.33
g3018.t1	754.8	1235.7	0.7	0.262	no	P53616	Probable secreted beta-glucosidase SUN4	SUN4 SCW3 YNL066W N2411 YNL2411W	3.2.1.-
g2805.t1	25.3	39.4	0.6	0.430	no	P08019	Glucoamylase, intracellular sporulation-specific	SGA1 SGA YIL099W	3.2.1.3
g1329.t1	48.5	69.2	0.5	0.491	no	Q12168	Endo-1,3(4)-beta-glucanase 2 (Laminarinase-2)	ACF2 ENG2 PCA1 YLR144C L3180	3.2.1.6
g4202.t1	192.4	242.9	0.3	0.662	no	P32486	Beta-glucan synthesis-associated protein KRE6	KRE6 CWH48 YPR159W	
g2867.t1	136.4	148.0	0.1	0.904	no	P53189	Probable family 17 glucosidase SCW11	SCW11 YGL028C	3.2.1.-
g1495.t1	77.8	67.8	-0.2	0.829	no	P38138	Glucosidase 2 subunit alpha	ROT2 GLS2 YBR229C YBR1526	3.2.1.84
g2866.t1	51.4	42.4	-0.3	0.759	no	P53008	Mannosyl-oligosaccharide glucosidase (Glucosidase I)	CWH41 GLS1 YGL027C	3.2.1.106
g3564.t1	241.1	164.6	-0.6	0.441	no	P53753	Endo-1,3(4)-beta-glucanase 1 (Laminarinase-1)	DSE4 ENG1 YNR067C N3547	3.2.1.6

Sexual reproduction and invasive growth

Yeasts that reproduce sexually through mating have a pheromone sensing system whereby the haploid α-cells secrete a pheromone, mating factor α-1, which is sensed by the a-cells, a signal that suppresses DNA synthesis in the a-cells, thus synchronizing them with α-cells for conjugation**.** Mating factor α-1 (*MF(ALPHA)1*, *YPL187W*) was up-regulated 53-fold on xylose, whereas barrierpepsin (*BAR1*, *YIL015W*) was up-regulated 17-fold. Barrierpepsin likely cleaves mating factor α and fine-tunes the concentration of the pheromone for optimal conjugation [1]. Similarly, protein *SST2* (*YLR452C*), which also responds to α-pheromone to desensitize cells to α-pheromone, was 3.3-fold down-regulated [2]. A functional interaction between cells during conjugation is made by agglutinins in response to pheromones, of which α-agglutinin was up-regulated 5.7-fold. The flocculation proteins *FLO1* (*YAR050W*), *FLO11* (*YIR019C*), and *FLO9* (*YAL063C*) were also up-regulated. They form functional interactions by causing cell aggregation in liquid culture and are known to be involved in the formation of diploid pseudohyphae, which play a role in the adhesion of cells to substrates, invasive growth by haploid cells (see [3] for a review). It was shown that FLO11 is dominant in invasive growth whereas FLO1 is dominant in flocculation, while they are both under the control of the transcription factor encoded by the MSS1 gene [3]. In summary, both sexual reproduction was up-regulated as well as a variety of genes involved with morphological changes that may be beneficial for penetration into a solid substrate. We observed only few pseudohyphae on xylose towards the stationary phase and none on glucose, indicating that the response is not manifested completely as a phenotype under these conditions, but likely further stimuli are needed to make a transition to invasive growth.

References

1. MacKay V. Welch SK, Insley MY, Manney TR, Holly J, Saari GC, et al. The *Saccharomyces cerevisiae* BAR1 gene encodes an exported protein with homology to pepsin. Proc. Natl. Acad. Sci. USA. 1988;85: 55-59.
2. Apanovitch DM, Slep KC, Sigler PB, Dohlman HG. Sst2 Is a GTPase-activating protein for Gpa1: Purification and characterization of a cognate RGS−Gα protein pair in yeast. Biochemistry. 1998;37(14): 4815-4822.
3. Bester MC, Pretorius IS, Bauer FF. The regulation of *Saccharomyces cerevisiae* FLO gene expression and Ca^{2+}-dependent flocculation by Flo8p and Mss11p. Curr. Genet. 2006;49: 375-383.

Chapter 4

Identification of major transcriptional regulators in central carbon metabolism – the enumerative approach

Abstract

In the previous chapters, the construction of a draft genome and an RNA-seq transcriptomics analysis of the differential response of *Kluyveromyces marxianus* strain UFS-Y2791 to glucose or xylose were described. The analysis revealed several interesting aspects, including major up-regulation of central metabolic routes and up-regulation of enzymes involved in alternative carbon source utilisation. Particularly, the peroxisomes and peroxisomal β-oxidation were drastically up-regulated in a xylose medium. The pattern is reminiscent of glucose derepression in the xylose medium. The genetic basis for the differential expression was investigated by testing the hypothesis that the same key transcription factors as identified in *S. cerevisiae* is the basis for the response. The enumerative method of over-represented heptamers was used to reveal Adr1 as the most probable transcription factor activating genes in *Kluyveromyces marxianus* in the xylose medium.

Introduction

When *Saccharomyces cerevisiae* is cultivated in a medium containing glucose above a low threshold, the cell represses a large number of genes. One quarter of the 6 000 genes are differentially regulated by the presence of glucose [Young et al. 2003]. Glucose transcriptional repression in *S. cerevisiae* involves six signalling pathways and also involves cross-talk between them [Broach 2012, Chapter 1]. These signalling pathways that control growth and development have recently been reviewed by Broach [2012]. The major pathway responsible for the majority of the changes is the Ras/protein kinase A pathway. The most drastic changes to expression levels in central carbon metabolism occur due to the action of the Snf1 kinase pathway, however. Snf1 is a master regulator kinase in yeast and has a homolog in mammals known as the AMP-activated protein kinase (AMPK). AMPK is understood as a guardian of energy homeostasis. It activates glucose uptake and oxidation, and fatty acid oxidation, while suppressing anabolic reactions [Broach 2012]. In *S. cerevisiae* too, it is involved with energy metabolism, and additionally regulates the utilisation of alternative carbon sources. In mammals, AMPK responds to the energy state by the breakdown products of glucose and not by glucose itself. AMP binds to AMPK and regulates its activity. In *S. cerevisiae*, AMP does not seem to be the signal, but more likely glucose signals through an unknown mechanism via upstream kinases such as Elm1, Tos3 and Sak1, or via the Reg1/Glc7 complex to the Snf1/Snf4 complex, inhibiting its activity. Snf1 activates the transcription factors (TFs) Cat8 via deactivation of Mig1 [Schöler and Schüller 1994, Soontorngun et al. 2007], and activates Sip4 [Schüller 2003] and Rds2 [Soontorngun et al. 2007]. Also, Snf1 indirectly activates Adr1 by either activating some dephosphorylase that removes the phosphate from the regulatory domain of Adr1, or by deactivating some kinase that phosphorylates Adr1 [Ratnakumar et al. 2010]. These TFs function primarily as activating TFs, whereas Rds2, a major regulator of gluconeogenesis, functions both as an activator and a repressor [Soontorngun et al. 2007]. The importance of Cat8 and Sip4 in activating gluconeogenesis has previously been emphasised [Barnett and Entian 2005, Carlson 1999, Schüller 2003].

Mig1 is a known repressor of gene expression of a variety of genes in *S. cerevisiae*, including the CAT8 gene [Hedges et al. 1995], GAL1 for galactose utilisation [Nehlin et al. 1991] and likely the INU1 gene in *K. marxianus* that encodes inulinase [Lertwattanasakul et al. 2011]. Inulinase allows *K. marxianus* to utilise either sucrose or inulin, a fructan stored at a high concentration in some plants [Chi et al. 2009]. Mig1 recruits the proteins Ssn6 and Tup1 to repress genes [Treitel and Carlson 1995]. The GAL1 gene was up-regulated in *K. marxianus* in a xylose medium, and INU1 was one of the most strongly

up-regulated genes [Schabort et al. 2016, Chapter 3]. Lertwattanasakul et al. [2011] also found two putative Mig1 binding sites in the INU1 promoter of the DMKU3-1042 strain.

Dramatic responses in the up-regulation of peroxisomes and β-oxidation were seen in the previous work [Schabort et al. 2016, Chapter 3], along with various other genes in central carbon metabolism. It has previously been shown by microarray studies with knockout strains of *S. cerevisiae* that Adr1 and Cat8 played a major role in the up-regulation of genes for utilisation of alternative carbon metabolism, as well as peroxisomal organisation and β-oxidation [Young et al. 2003]. The latter study compared the expression levels of genes in the knockout strains and the wild-type strain during the oxidative shift following the depletion of glucose. The affected targets of the SNF1 knockout also corresponded with the combination of affected targets of Adr1 and Cat8. For some of these targets, putative binding sites could be found based on consensus motif searches and chromatin immunoprecipitation (ChIP). It was found that Adr1 was more important than Cat8 [Young et al. 2003].

The consensus pattern for Adr1 has been described by Cheng et al. [1994]. The zinc finger Adr1 binds to a core motif [TGA][TC]GG[AG]G. It usually binds as a dimer in opposite directions and the full motif can be described as C[CT]CC[GA][TCA]N{2-36}[TGA][TC]GG[AG]G; the reverse is identical. The core recognition site for Mig1 has been described as [GC][TC]GG[GA]G (and C[CT]CC[AG][GC] in the opposite direction), also known as the GC-box [Nehlin et al. 1991]. However, a significant bias of the sites on the 5'-end of the GC-box to A and T has been described, the so-called AT-box [Lundin et al. 1982], resulting in a more precise motif [ATG][AT][AT][AT][ATG]N[GC][TC]GGGG and the reverse of this pattern is CCCC[GA][GC]N[TAC][AT][AT][AT][CAT]. This pattern was specified as four consecutive guanines (or cytosines in the opposite direction), not allowing an adenine at position 5, making it more distinct from the Adr1 motif. The zinc cluster TFs Cat8 and Sip4 bind sequences containing CC, followed by GCC, separated by a few nucleotides. The Cat8 motif has been described as YCCNYTNRKCCG and that of Sip4 as TCCATTSRTCCGR [Roth et al. 2004]. Rds2 recognises very similar sites as Cat8 and Sip4 [Soontorngun et al. 2007].

In this chapter, a mechanistic basis is given for the differential expression pattern in a xylose medium based on two approaches. Firstly, by the correspondence with target genes of main transcriptional regulators in glucose derepression, as identified previously in the model species and using enrichment statistics. Secondly, the enumerative method of over-represented heptamers was used to reveal putative transcription factor bindings sites in an unbiased manner. Methods for network mapping of

heptamers and for an Occam's razor approach to the enumerative method is also demonstrated, which may be valuable additions to data exploration in computational biology.

Materials and Methods

Strains and cultivation

K. marxianus UFS-2791 was cultivated in a defined mineral medium containing glucose or xylose in aerobic shake flasks. RNA was extracted in mid-exponential phase. Protocols were described in Chapter 3 and in Schabort el at. [2016].

RNA-seq data

RNA-seq data were generated in previous work for the strain UFS-Y2791 [Schabort et al. 2016, Chapter 3]. It was found that these data could be efficiently read-mapped to the recently published complete genome of a different stain, namely *K. marxianus* DMKU3-1042 [Lertwattanassakul et al. 2015] using TopHat [Trapnell et al. 2009]. Differential expression was calculated using CuffDiff [Trapnell et al. 2013]. Fold changes were defined as the value in the xylose medium divided by the value in the glucose medium, and only applied to genes with q-values below 0.05.

Gene set enrichment statistics

The set of genes targeted by Adr1 in *S. cerevisiae* were identified as those that failed to be up-regulated during glucose derepression in an Adr1 knockout strain [Young et al. 2003]. It was assumed that this gene set was the same also in *K. marxianus.* The hypergeometric distribution was used to estimate the probability of finding the same number or more of genes up-regulated (18) in a randomised draw of homologs (46) in *K. marxianus* in a background of 4 093 protein coding genes, with 323 being up-regulated.

Motif enrichment statistics using the enumerative method

All overlapping heptamers in the upstream regions of the *K. marxianus* DMKU3-1042 genome, up to 1 000 bp from the translation start site of all up-regulated genes, were counted. This number for each heptamer was compared to the number found for all genes by using the binomial distribution. Since the background numbers were very high, nearly identical results were obtained compared to when the hypergeometric distribution was used. The background was taken to be the count of the same heptamer divided by the number of all heptamers, in all upstream regions. The observed heptamer frequency was the count of any particular heptamer, and the number of draws taken to be the number

of heptamers in the upstream regions of all up-regulated genes. The binomial distribution results in a p-value for finding the same number or more of a heptamer by chance. The p-values from multiple comparisons were corrected by multiplying with the number of comparisons, taken as the total number of possible heptamers (16 834), resulting in q-values.

Motif matching

Heptamers were scanned against motifs from the JASPAR database [Sandelin et al. 2004]. Motifs were downloaded as positional probability matrices (PPMs) and the scoring performed in algorithms developed for *Reactomica* [Schabort et al. 2016, Chapter 3]. For PPMs that were either longer or shorter than a heptamer, the best local match was used. To calculate the motif score, the relevant scores in the PPMs were summed since the genomic context (such as distance from the translation start site), on which the background nucleotide frequencies depend, could not be used for a complete probabilistic methodology in this position independent method. The score was normalised by the maximum possible score that could be obtained, which was equal to the length of either the heptamer or the PPM, whichever was the shortest. Clustering of PPMs was done by calculating a distance matrix, based on the distance at the best alignment between a pair of PPMs. The distance matrix was used to construct a distance tree.

The Occam's razor motif

The hypothesis of the simplest explanation for over-represented sequences was used to explain the data of heptamer over-representation, and hence could be termed an Occam's razor approach to the heptamer frequency problem. The hypothesis states that the up-regulated response was generated by a single, strong-acting TF, which dominated over the effects of minor transcription factors in this particular response. The alternative model is that multiple transcription factors contributed significantly to this particular response. To construct the Occam's razor PPM, the top 30 over-represented heptamers were aligned, and as the reverse complement where relevant. No gaps were allowed during the alignment. Missing values on either side of the alignment were filled in as gaps. To calculate a frequency matrix, each observed nucleotide was counted as unity, while each gap was counted as one of each of the four nucleotides to simulate counts for the unknown bases. The counts were converted into a PPM, which is defined as the Occam's razor motif. This motif was compared to each of the short zinc finger TFs for identification. The process is show in Figure 1.

Figure 1. The method of converting over-represented heptamers into a PPM for the Occam's razor approach. The PPM is most closely matched the pattern for Adr1, followed by YPR022C and Mig1.

Sequence bias in neighbouring bases

Over-represented heptamers were mapped to the upstream regions of all up-regulated genes. For each match, the 30 bp upstream and downstream of the heptamers were extracted and joined to the heptamer, and a list compiled. The lists were converted to frequency matrices, and character bias and the G-statistic calculated.

Results

To test the hypothesis that the same major regulator (Adr1) activates genes under glucose derepression both in *S. cerevisiae* and *K. marxianus*, differential expression values between the two species were compared, where fold changes in the values for *S. cerevisiae* were due to the knockout of the Adr1 TF gene (observed under glucose derepression), whereas changes in the values for *K. marxianus* were due to differences in the culture media used, namely using glucose or xylose as carbon source. The genes identified by Young et al. [2003] to be targets of Adr1 and their fold changes (due to knockout, intact ADR1 divided by knockout adr1) are given in Table 1, along with fold changes in RNA-seq values of *K. marxianus* from this study, calculated as the value in the xylose medium divided by the value in the glucose medium. The vast majority of these genes were conserved between the two yeasts, allowing a comparison. Notably, out of the nine peroxisomal targets of Adr1 with homologs in *K. marxianus,* eight were up-regulated in the xylose medium. These included the enzymes of β-oxidation, fatty acid transporters and the peroxisomal catalase gene CTA1. The genes PXA2 (peroxisomal long-chain fatty acid import protein 1 / peroxisomal ABC transporter 2) and FAA2 (long-

chain fatty acid-CoA ligase 2 / long-chain acyl-CoA synthetase 2) were reported to fall just below the statistically significant level by Young et al. [2003], and these too were strongly up-regulated in *K. marxianus* according to RNA-seq data, at 21-fold and 84-fold, respectively.

Of the 16 genes involved in non-fermentative carbon metabolism with homologs, eight were up-regulated and four were down-regulated in the xylose medium. Notably, among these up-regulated genes were CIT3 (citrate synthase 3) and ICL2 (2-methylisocitrate lyase) of the methylisocitrate pathway. Also up-regulated were ACS1 of the pyruvate dehydrogenase bypass and required for acetate utilisation, as well as the mitochondrial aldehyde dehydrogenase gene ALD4. The genes encoding both enzymes responsible for glycerol utilization, GUT1 and GUT2, were also up-regulated in the xylose medium, and were thus likely regulated by Adr1.

Table 1. Comparison of differential gene expression in *K. marxianus* UFS-Y2791 and *S. cerevisiae* for putative Adr1 targets. In column 6, '1' refers to data from Young et al. 2003; '2' refers to data from Young et al. 2002. In column four, a value of 1 was used to indicate that there was no significant change.

gene	gene ID (*K. marxianus* UFS Y2791)	ADR1/adr1 (*S. cerevisiae*) [Young et al. 2003]	Xyl/Glc (*K. marxianus*) [Schabort et al. 2016]	Signi-fincant	ADR1 site (ChIP)	Function
Non-fermentative carbon metabolism						
FDH2	None	64	-	-	1	Formate dehydrogenase
FDH1	None	52	-	-	-	Formate dehydrogenase
CIT3	g3431.t1	10	293.2	yes	1	Citrate synthase
YML131W	None	9.1	-	-	-	Quinone oxidoreductase homolog
ACS1	g3271.t1	9	245.3	yes	2	Acetate-CoA ligase
ADH5	None	8.2	-	-	-	Alcohol dehydrogenase
GLO4	g1981.t1	7.7	1	no	-	Hydroxyacylglutathione hydrolase
ICL2	g2681.t1	7	28.2	yes	1	2-Methylisocitrate lyase
ADH2	g157.t1	6.8	0.004	yes	2	Alcohol dehydrogenase
DIC1	None	6.3	-	-	-	Dicarboxylate transport
ALD4	g2199.t1	5.5	83.0	yes	1	Aldehyde dehydrogenase
CYB2	g3255.t1	4.6	5.4	yes	-	L-Lactate dehydrogenase
YPL201C	None	4.5	-	-	-	Glycerol metabolism?
YPL113C	g539.t1	4.4	1	no	-	Lactate dehydrogenase homolog
ALD5	g2677.t1	3.5	0.006	yes	-	Aldehyde dehydrogenase
YCP4	g2686.t1	3.4	2.372	yes	-	Flavodoxin
OAC1	g1400.t1	3	0.18	yes	-	Oxaloacetate transport
YGR043C	None	2.6	-	-	-	Transaldolase homolog
YHL008C	g366.t1	2.6	1	no	-	Formate/nitrite transport

gene	gene ID (*K. marxianus* UFS Y2791)	ADR1/adr1 (*S. cerevisiae*) [Young et al. 2003]	Xyl/Glc (*K. marxianus*) [Schabort et al. 2016]	Signi-fincant	ADR1 site (ChIP)	Function
GUT1	g2372.t1	2.4	8.3	yes	2	Glycerol kinase
MSS2	g1103.t1	2.4	1	no	-	Cox1 pre-mRNA splicing factor
GUT2	g3514.t1	2.3	4.442	yes	-	Glycerol-3-P dehydrogenase
CTP1	g4889.t1	2.1	0.14	yes	-	Citrate transport
Peroxisome biogenesis and oxidation						
POX1	g3143.t1	55	103.8	yes	1	Acyl-CoA oxidase
SPS19	g103.t1	17	269.4	yes	-	2,4-Dienoyl-CoA reductase (NADPH)
CTA1	g679.t1	16	11.7	yes	2	Catalase
FOX2	g2436.t1	16	292.6	yes	-	3-Hydroxyacyl-CoA dehydrogenase, enoyl-CoA hydratase
POT1	g4245.t1	14	253.3	yes	2	Acetyl-CoA C-acyltransferase
YMR018W	None	6.5	-	-	-	Pex5 homolog; putative pts1 receptor
PXA1	g1855.t1	5.7	20.0	yes	-	Peroxisome ABC transporter
IDP3	None	5	-	-	-	Isocitrate dehydrogenase (NADP+)
DCI1	None	4.7	-	-	-	Dodecenoyl-CoA -isomerase
PEX11	g2902.t1	3.3	4.2	yes	-	Peroxisomal membrane protein
YOR389W	g141.t1	2.7	1	no	-	Pex21 interaction by two-hybrid
PCD1	g1340.t1	2.2	28.4	yes	-	Peroxisomal nudix hydrolase
Meiosis and sporulation						
ADY2	None	20	-	-	1	Transporter, nitrogen utilization
YPL033C	None	15	-	-	-	Meiosis
DMC1	g1900.t1	7	1	no	-	Meiotic recombination
ATO3	g680.t1	6.4	0.30	yes	-	Ammonia transport/Ady2 homolog
SPO20	None	5.6	-	-	-	Pro-spore membrane ɣ-SNARE
BNS1	g651.t1	3.2	1	no	-	Meiosis
SPS4	g1840.t1	3.1	1	no	-	Meiosis
CSM4	None	3	-	-	-	Chromosome segregation meiosis
SPR6	None	2.1	-	-	-	Sporulation
Amino acid transport and metabolism						
YLR126C	None	7.7	-	-	-	Gln amidotransferase motif
LEU1	g1737.t1	6.3	1	no	-	Leucine metabolism
ALP1	g3072.t1	4.9	1	no	-	Amino acid transport
BAG7	None	3.8	-	-	-	General amino acid permease
CAR2	g1365.t1	3.6	1	no	-	Arg metabolism
SSU1	None	3.3	-	-	-	Sulfite transport
YDR111C	None	3.2	-	-	-	Asp aminotransferase homolog
ARO9	g3678.t1	2.8	3.6	yes	-	Aromatic amino acid aminotransferase
BAT1	g2826.t1	2.8	0.33	yes	-	Branched amino acid aminotransferase

gene	gene ID (*K. marxianus* UFS Y2791)	ADR1/adr1 (*S. cerevisiae*) [Young et al. 2003]	Xyl/Glc (*K. marxianus*) [Schabort et al. 2016]	Signi-fincant	ADR1 site (ChIP)	Function
PUT4	g988.t1	2.8	1	no	-	Neutral amino acid transport
PTR2	g4948.t1,	2.5	-	no	-	Peptide transport
DAL3	g3892.t1	2.2	1	no	-	Allantoin Met/ureido metabolism
Transcriptional regulation and signal transduction						
YDL156W	None	4.6	-	-	-	Tup1 homolog
IME1	None	4.5	-	-	-	Meiosis transcription factor
SLZ1	None	4.3	-	-	-	Meiosis transcription factor
NRG1	g4637.t1	3.6	1	no	-	Glucose repression/invasive growth
GIP2	g3085.t1	3.6	1	no	1	Glc7-regulator
TEC1	g4465.t1	2.2	4.8	yes	-	TEA/ATTS family
RRD1	g3512.t1	2.1	1	no	-	PP-2A regulator
Other						
ETR1	None	3.2	-	-	1	Dodecenoyl-CoA δ-isomerase, Fatty acid biosynthesis
FOB1	g894.t1	2.9	1	no	-	rDNA recombination
DIA3	g3171.t1	2.6	0.048	yes	-	Acid phosphatase, pseudohyphal growth
DBR1	g3507.t1	2.6	1	no	-	RNA lariat debranching enzyme
ECM8	None	2.5	-	-	-	Cell wall organization and biogenesis
GSY1	None	2.4	-	-	-	Glycogen synthase
TIR1	None	2.4	-	-	-	Structural constituent of cell wall
TRF4	None	2.4	-	-	-	Mitotic chromosome condensation
GPT2	g1977.t1	2.1	1	no	-	Glycerol-3-phosphate O-acyltransferase, phospholipid Biosynthesis
YPC1	None	2.1	-	-	-	Ceramidase
BTN2	g2051.t1	2.1	1	no	-	Regulation of pH

By contrast, enzymes involved in meiosis and sporulation, amino acid transport and metabolism, transcriptional regulation, signal transduction and other processes that have been associated with Adr1 in the knockout microarrays study, were mostly not differentially regulated. Further, a substantially lower fraction of these genes were conserved between the strains as opposed to those in central carbon metabolism.

In total, the 46 homologs in *K. marxianus* had 18 up and 7 down-regulated genes, resulting in a highly significant enrichment score of 9.9 standard deviations away from a random mean, and thus a vanishingly small probability of being a randomised sample of 46 genes, given the background. Using only the peroxisomal and non-fermentative carbon metabolism targets resulted in an enrichment

score of 10.8. Adr1 is, therefore, a putative regulator of the glucose de-repressing response in the defined xylose mineral medium, considering experimental evidence from the model species.

A good correspondence was also found between the 40 most significantly up-regulated genes in *S. cerevisiae* under derepressing conditions [Young et al. 2003] and their homologs in *K. marxianus* (Tables 2 and 3). RNA-seq data were both read-mapped to the UFS-Y2791 draft genome from previous work [Schabort et al. 2016, Chapter 3] and to the complete genome of strain DMKU3-1042 from Lertwattanassakul et al. [2015]. Annotations proved very similar, as did the differential expression statistics. Of the 40 genes in *S. cerevisiae*, 27 and 25 genes could be mapped to the UFS Y2791 and DMKU3-1042 genomes, respectively. Of these, 63% and 64% were also very strongly up-regulated. Differences were found with PCK1, SFC1, YKL187C, YAT2, MPC3, IDP2, SUE1, which were constitutively expressed under both the glucose and xylose conditions. In addition, ATO3 and ICL1 were both down-regulated.

Table 2. The 40 most significantly glucose-derepressed genes in *S. cerevisiae* in comparison with the glucose-xylose response of homologs in *K. marxianus*. Transcriptomic data of knockout *S. cerevisiae* strains were from Young et al. [2003]. RNA-seq data from our study were both read-mapped to the UFS-2791 draft genome assembly and to the DMKU3-1042 complete genome assembly from Lertwattanassakul et al. [2015].

ORF	Gene	Gene (*K.marx* UFS)	Gene (*K.marx* DMKU)	DR/R (*S.cer*) [Young et al. 2003]	ADR1/ adr1 (*S.cer*) [Young et al. 2003]	SNF1/ snf1 (*S.cer*) [Young et al. 2003]	CAT8/ cat8 (*S.cer*) [Young et al. 2003]	Xyl/Glc RNA-seq (on *K.marx* UFS draft) [Schabort et al. 2003]	Xyl/Glc RNA-seq (on *K.marx* DMKU) [Schabort et al. 2003]	regulators predicted
YLR377C	FBP1	g2042.t1	gene4146	**130**	0.8	**36**	**7.4**	**27.3**	**25.4**	Cat8
YKL217W	JEN1	g4765.t1	gene1090	**98**	3.3	**28**	**2.1**	**156**	**145**	Cat8
YGR236C	SPG1			**92**	**15**	**5**	0.5			Adr1
YKR097W	PCK1	g3511.t1	gene2388	**92**	1.4	**180**	**3.6**	1	1	Cat8
YJR095W	SFC1	g2212.t1	gene4880	**78**	1.1	**170**	**6.2**	1	1	Cat8
YIL057C	RGI1	g3353.t1	gene4421	**77**	**16**	**8.9**	0.8	**3.1**	**3.2**	Adr1
YMR107W	SPG4	g369.t1	gene1771	**75**	2.3	**18**	0.5	**197**	**182**	
YCR010C	ADY2		gene842	**72**	**20**	**32**	**9.3**		**620**	Adr1,Cat8
YDR384C	ATO3	g680.t1	gene1415	**55**	**6.4**	**2**	**4.6**	**0.30**	**0.29**	Adr1,Cat8
YPL276W	FDH2			**50**	**64**	**5.1**	1.8			Adr1
YPR001W	CIT3	g3431.t1	gene4678	**50**	**10**	1.4	0.7	**293**	**302**	Adr1
YGL205W	POX1	g3143.t1	gene1095	**44**	**55**	0.8	**2.3**	**104**	**71.4**	Adr1,Cat8
YPR002W	PDH1	g3432.t1	gene4677	**44**	2.1	1.1	1	**32.8**	**33.8**	?
YAL054C	ACS1	g3271.t1	gene2271	**43**	**9**	**30**	**3.5**	**245**	**235**	Adr1,Cat8
YOR388C	FDH1			**35**	**52**	**4.6**	1.8			Adr1
YKL187C	YKL187C	g3857.t1	gene2946	**34**	2.7	**6.2**	**3.2**	1	1	Cat8

ORF	Gene	Gene (*K.marx* UFS)	Gene (*K.marx* DMKU)	DR/R (*S.cer*) [Young et al. 2003]	ADR1/ adr1 (*S.cer*) [Young et al. 2003]	SNF1/ snf1 (*S.cer*) [Young et al. 2003]	CAT8/ cat8 (*S.cer*) [Young et al. 2003]	Xyl/Glc RNA-seq (on *K.marx* UFS draft) [Schabort et al. 2003]	Xyl/Glc RNA-seq (on *K.marx* DMKU) [Schabort et al. 2003]	regulators predicted
YDR256C	CTA1	g679.t1	gene4095	**34**	**16**	**7.6**	**3.3**	**11.7**	**11.5**	Adr1,Cat8
YKR009C	FOX2	g2436.t1	gene4660	**33**	**16**	**2.4**	1.5	**293**	**285**	Adr1
YGR067C	YGR067C	g4524.t1		**33**	0.9	**39**	**7.1**	**16.2**		Cat8
YNL195C	YNL195C			**29**	**3**	**7.7**	0.8			Adr1
YER065C	ICL1	g64.t1	gene4423	**28**	0.4	**77**	**20**	**0.12**	**0.14**	Cat8
YER024W	YAT2	g135.t1	gene4133	**26**	0.6	**5.6**	**2**	1	1	Cat8
YGR243W	MPC3	g3290.t1		**26**	2.8	**6.1**	0.6	1		
YAR035W	YAT1	g3828.t1	gene181	**25**	0.9	**7.4**	1.5	**3.6**	**3.5**	
YMR206W	YMR206W	g4810.t1	gene4161	**25**	1.2	**5.9**	1	**7.2**	**8.1**	
YIL160C	POT1	g4245.t1	gene1055	**22**	**14**	1.5	0.4	**253**	**218**	Adr1
YPR150W	YPR150W			**22**	**4.3**	**2**	1.3			Adr1
YLR174W	IDP2	g4616.t1	gene1816	**21**	0.9	**16**	**2.5**	1	1	Cat8
YLR126C	YLR126C			**21**	7.7	1.5	1			?
YBR050C	REG2			**21**	1.5	**42**	**5**			Cat8
YEL008W	YEL008W			**21**	1.1	**7.5**	1.5			
YPR006C	ICL2	g2681.t1	gene4663	**21**	**7**	**4**	**2.1**	**28.2**	**31.2**	Adr1,Cat8
YHL032C	GUT1	g2372.t1	gene5110	**20**	**2.4**	**6.8**	**2.3**	**8.3**	**7.9**	Adr1,Cat8
YHR139C	SPS100			**19**	0.5	1	0.7			?
YPR151C	SUE1	g1464.t1	gene1368	**18**	**3.8**	1.9	1.2	1	1	Adr1
YLR267W	BOP2			**17**	0.4	0.9	0.1			?
YNR002C	FUN34/ATO2	g766.t1		**16**	1	1.3	0.6	**607**		?
YNL009W	IDP3			**15**	**5**	**2**	1.6			Adr1
YNL013C	YNL013C			**15**	3	1.6	0.1			?
YER179W	DMC1	g1900.t1	gene1006	**14**	**7**	1.8	0.9	1	1	Adr1

Table 3. Summary of the most significantly glucose-derepressed genes in *S. cerevisiae* in comparison with the glucose-xylose response of homologs in *K. marxianus*.

	ΔAdr1 (*S. cerevisiae*) [Young et al. 2003]	ΔSnf1 (*S. cerevisiae*) [Young et al. 2003]	ΔCat8 (*S. cerevisiae*) [Young et al. 2003]	Xyl/Glc RNA-seq (on *K. marxianus* UFS-Y2791 draft genome) [Schabort et al. 2003]	Xyl/Glc RNA-seq (on *K. marxianus* DMKU3-1042 genome) [Schabort et al. 2003]
total	40	40	40	27	25
up(significant)	19	29	17	17	16
down(significant)				2	2
%up	48	73	43	63	64
%down				7	8

Based on the transcriptome data, there is thus substantial evidence for the role of Adr1 but also Cat8 in the xylose response, given the RNA-seq data from glucose and xylose cultivations and considering the prior information of their role in glucose repression in *S. cerevisiae*, in particular for the peroxisomal genes. The next objective was to find evidence for the binding sites of these TFs in the regulatory regions of up-regulated expressed genes.

The enumerative approach

The statistical test for over-representation of k-mers in all up-regulated genes (323) with subsequent correction for multiple comparisons resulted in five heptamers with q-values below 0.05: CCCCACA, ACCCCAG, TGGAGAA, TTCCCCA and GCCGCGC. These were mapped to DNA binding motifs and are presented in Table 4. Notably, Adr1 and Mig1 were among the TFs mapping to these heptamers. Multiple PPMs may map to the same heptamer. A useful representation is given by the k-mer network in Figure 2. This map allows inspection of all motifs matching to a heptamer and reveals overlapping specificity. From the k-mer network it can be seen that the stress response factors Msn2 and Msn4, and Rgm1 map to all three heptamers, while Adr1 and Mig1 both map to only one heptamer, ACCCCAG. The similarity between the four over-represented heptamers is evident, and the challenge is to identify which TF best fits the heptamers. A pattern common to four of the five heptamers is a repeat of four cytosines, adjacent to adenine or thymidine, typical of binding by zinc finger proteins. The pattern TGGAGAA (or TTCTCCA on the opposite strand) is almost identical to TTCCCCA. Decreasing the threshold of matching between heptamers and PPMs to 0.5 revealed no additional matches between Adr1 and the four heptamers.

Table 4. Over-represented heptamers in upstream regulatory regions of 323 up-regulated genes, with a q-value smaller or equal to 0.05 after correcting p-values for multiple comparisons using the total number of heptamers (16 384) as multiplier. The minimum fractional match score between heptamer and PPM was set to 0.7.

JASPAR Matrix ID	TF	Heptamer	p	q	motif score/ potential	pattern
9680	YPR022C	CCCCACA	4.17E-09	6.83E-05	0.84	CCCCAC[CG]
9675	YML081W	CCCCACA	4.17E-09	6.83E-05	0.82	[AC]CCCC[GT]C[AT][CT]
9669	YGR067C	CCCCACA	4.17E-09	6.83E-05	0.81	[AG]CCCC[AG]C[AT][CT][CT][AGT][GT][CGT][AG]
9585	MSN2	CCCCACA	4.17E-09	6.83E-05	0.80	AGGGG
9586	MSN4	CCCCACA	4.17E-09	6.83E-05	0.80	AGGGG
9685	ZMS1	CCCCACA	4.17E-09	6.83E-05	0.79	T[AT]CCCCGC[AT]
9582	MIG2	CCCCACA	4.17E-09	6.83E-05	0.78	CCCCGC[ACG]

9583	MIG3	CCCCACA	4.17E-09	6.83E-05	0.77	CCCCGC[AG]
9610	RGM1	CCCCACA	4.17E-09	6.83E-05	0.75	AGGGG
9577	MET31	CCCCACA	4.17E-09	6.83E-05	0.75	[AG][CG]TGTGGCG
9578	MET32	CCCCACA	4.17E-09	6.83E-05	0.73	[AC]GCCACA
9576	MET28	ACCCCAG	1.83E-07	3.00E-03	0.83	CTGTGG
9585	MSN2	ACCCCAG	1.83E-07	3.00E-03	0.80	AGGGG
9586	MSN4	ACCCCAG	1.83E-07	3.00E-03	0.80	AGGGG
9610	RGM1	ACCCCAG	1.83E-07	3.00E-03	0.75	AGGGG
9657	USV1	ACCCCAG	1.83E-07	3.00E-03	0.73	[AT][AT][AT]TTCCCCCTGAA[CT][CT][AT][GT][GT][CG]
9581	MIG1	ACCCCAG	1.83E-07	3.00E-03	0.72	[AC]CCCC[AG]C
9512	ADR1	ACCCCAG	1.83E-07	3.00E-03	0.71	[AC]CCCCAC
9676	YNR063W	TGGAGAA	4.35E-07	7.12E-03	0.72	TCGGAGAT
9598	PDR8	TGGAGAA	4.35E-07	7.12E-03	0.71	[AG]CGGAGAT
9585	MSN2	TTCCCCA	5.99E-07	9.81E-03	0.80	AGGGG
9586	MSN4	TTCCCCA	5.99E-07	9.81E-03	0.80	AGGGG
9604	RDR1	TTCCCCA	5.99E-07	9.81E-03	0.77	TGCGGAA[AC]
9610	RGM1	TTCCCCA	5.99E-07	9.81E-03	0.75	AGGGG
9685	ZMS1	TTCCCCA	5.99E-07	9.81E-03	0.74	T[AT]CCCCGC[AT]
9555	HAL9	TTCCCCA	5.99E-07	9.81E-03	0.73	CGGAA
9607	REB1	TTCCCCA	5.99E-07	9.81E-03	0.71	[AG]TTACCCGG
9575	MCM1	TTCCCCA	5.99E-07	9.81E-03	0.70	CC[CT][AT]ATT[AG]GGAA
9534	DAL81	GCCGCGC	3.06E-06	5.01E-02	0.86	AAAAGCCGCGGGCGGGATT
9656	UME6	GCCGCGC	3.06E-06	5.01E-02	0.75	TCGGCGGCTAA[AT]T
9619	RSC30	GCCGCGC	3.06E-06	5.01E-02	0.74	[ACG][CG]CGCGCG
9597	PDR3	GCCGCGC	3.06E-06	5.01E-02	0.71	TCCGCGGA
9639	STP2	GCCGCGC	3.06E-06	5.01E-02	0.71	[CT][AG][AG][AT][CT]GGCGCCGCA[CT][CG][AC][AC][GT][AT]

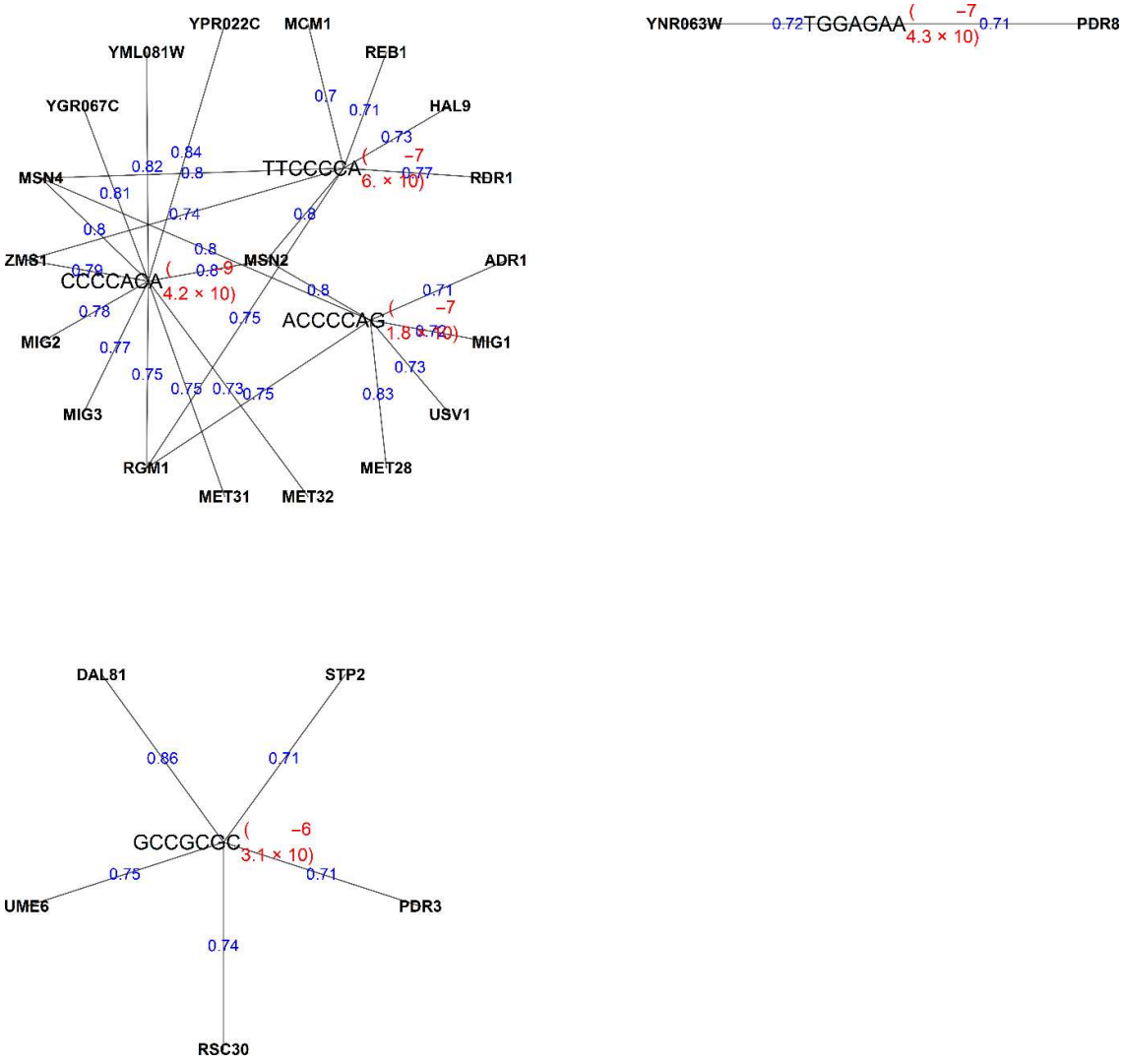


Figure 2. Heptamer network of over-represented heptamers in upstream regulatory regions of 323 up-regulated genes, with a q-value smaller or equal to 0.05 after correction for multiple comparisons using the total number of heptamers (16 384) as multiplier. The minimum fractional match score between heptamer and PPM was set at 0.7.

The best match to the CSRE motif of Cat8, with consensus CCGGA[AG], was a weak match to ACCCCAG with a matching score of 0.58, violating the consensus. Thus, Cat8 did not seem to be a major player in the glucose to xylose response. Other Snf1 dependent TFs Rsa2, Sip4 and Hap4 were absent from the list. The top four heptamers were so closely related that most, if not all, could possibly be matched by Adr1. A clearly different pattern was observed with the heptamer GCCGCGC, which had an over-representation score q of 0.05 after correction for multiple comparisons. The alignment of the top four is shown in Figure 3. Mig1 may also be relevant along with Adr1. The consensus core GC-box binding site for Mig1 was described as [GC][CT]GG[GA]G [Nehlin et al. 1991]. The reverse GC-box can

be described as C[CT]CC[GA][GC]. The best two heptamers in our analysis, CCCCACA, ACCCCAG, TTCCCCA, and TTCTCCA, were consistent with the reverse GC-box and with Adr1. The heptamers TCCCCGC and CTGGGGT (ACCCCAG) were very close to significance after the severe correction for multiple comparisons at q-values of 0.08 and 0.17, both matching Mig1.

K-mer	p after correction (q)	
--CCCCACA	6.8*10^-5	
-ACCCCAG-	3.0*10^-3	
TTCTCCA--	7.1*10^-3	
TTCCCCA--	9.8*10^-3	
C[CT]CC[GA][TCA]	Adr1	[Cheng et al. 1994]
CCC[GA][GC]N[TAC][AT][AT][AT][CAT]	Mig1	[Lundin et al. 1982]

Figure 3. Four heptamers over-represented in upstream regions of all up-regulated genes in *K. marxianus*, after correction for multiple comparisons, along with the consensus motifs for Adr1 and Mig1.

However, a significant bias of the sites on the 5'-end of the GC-box towards A and T has been described, the so-called AT-box, and the bias of the third pyrimidine was strongly towards G and not A [Lundin et al. 1982]. The forward consensus motif is, therefore, perhaps better described as [ATG][AT][AT][AT][ATG]N[GC][TC]GGGG and the reverse of this pattern is CCCC[GA][GC]N[CAT][AT][AT][AT][CAT]. On searching the list of top heptamers for an AT-box, it was surprising to find that almost none of the top 100 over-represented heptamers could be classified as AT-rich. Only poly-A and poly-T was in this list. Poly-T was very close to the significance threshold after correction for multiple comparisons at q = 0.056 ($p = 3.44\times10^{-6}$), while poly-A was at q = 1.3 ($p = 7.93\times10^{-5}$). It is important to mention that the simple correction used here could be interpreted as very severe. Either poly-A or poly-T may serve as an AT-box next to a GC-box. However, as the AT-box is less well conserved compared to the GC-box, the enumerative method might not reveal an AT-box.

The highest scoring heptamer by far was CCCCACA with a q-value of 6.83×10^{-5}. It was best matched by a lesser known zinc finger protein YPR022C, followed by YML081W, a probable TF TDA9 (Topoisomerase I damage affected protein 9), and zinc finger protein YGR067C. There is currently no conclusive evidence for the role of any of these in *S. cerevisiae*. Other matches were to better known TFs MSN2, MSN4, ZMS1, MIG2, MIG3, RGM1, MET31 and MET32 which are associated with various biological processes. A clustering of motifs is presented in Figure 4 to show the relationships among some of the zinc finger DNA binding motifs highlighted in this Chapter.

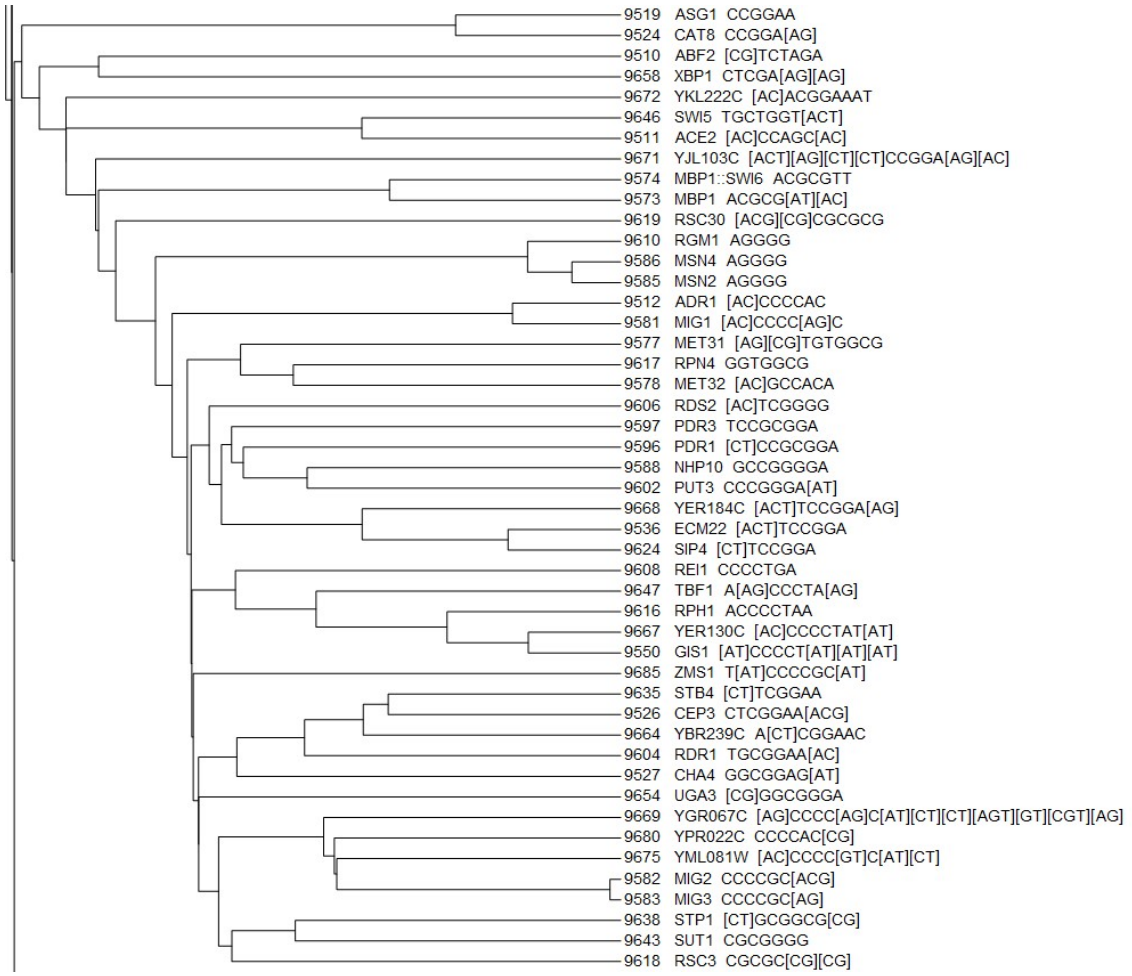


Figure 4. Similarity between selected PPMs. A local minimum distance was used as the criterion for clustering. The regular expressions indicate sequence specificity only, while clustering was performed using the local minimum distance criterion between PPMs. The regular expressions were generated by taking all high-scoring characters in each position, of which the total probability sums to at least 0.85.

Discovery of an Occam's razor motif

The short k-mer length rendered it difficult to decide on the best fit to a PPM. On close inspection it was found that in the top 30 heptamers, many contained a stretch of four cytosines, possibly with one replacement at the second C to a T, and that these, along with reverse complements, aligned well (Table 5). The process of constructing the Occam's razor PPM is illustrated in Figure 1.

Table 5. The top 30 over-represented heptamers in the upstream regulatory regions of the genes up-regulated in the xylose medium. The GC-box of the zink finger core binding site is indicated in bold, suggestive of Adr1 or Mig1.

K-mer	p	q
CCCCACA	4.17E-09	6.83E-05
A**CCCC**AG	1.83E-07	0.003
T**GGAG**AA	4.35E-07	0.007
TT**CCCC**A	5.99E-07	0.010
GCCGCGC	3.06E-06	0.050
TTTTTTT	3.44E-06	0.056
T**CCCC**GC	4.89E-06	0.080
TT**GGAG**A	5.41E-06	0.089
TGCTACG	6.26E-06	0.102
CT**GGGG**T	1.09E-05	0.179
TAT**GGGG**	1.22E-05	0.199
GCGACAG	1.22E-05	0.199
AA**CCCC**A	1.35E-05	0.221
TCTCCTC	1.97E-05	0.323
TGCCCAG	2.48E-05	0.407
CCAGGCA	2.66E-05	0.435
CCCCAGA	4.13E-05	0.677
TAGCAAA	4.32E-05	0.708
CCCCCCC	4.91E-05	0.804
CGCCGCG	5.29E-05	0.866
GTTGCTA	5.58E-05	0.914
GTGCGCC	5.75E-05	0.942
AAAAAAA	7.93E-05	1.299
T**GGGG**TA	1.00E-04	1.641
TGCGCCT	1.14E-04	1.874
TGGCACG	1.18E-04	1.925
TA**CCCC**A	1.25E-04	2.051
GTTACGT	1.28E-04	2.095
CAGGCAC	1.35E-04	2.205

Remarkably, the Adr1 motif closely resembled the PPM from combined heptamers. The strong bias to adenine at position seven was consistent with the description of the Adr1 motif, as were the allowance of a thymidine at position four and the slight bias towards A or T in the first two positions [Cheng et al. 1994]. Taking the combination of the top heptamers was necessary, as Figure 5 (top panel) shows that when the heptamer was taken on its own and mapped to the genome, there was a lack of sequence bias immediately upstream or downstream of the heptamer.

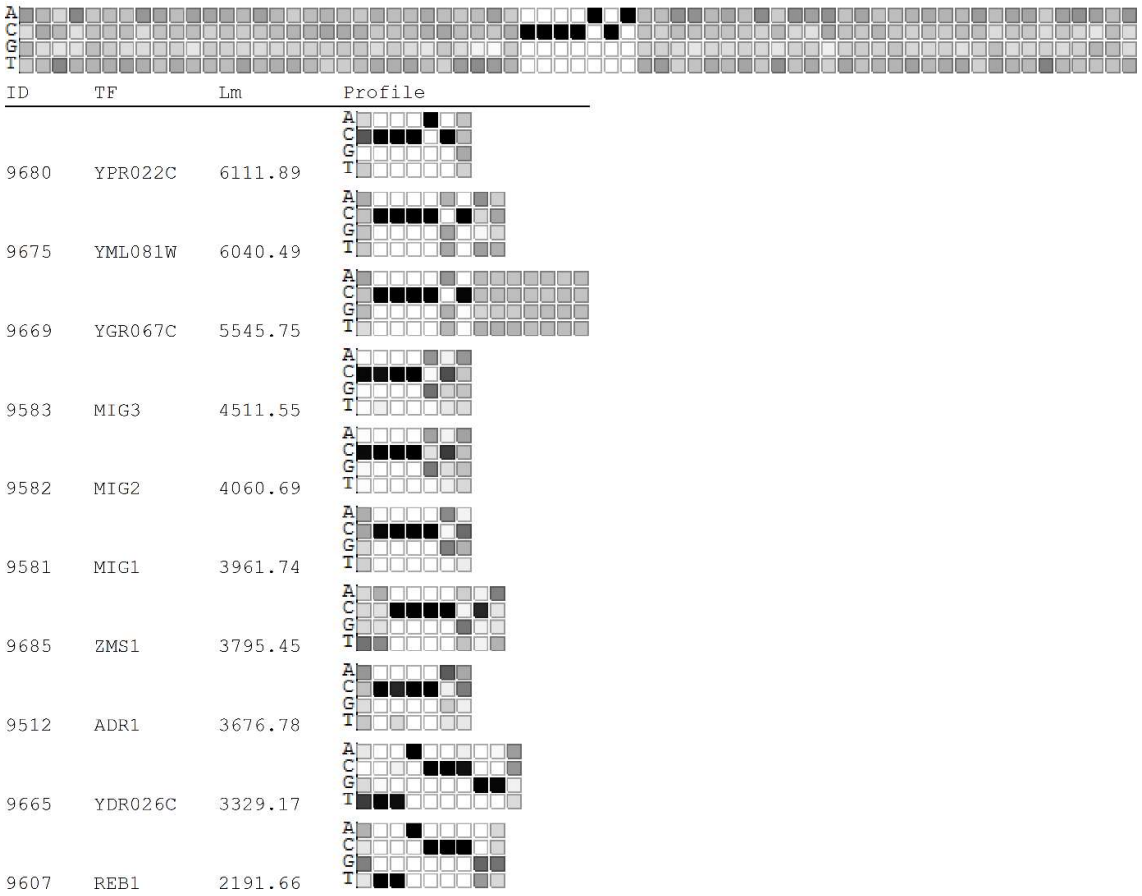


Figure 5. PPMs matching to the top scoring heptamer CCCCACA. The lack of sequence bias immediately upstream or downstream is evident in the top panel.

The Adr1 gene is strongly up-regulated in xylose medium

In support of the involvement of Adr1, its transcript levels were 36-fold up-regulated. Although the activity of TFs involved in glucose derepression in *S. cerevisiae* seem to be mostly regulated by post-translational modifications (Chapters 6 and 7), it is unlikely that such a high up-regulation in transcript levels would not lead to increased activity. To the contrary, CAT8 and RDS2 were constitutively expressed and so also were components of the HAP complex HAP2, HAP3 and HAP5, whereas HAP1 was down-regulated. It is interesting to note that in addition, SIP4 was up-regulated moderately at 5.5-fold and MIG1 was down-regulated 3.9-fold. The latter two TFs would likely play a role, but likely in fewer genes than Adr1. As is typical for the kinases, the gene of Snf1, which controls the activity of all these TF, was constitutively expressed in both the glucose and xylose media.

Sequence bias in neighbouring bases

The five over-represented heptamers were mapped to the upstream regions of the 323 up-regulated genes. Figures 6 to 10 show the alignments of the regions, along with character bias and the G-statistic. Although there was no bias immediately adjacent to the heptamers, which would help with more accurate TF identification, there was some evidence of bias towards the same Adr1-type GC-box pattern in the neigbouring regions. This pattern is suggestive of these TFs, likely Adr1, binding as dimers, although this was not conclusive. Notably, there was a lack of a putative AT-box, which would have suggested that Mig1 would be the TF and not Adr1. Also, the heptamer GCCGCGC seems to be located in GC-rich regions.

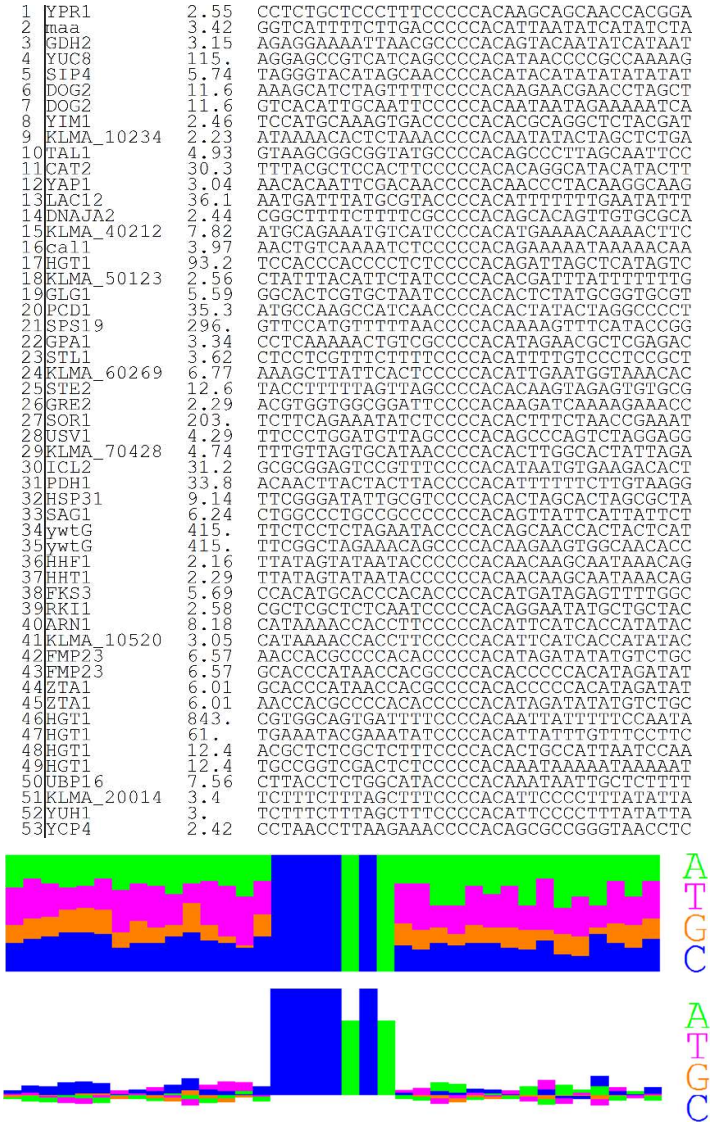


Figure 6. Sequence bias in bases neighbouring the heptamer CCCCACA. Bias for C towards the 5' side of the reverse GC-box suggests that the TF might bind as a dimer.

1	TY2B-GR2	2.55	ACAGAGGATAACAAAACCCCAGAAACCCCAGAACCTG
2	TY2B-GR2	2.55	AACAAAACCCCAGAAACCCCAGAACCTGCAGAGAAGA
3	JEN1	145.	CCGCAGTCTCATCAAACCCCAGGCAAGACCCCAGGCC
4	JEN1	145.	CAAACCCCAGGCAAGACCCCAGGCCAAATCCCCAGGC
5	GAL10	5.95	AGACTGTACTAACTAACCCCAGCCGATCGCCAGCTGG
6	GAL1	3.75	AGACTGTACTAACTAACCCCAGCCGATCGCCAGCTGG
7	YPR1	2.55	CACTAGTACCTCAGTACCCCAGTACCACCAGCACCAC
8	YPR1	2.55	GAAAATAACGCAGGAACCCCAGAAAAAACGGCCTTGA
9	KLMA_30050	3.34	GAAATATAGGGGTTTACCCCAGCCATACACCTAGAAA
10	MIF2	3.55	GCATACCCCAGGCATACCCCAGGCTATGCCTGGGCCA
11	MIF2	3.55	CCAGACCTGGGGCATACCCCAGGCATACCCCAGGCTA
12	SPG4	182.	ACAAGGGGCAACAATACCCCAGGGATACCCCTTTGGT
13	Acot10	3.54	GACAATGGGAGGGAGACCCCAGTCTGAATTAAGCCTA
14	STL1	15.	ATTTTATTTCTTTCAACCCCAGATTAATAGCATATAA
15	KLMA_30624	10.1	TGAAGTAAAGAATGAACCCCAGATCAAAGTCACGTGA
16	CAT2	30.3	CCCTCTCTCAAACCTACCCCAGGTTTAATGGAGAAAC
17	HOP1	4.66	AGGTTACCTTCCAATACCCCAGACTGCATATCTCTTT
18	KLMA_40412	4.1	AGGTTACCTTCCAATACCCCAGACTGCATATCTCTTT
19	Nqo2	2.61	TAACACTCTAGTAGCACCCCAGTTAAACCCCAGTTAA
20	Nqo2	2.61	TAGCACCCCAGTTAAACCCCAGTTAAACCCTAGTAAC
21	KLMA_50023	63.9	AATTAAGGAGTTTTGACCCCAGACTTTTTTCTTCAAA
22	PIR3	4.21	GTGGCAAAAAAAAAAACCCCAGCAAAAGAAACGTCAT
23	DBF2	3.23	CAAGTACCAAGTACCACCCCAGAACACCCCAGAAATG
24	DBF2	3.23	GTACCACCCCAGAACACCCCAGAAATGTTCTCCAACG
25	KLMA_50451	24.6	CGTAACCCAATTGGAACCCCAGAGTAGTTTTATTCCT
26	AXL1	5.5	AAATGCGATGAGCTAACCCCAGCTAGCCTCACAAACG
27	JEN1	3.22	TGTGAGTCGGACAAAACCCCAGTGGGGAAAATGCCGC
28	DIP5	3.86	CGGATTACAGGCCAGACCCCAGCATTAAATAGGTGGT
29	CTA1	11.5	ACCTTAGGAAATAATACCCCAGGTTGCAAGCCATATA
30	FBP1	25.4	AAAACGATAAGCGTCACCCCAGAGCCGGGAATCGGAT
31	PST2	3.36	TTATAGGGTTGTTATACCCCAGTTCCTTCCCTTCCTT
32	KLMA_70005	4.67	CTGTTGGCTTAAAAGACCCCAGACTTTGAACAGGAAG
33	KLMA_70030	2.87	GTTTCTCAACCTACTACCCCAGATCCTGATCGCCAAT
34	CLB4	2.33	TGCATACACATACATACCCCAGCATCTATATCTCAAA
35	MNN5	2.28	TGCATACACATACATACCCCAGCATCTATATCTCAAA
36	KLMA_80008	8.16	TGCTGCTGCTGCTAAACCCCAGTTTTCTGCCTGCCCG
37	HSP31	9.14	CACAGCCTGATCCAAACCCCAGTAATAATCATAATAA
38	KLMA_80060	48.6	AGGTGACTCTGGGCAACCCCAGTGAAGGAGGTGGGCG
39	HSP10	2.58	TAAGGGCCCTTCTGGACCCCAGACCTGGGGTCAGGAC
40	HSP10	2.58	TTAGTCGTGGTCCTGACCCCAGGTCTGGGGTCCAGAA
41	ywtG	415.	CAATCCAATCCAATTACCCCAGTATGGGGTTGAGCCG
42	ywtG	415.	GGAGATGGAGCCAGTACCCCAGGAAAATAGCGGATCT
43	HHF1	2.16	AGAATGCCCCGGGTAACCCCAGTTTCTCTGGGCAACC
44	HHT1	2.29	AGAATGCCCCGGGTAACCCCAGTTTCTCTGGGCAACC
45	INU1	90.4	AAATGGGGCGTTGTTACCCCAGGTATCCGGTTGTAGT
46	INU1	90.4	GTAGGTGCGGCAGACACCCCAGGAAAAATGCACACAG
47	FMP23	6.57	TAAACTCGTAAGGAAACCCCAGCAAGACAGATCGTCG
48	ZTA1	6.01	TAAACTCGTAAGGAAACCCCAGCAAGACAGATCGTCG
49	KLMA_10560	2.28	ATATACGACAGGGTAACCCCAGAGTTTGGTTTTCATG
50	SIP5	2.34	CCTCACACGAGCTCAACCCCAGCCTTGTTTCCAGGTC
51	SOU2	8.46	CTTTTCTTTGGGTCTACCCCAGATTTTTCGAGACCTA
52	ADR1	36.9	GCTGGGGCTTTTGCTACCCCAGACGAATTTTCGAAGG
53	GSM1	3.54	TATCAGGTTTCATAGACCCCAGTAAACAGCCTAACTT

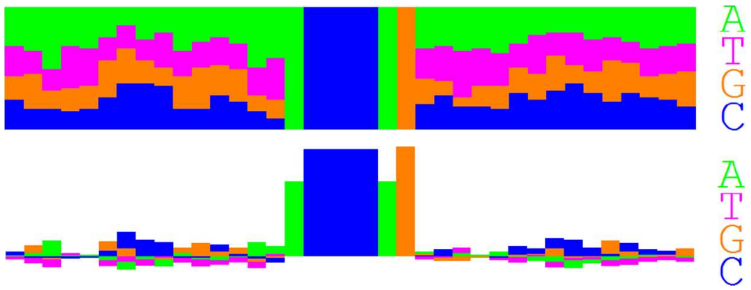


Figure 7. Sequence bias in bases neighbouring the heptamer ACCCCAG. Bias for C towards the 5' and 3' sides of the reverse GC-box suggests that the TF might bind as a dimer.

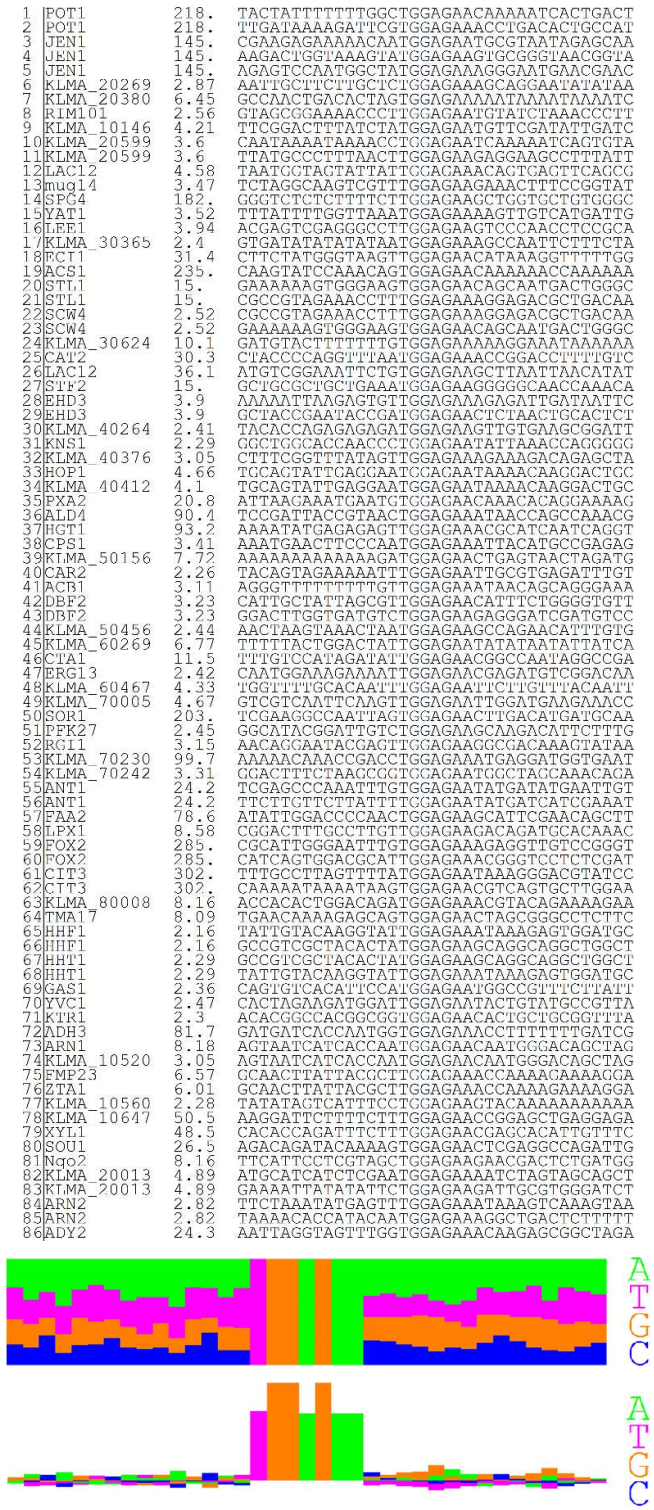


Figure 8. Sequence bias in bases neighbouring the heptamer TGGAGAA. Bias for C towards the 3' side of the GC-box suggests that the TF might bind as a dimer.

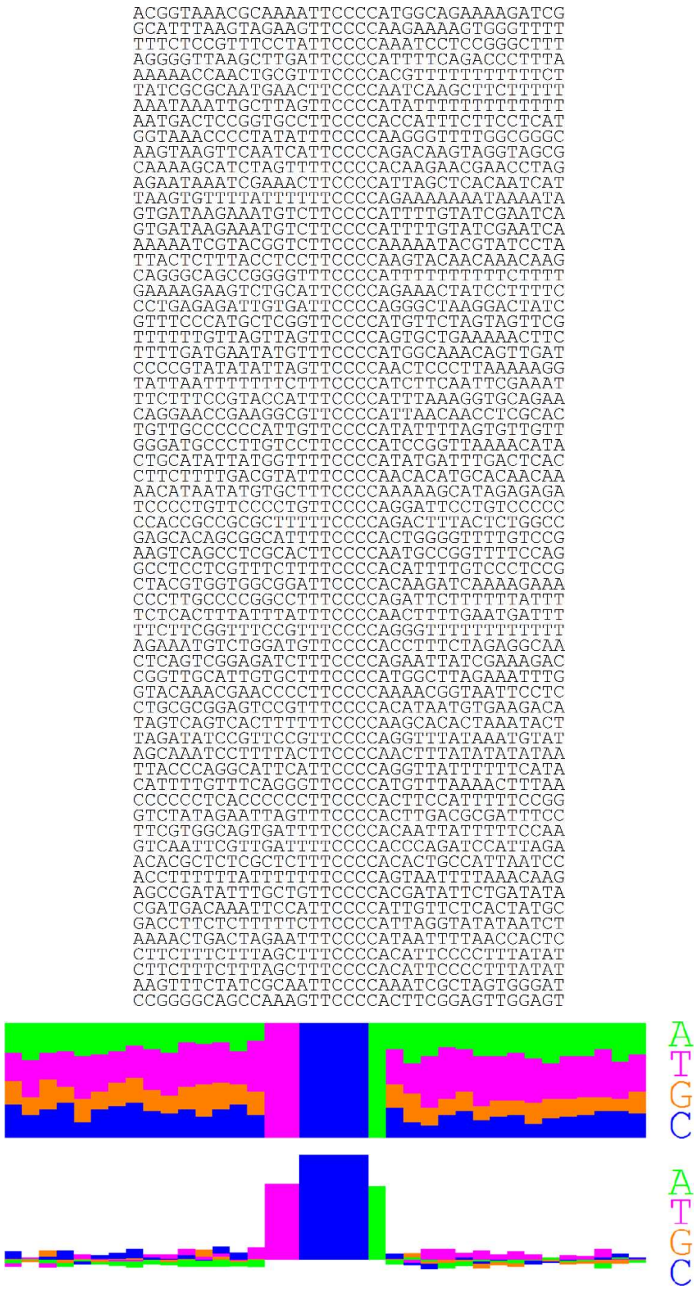


Figure 9. Sequence bias in bases neighbouring the heptamer TTCCCCA. No indication of bias is observed.

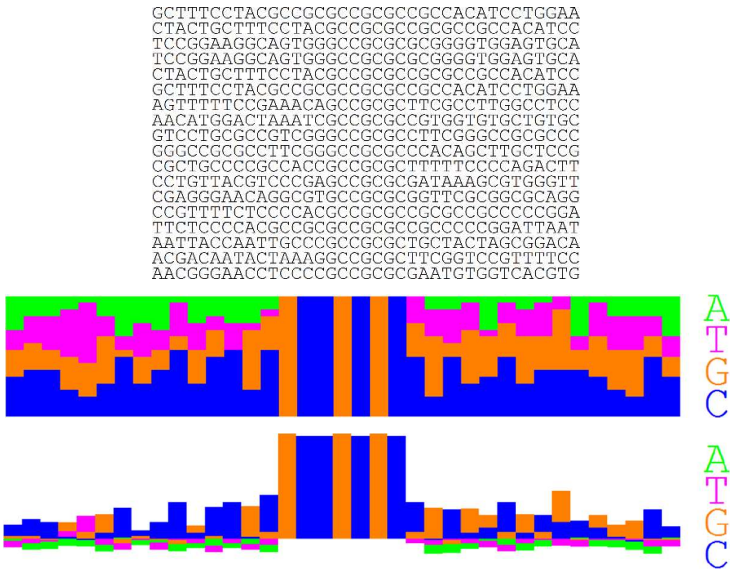


Figure 10. Sequence bias in bases neighbouring the heptamer GCCGCGC. The heptamer is found in a GC-rich region.

In *S. cerevisiae*, the glucose derepression response in genes for the utilisation of alternative carbon sources is regulated mostly by the action of the Adr1 and Cat8 TFs, which are in turn regulated by the Snf1 kinase [Young et al. 2003, Broach et al. 2012]. Evidence is provided here that also in *K. marxianus* this response is likely mediated by Adr1, with Mig1 having a lesser effect. Efforts to reveal TF binding sites are complicated by the fact that TFs bind only short sequences, with significant variation in the sequences they bind. Through both using the biological prior information of the target gene set in the model species, as well as the unbiased enumerative approach, Adr1 involvement was highlighted. Other TFs Cat8, Sip4 and Rds2 do not seem to make up the larger part of the response, although these may likely play a lesser role not revealed by the Occam's razor approach.

Assuming that the response to the absence of glucose and the presence of xylose was analogous to the condition of glucose depletion in *S. cerevisiae* at the end of the exponential phase [Young et al. 2003], there is evidence of transcriptional rewiring that has taken place during genome evolution. The biological explanation is not clear for most of the genes that were not up-regulated in *K. marxianus* as opposed to *S. cerevisiae* during glucose derepressed conditions. However, the genes PCK1 (phosphoenolpyruvate carboxykinase), SFC1 (succinate/fumarate mitochondrial transporter) and ICL1 (isocitrate lyase) are typical gluconeogenic enzymes, and their constitutive expression pointed to the requirement for an additional activator for some gluconeogenic genes, including those of the glyoxylate cycle. This activator might be Rds2, a known activator of gluconeogenesis [Soontorngun et al. 2007]. The targets of Rds1 include PCK1 and SFC1, as well as IDP2 (isocitrate dehydrogenase 2)

which was also not up-regulated, as opposed to the case in *S. cerevisiae* derepression. Among the putative Adr1 targets of *S. cerevisiae*, the best conserved genes were those of the peroxisome and central carbon metabolism, for which the vast majority of genes were also strongly up-regulated in *K. marxianus* grown in the xylose medium as opposed to the glucose medium.

Conclusions

There is substantial evidence that Adr1 plays a major part in the up-regulation of peroxisomal genes in a xylose medium, as well as of various other genes involved in the utilisation of alternative carbon sources. Although both biological prior information and genomics evidence corresponded with Adr1 as the main effector, the Occam's razor approach and heptamer counting in general suffered from the shortness of the binding site description. In particular, it can in principle not reveal TFs with relatively few binding sites and, therefore, misses other TFs. To this end, the higher length of some PPMs should be used to full advantage by motif scans. In addition, the wealth of modern experimental data on the model species *S. cerevisiae* may be used to reveal complete genome-scale gene regulatory networks. As part of the study, a likelihood framework was designed for this purpose and is the subject of follow-up work.

References

Barnett JA, Entian KD. A history of research on yeasts—9: regulation of sugar metabolism. Yeast. 2005;22: 835–894.

Broach JR. Nutritional control of growth and development in yeast. Genetics. 2012;192: 73-105.

Carlson M. Glucose repression in yeast. Curr Opin Microbiol. 1999;2: 202–207.

Cheng C, Kacherovsky N, Dombek KM, Camier S, Thukral SK, Rhim E, Young ET. Identification of potential target genes for Adr1p through characterization of essential nucleotides in UAS1. Mol Cell Biol. 1994;14(6): 3842-3852.

Chi Z, Chi Z, Zhang T, Liu G, Yue L. Inulinase-expressing microorganisms and applications of inulinases. Appl. Microbiol. Biotechnol. 2009;82: 211–220. doi: 10.1007/s00253-008-1827-1.

Hedges D, Proft M, Entian K. CAT8, a new zinc cluster-encoding gene necessary for derepression of gluconeogenic enzymes in the yeast *Saccharomyces cerevisiae*. Mol Cell Biol. 1995;15(4): 1915-1922.

Lertwattanasakul N, Kosaka T, Hosoyama A, Suzuki Y, Rodrussamee N, Matsutani M, Murata M, Fujimoto M, Suprayogi S, Tsuchikane K, Limtong S, Fujita N, Yamada M. Genetic basis of the highly efficient yeast *Kluyveromyces marxianus*: complete genome sequence and transcriptome analyses. Biotechnol Biofuels. 2015;8(47). doi: 10.1186/s13068-015-0227-x.

Lertwattanasakul N, Rodrussamee N, Suprayogi LS, Thanonkeo P, Kosaka T, Yamada M. Utilization capability of sucrose, raffinose and inulin and its less-sensitiveness to glucose repression in thermotolerant yeast *Kluyveromyces marxianus* DMKU 3–1042. AMB Express. 2011;1(20): 1–11. doi: 10.1186/2191-0855-1-20.

Lundin M, Nehlin JO, Ronne H. Importance of a flanking AT-rich region in target site recognition by the GC box-binding zinc finger protein MIG1. Mol Cell Biol. 1982;14(3): 1979-1985.

Nehlin O, Carlberg M, Ronne H. Control of yeast GAL genes by MIG 1 repressor: a transcriptional cascade in the glucose response. EMBO. 1991;10(11): 3373-3377.

Ratnakumar S, Young ET. Snf1 dependence of peroxisomal gene expression is mediated by Adr1. J Biol Chem. 2010;285(14): 10703–10714.

Roth S, Kumme J, Schüller H. Transcriptional activators Cat8 and Sip4 discriminate between sequence variants of the carbon source-responsive promoter element in the yeast *Saccharomyces cerevisiae*. Curr Genet. 2004;45(3): 121-128.

Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004;32:D91-D94.

Schabort DWP, Letebele PK, Steyn L, Kilian SG, du Preez JC. Differential RNA-seq, multi-network analysis and metabolic regulation analysis of *Kluyveromyces marxianus* reveals a compartmentalised response to xylose. PLoS ONE. 2016;11(6): e0156242. doi:10.1371/journal.pone.0156242.

Schöler A, Schüller H. A carbon source-responsive promoter element necessary for activation of the isocitrate lyase gene ICL1 is common to genes of the gluconeogenic pathway in the yeast *Saccharomyces cerevisiae*. Mol Cell Biol. 1994;14(6): 3613-3622.

Schüller H. Transcriptional control of nonfermentative metabolism in the yeast *Saccharomyces cerevisiae*. Curr Genet. 2003;43(3): 139-160.

Soontorngun N, Larochelle M, Drouin S, Robert F, Turcotte B. Regulation of gluconeogenesis in *Saccharomyces cerevisiae* is mediated by activator and repressor functions of Rds2. Mol Cell Biol. 2007;27(22): 7895–7905.

Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9): 1105-1111.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimental H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2013;7(3): 562–578. doi: 10.1038/nprot.2012.016.

Treitel MA, Carlson M. Repression by SSN6-TUP1 is directed by MIG1, a repressor/activator protein. Proc Natl Acad Sci USA. 1995; 92: 3132-3136.

Young ET, Dombek KM, Tachibana C, Ideker T. Multiple pathways are co-regulated by the protein kinase Snf1 and the transcription factors Adr1 and Cat8. J Biol Chem. 2003;278(28): 26146–26158.

Young ET, Kacherovsky N, Van Riper K. Snf1 protein kinase regulates Adr1 binding to chromatin but not transcription activation. J Biol Chem. 2002;277: 38095–38103.

Chapter 5

A likelihood framework for gene regulatory networks

Abstract

Previously in this study, a draft genome was assembled for *Kluyveromyces marxianus* strain UFS-Y2791 and a detailed transcriptomic analysis performed to study the differential genetic response in glucose and xylose defined culture media. It was also shown by using the enumerative approach of heptamer frequency comparisons that Adr1 was a likely regulator for the genes that were up-regulated in the xylose medium. This enumerative method lacks in that it can only reveal a small subset of potentially important transcription factors in the differential regulation, and that it does not make full use of the knowledge of DNA binding specificity of transcription factors. To this end, a likelihood method was developed in this chapter for the construction of gene regulatory networks. The likelihood based approach makes use of several sources of evidence for a given transcription factor-gene interaction. Likelihoods were assigned for motif strength, motif conservation among sister species, common Gene Ontology terms, and experimental evidence for the interaction in the model species *Saccharomyces cerevisiae*. Finally, reporter transcription factor enrichment was used to reveal the differentially active transcription factors. The transcription factors Gcn4 and Gcr2 were elucidated as the most significantly differentially active regulators associated with genes that were down regulated in the xylose medium. This observation is consistent with prior knowledge of functioning of these activating transcription factors. It was notable that the Gcn4 gene itself was constitutively expressed, suggesting that the method is superior in detecting differential activity of transcription factors as opposed to the less mechanistic reverse engineering methods.

Introduction

The study of gene regulation in model organisms such as *Escherichia coli*, *Saccharomyces cerevisiae* and humans is an active field of research. Many different approaches have been proposed to construct gene regulatory models. The most established method is the reverse-engineering approach in which multiple gene expressions datasets such as microarrays are used to discover both the underlying interaction network and to fit parameters of activation or repression. The reverse-engineering approach may be based on dynamic Bayesian networks [Sachs et al. 2005], regression models [Haury et al. 2012] or mutual information [Basso et al. 2005, Wang et al. 2009]. An international modelling challenge, the DREAM5 project (Dialogue on Reverse Engineering Assessment and Methods) was organised to compare the prediction accuracies of a number of methods in a competition [Marbach et al. 2013]. It was found that it was much more difficult to infer eukaryotic networks in the reverse engineering approach compared to bacteria. This is most probably due to the multiple layers of gene regulation found in eukaryotes, which includes regulation of chromatin condensation. A shortcoming of the reverse engineering approach is that the level of activity of a transcription factor (TF) cannot necessarily be predicted by either the expression level of its mRNA [de Sousa et al. 2009] or even that of the protein, while differential expression of the transcription factor gene itself is required to make an inference of an interaction in these methods. TFs, like all proteins, are post-translationally modified by mechanisms such as phosphorylation which alter their state of activity, and accordingly a linear relationship between the measured transcript levels and the activity of the TF does not exist, as seen in the expression levels of its targets. Another drawback of the reverse engineering approach is also that the method relies heavily on a large number of transcriptome datasets, making it a costly exercise. Another potential pitfall is that many of the putative interactions from large-scale transcriptome-based network construction may arise from secondary transcriptional effects. For instance, the gene knockout of a TF may affect another major TF, resulting in false assignment of a TF-target interaction. Most important, it has also been shown that co-expression is highly confounded by correlations with master regulators such as PCNA and other cell cycle related proteins [Venet et al. 2011]. It is thus imperative to approach the problem by using mechanistic information, such as the presence of DNA binding motifs, to its full extent, as well as evidence of direct physical interaction between a TF and its target DNA.

For model species, many years of research have led to the elucidation of specific interactions between TFs and their targets. High-throughput methods such as Chip-chip and Chip-seq, producing large datasets of direct physical TF-target interactions, have also been developed in recent years [Bulyk

2006]. Other important data types include microarray analyses of TF knockout strains, in which the differentially expressed targets become likely candidates as targets for the knockout TF [Harbison et al. 2004, Hu et al. 2007]. Such datasets are captured in large databases such as the *Saccharomyces* Genome Database (SGD) [Cherry et al. 2011]. These above-mentioned datasets are so large that the quality of the content might be underestimated simply due to the size. These hold tremendous potential for integrative network analysis, revealing new insights into regulation [Wang et al 2012, Gitter et al. 2009]. A gap in methods to fully employ such data into a compact representation of regulatory interactions currently exists, and especially to reveal how they relate to known pathways and processes.

For non-model species, including the industrially important yeast *Kluyveromyces marxianus*, genome-wide gene regulatory design remains unexplored, as it is for the vast majority of non-model species. To enable engineering of such species for the efficient production of biochemicals, a more detailed understanding of their gene regulation is required. Discovering individual interactions by knockouts or gel-shift assays might be the most accurate, but is, however, not cost effective at the genome scale, nor will such a project be completed in a reasonable amount of time. High-throughput experiments are a better starting point; however, such experiments are also very expensive. A different approach may be to use the wealth of information on model species, including the high-throughput data available for these [Cherry et al. 2011], and by using the evidence for DNA-binding sites in the species of interest together with the proper statistics to infer the gene regulatory interactions in the species of interest. This may be the best starting point for a species with a newly sequenced genome. Such a gene regulatory network can already make predictions about TFs that are important in a certain response. Ideally, master regulators may be found which may be manipulated to cause a drastic shift towards a more preferable metabolic pattern [Sonderegger et al. 2004]. For instance, finding a regulator that might increase fermentative metabolism in Crabtree negative, xylose utilising yeasts would be an attractive outcome for biofuel production.

The first aim of this chapter was to develop a method for constructing a gene regulatory network in a non-model species (*K. marxianus*) by using the draft genome of the UFS-Y2791 strain of *K. marxianus* (Chapter 2), predicting TF binding sites, and combining this information with high-throughput datasets of TF-target interactions in a model species (*S. cerevisiae*) to construct a gene regulatory network. This required a significant amount of code development to perform genome-scale motif finding, handling large datasets, and especially to control every step of the integration process. The software developed here was developed to enable the use of a fragmented draft genome, as would originate from

sequencing and *de novo* assembly of a new genome. The gene regulatory networks that were constructed were based on a statistical likelihood framework that combines multiple sources of evidence.

The second aim was to elucidate TFs that were likely important in the differential genetic response from glucose to xylose as the carbon source. For this, the gene set enrichment approach was followed, otherwise known as the reporter TF method [Patil and Nielsen 2005], but using high-quality RNA-seq data. Instead of using the correlations between TFs and targets in transcriptomic data as a means to elucidate the regulatory interactions, the network was derived by a more mechanistic explanation based on the DNA binding motifs and experimental evidence of an interaction in the model species, while the RNA-seq transcriptomic data was used to select the best model based on enrichment statistics.

It is important to note that the gene set enrichment approach taken here to draw the conclusions about active TFs is inherently robust to potential errors that may arise in the assignment of TF-target interactions, as multiple genes in the target gene set contribute to the conclusion. The conclusion about the differential activity of a TF is more important than the presence of any individual interaction. In the same fashion of other chapters, the networks are visualised to gain insight into both global regulatory features and regulation of individual pathways. This chapter focusses on the methods developed and the results are based on the initial draft genome [Schabort et al. 2016].

Materials and Methods

RNA-seq data

RNA-seq data were generated in previous work for the strain UFS-Y2791 [Chapter 3, Schabort et al. 2016]. The data were mapped to the draft genome of *K. marxianus* UFS-Y2791 [Chapter 2] using TopHat [Trapnell et al. 2009]. Differential expression was calculated using CuffDiff [Trapnell et al. 2013].

Motif scans

Motif scans were performed for all regulatory regions in *K. marxianus* UFS-Y2791, defined as the 1 000 bp upstream of a gene, which were not allowed to overlap with a neighbouring gene, allowing a match in either direction. A motif likelihood *Lm* was calculated for a motif match by calculating the ratio of the probability of the sub-sequence matching the model, divided by the probability of a sub-sequence

matching the background model of nucleotide frequencies in the upstream regions, specific to that distance from the translation start site (TLSS). Each motif likelihood score *Lm* was calculated as the product of independent occurrences as below, where *m* is the probability from the PPM matching to the character at position *i*, and *b* is the background frequency of the character at position *i*.

$$L\mathrm{m} = \frac{\prod_i^n m\,[i]}{\prod_i^n b\,[i]}$$

In this likelihood formulation, the background frequency of each nucleotide A, C, G and T is taken into account to give an estimate of the statistical significance of a motif match. The background nucleotide frequencies were calculated in a sliding window of 30 base pairs in all upstream regions at various distances from the TLSS. Effectively, the motif score was normalised to the background frequencies at the relevant distance from the TLSS. In this likelihood formulation, the background frequency of each nucleotide A, C, G and T is taken into account to give an estimate of the statistical significance of a motif match. Moreover, the nucleotide composition changes as a function of distance from the translation start site (TLSS). For convenience, an interpolating function was fitted for the background frequency of each nucleotide base as a function of position from the TLSS, in which the distances are negative values. Values further than 1 000 bp from the TLSS were assumed to have the same frequency as at 1 000 bp.

The JASPAR database [Mathelier et al. 2014] contains a large number of DNA binding motifs, applicable to many different species and represented in the form of positional probability matrices (PPM). Of these, 177 are specified as fungal DNA binding sites, mapping to fungal transcription factor proteins from *S. cerevisiae*. Some PPMs in the database have a high degeneracy (high motif entropy) and are thus not very restrictive with regard to the sequences they bind. Others in the database have been specified as highly precise. The latter case could have arisen from a very small dataset containing the few DNA sequences observed to be bound by a TF. Absolutely precise PPMs (consensus sequences) would likely result in too a conservative matching criterion. Pseudocounts are usually applied in such a case, in which the non-unity nucleotide scores are altered to a small number to allow close matches. The choice of the pseudocount is, however, not arbitrary, and its effect has been previously explored and an optimal value estimated as a pseudocount for a given motif entropy [Nishida et al. 2008]. A proper algorithm to decide on the choice of the pseudocount was designed as part of the work in this chapter, based on the motif entropy. However, since the accuracy with which each of the PPMs have been constructed is not evident, pseudocounts for the motifs were not included.

In effect, (*a*) the motif entropy at each position in the PPM, (*b*) the length of a PPM, and (*c*) the background nucleotide frequencies all contribute to the final likelihood score. To demonstrate this effect and to be able to understand the results obtained from the genome-wide motif scan, a simulation was performed. *Sn* was defined as the average nucleotide score at any position along the length of the PPM. High values of *Sn* (such as 1) would correspond to highly precise PPMs (low entropy). Simulated values for *Lm* were calculated as follows, where *n* is the length of the PPM and *b* was assumed to be 0.25, which assumes that background nucleotide frequencies were all equal.

$$Lm = \frac{\prod_i^n Sn}{\prod_i^n b}$$

Lm was calculated for every PPM length in the JASPAR database, from 5 to 21 bp, and at various values for *Sn*, from 0.3 to 1 (see Addendum 3, Figure 1). Based on these results, an initial cut-off for *Lm* was decided at 10 in order to save computational time.

It is important to note that transcription of a gene is directional, and intergenic regions should be chosen according to direction of transcription. This was one of the most complicated aspects in this work, since positions with respect to the direction of transcription and those of the contig needed to be kept track of. Motifs were scanned both against the forward and reverse direction of each intergenic region. To limit the amount of data captured in this initial phase, only putative motifs with an *Lm* at or above 10 were considered. Only those within 1 000 bp from the TLSS were further considered and no lower limit for the size of an intergenic region was set. It was not considered a good strategy to use the number of times a motif occurred for the same gene as a measure of likelihood, as one could assume that a single TF bound to the DNA in the regulatory region should be sufficient to drive or suppress expression.

Construction of a gene regulatory network

All data processing and visualisation were performed in algorithms developed for *Reactomica* [Chapter 3, Schabort et al. 2016] implemented in the Wolfram language. A likelihood framework was developed to incorporate multiple sources of evidence, which is partly based on the idea of a naïve Bayesian network using Bayesian classifiers. A Bayesian classifier calculates the likelihood that one hypotheses (H_1) is more likely, given the data, compared to a competing hypothesis (H_0) [Duda et al. 2001, Jansen et al. 2003, Collins et al. 2007, Chapter 1]. The likelihood *L* for an interaction was calculated as below where *Lm* is the motif likelihood, *Lc* the likelihood of motif conservation among sister species, *Lg* the Bayesian classifier for common GO terms, and *Li* the likelihood set when observing the same interaction in the model species or not.

$$L = Lm \times Lc \times Lg \times Li$$

All observations were ranked by the final likelihood and the best interactions chosen, where the number may be some estimate of the total number of interactions, such as 10 000. This naïve Bayesian formulation is more suitably called a likelihood rank ratio, as the true number of interactions are unknown and some of the classifiers used are arbitrarily set. Therefore, the values might not be interpretable as true likelihoods in the strict probabilistic sense. The method of calculating *Lm* was described above. For *Lc*, seven *Kluyveromyces* genomes were aligned using the progressiveMauve multiple genome aligner [Darling et al. 2010] and the alignment segments converted to conservation scores as described in Addendum 2. *Lc* was calculated as the ratio of the conservation score in the frame of a putative motif in *K. marxianus*, divided by that found in two 20 bp neighbouring regions around the motif, each normalised by the number of base pairs in the frames, as the two competing hypotheses. For calculating the appropriate *Lg* values as a function of the number *n* of GO terms in common between a TF protein and its target protein, a Bayesian classifier was used as below using Gold Standard positive and negative datasets, similar to what has been done for protein interaction datasets [Jansen et al. 2003, Collins et al. 2007, Chapter 1].

$$sensitivity = \frac{TP}{Positives}$$

$$false\ positive\ rate = \frac{FP}{Negatives}$$

$$Lg[n] = \frac{sensitivity}{false\ positive\ rate}$$

TP (number of true positives) indicate the number of TF-target pairs included in the Gold Standard (GS) positive dataset, and which had *n* GO terms in common. FP (number of false positives) indicate the number of TF-target pairs included in the GS negative dataset, and which had the same number of *n* GO terms in common. The set of TF-target interactions in SGD was considered as the GS positives, while the GS negatives were generated by assigning false targets to the same set of TFs. The *Lg* value at each number of common GO terms describes the likelihood ratio of a correct classification versus and incorrect classification. In this way, a likelihood *Lg* could be assigned as a function of the number of GO terms as obtained from the UniProt annotations. To counteract the effect of the small sample sizes for high values of *n*, the cumulative function of Li values was considered to be *Lg*, at a small expense of accuracy (see Results). For the likelihood based on conservation of an interaction in the models species *Li,* a simple scoring system was designed that weights the various sources of experimental evidence towards direct physical evidence (see Results). The methods of calculating *Lc*, *Lg* and *Li* are further explained in the Results section.

To construct a network, the *n* top scoring motifs were allowed to be included as TF-target interactions, taking only one motif per interaction. Variations in the calculation of the final likelihood based on *Lm*, *Lc*, *Lg* and *Li* were noted in the Results section and the final network based on the draft genome was constructed using the formula below.

$$L = Log[10, Lm] \times Lc \times Lg \times Li$$

Likelihood *Lc* based on sequence conservation among sister species

Motif conservation among sister yeast species was calculated in a relatively simple manner from the processed multiple genome alignment (see Addendum 2 for details). A convenient alignment viewer was designed to inspect motif matches as well as their sequence conservation in and around the motif region. The sequence conservation scores and the gap marker scores were calculated as described in Addendum 2, and is based on the alignment of the UFS-Y2791 genome as well as six other *Kluyveromyces* genomes. Average conservation scores were calculated as the sum of the values in these vectors wherein the two background vectors were combined, and normalising both by the number of bases. This method automatically corrects for the effect of fewer bases in cases where the flanking regions were at the edge of a contig (see Addendum 3 for further details). To convert conservation scores into a likelihood ratio for conservation, the average conservation score in the motif range C_m was divided by the average background conservation score C_b.

$$Lc = \frac{Cm}{Cb}$$

C_b was calculated from the combined background flanking regions, upstream and downstream of the motif, each consisting of 20 base pairs. It was considered to be a good strategy to treat this likelihood ratio partly as a qualifier, by setting a minimum *Lc* value at 0.5. This cut-off value would function similar to a pseudocount in motif scans, which avoids too strict a penalty. Another approach was to simply take the conservation score inside the motif regions for *Lc*, which is a value from 0 to 7.

Likelihood *Lg* based on GO terms in common

In order to estimate likelihood values *Lg* for GO terms that are common to the transcription factor and the target, a Bayesian classifier was calculated based on Gold Standards. The Gold Standard datasets were obtained by considering the TF-target interactions in the SGD as the Gold Standard positive dataset, and the GS negative dataset was generated by assigning false targets to the same set of TFs. In this way, a likelihood *Lg* could be assigned as a function of the number of GO terms in common with a TF protein and its target protein, as obtained from the UniProt annotations. Using four GO terms in

common applied to a very small fraction of examples (0.4%), thus involving a small sample size, could lead to a skewed representation. To obtain a smoother function and to allow any number of common GO terms, the cumulative scores were calculated as *n* GO terms or more in common. Herein, all values higher than what was obtained at four GO terms in common was taken as the value for four terms in common. During the final likelihood calculation for a motif in the DNA, the relevant *Lg* was taken from the distribution.

Likelihood *Li* based on occurrence of an interaction in the model species

In *S. cerevisiae*, 32 311 regulatory interactions have been captured in SGD. This very large interaction network was constructed mostly from microarray transcriptomics experiments. Code was developed for *Reactomica* to import these interactions, along with other types of interactions from YeastMINE, which is generic to other instances of InterMINE. When mapping the proteins encoded by these genes to proteins in *K. marxianus* by a conservative reciprocal BLASTP match with E-value cut-off of 1E-6, 13 799 interactions resulted, involving 119 TFs and 3 377 target genes. Essentially, this is the list of interactions that would be obtained regardless of a motif present or any other supporting evidence for an interaction in *K. marxianus*. This list was used to look up whether a TF-target interaction (from motif matching) also existed in *S. cerevisiae*.

Multiple types of high-throughput experiments were captured in the SGD [Cherry et al. 2011]. Some sources of evidence might be considered stronger than others. Direct physical evidence might be stronger evidence as opposed to microarray type of experiments which could result in many false positives due to secondary transcriptional effects. Since the true set of interactions is not known for *S. cerevisiae*, a true GS positive dataset does not exist and a strict Bayesian classifier approach could not be taken. Rather, a pragmatic approach was taken by assigning likelihoods based on a scoring system, weighted towards evidence for direct physical hier kort iets – physical wat? (See Table 1 in Addendum 3). The confidence spectrum ranged from weak support ("microarray RNA expression") to strong support ("chromatin immunoprecipitation-chip assay" and "chromatin immunoprecipitation-chip assay").

Combining all sources of evidence and construction of regulatory networks

The overall likelihood *L* was subsequently calculated for each motif-target pair by the following equation:

$$L = Lm \times Lc \times Lg \times Li$$

The list of final likelihoods was sorted and the interactions with the top *n* scores taken to construct a regulatory network. As an innovation, the enrichment statistic was used as a measure of confidence in any regulatory network constructed. The enrichment statistic of the target gene set of each TF is a measure of the total differential expression of the targets in comparison with the background differential expression, based on the RNA-seq data [Patil and Nielsen, Chapter 3, Schabort et al. 2016]. If it was assumed that a TF functioned mainly alone in a differential response, it is logical that the larger the enrichment statistic of a differentially active TF, the more consistent the network would be. At the same time, not all TFs would be active in the differentially expression response, and it would be better to find networks in which only a few TFs were active (with large enrichment scores), with the majority of TFs having low enrichment scores. Thus, the best network was chosen as the one with the highest enrichment statistic for any TF. Convenient methods were developed to explore the relationship between the enrichment statistic, number of TFs and the number of targets in a TF network. To calculate enrichment statistics, the method of simulated background distribution using randomly picked gene sets (the Z-score method) was used [Patil and Nielsen 2005, Schabort et al. 2016] where the enrichment score was calculated as below.

$$\mathrm{S} = \frac{\mathrm{Z(total, Test)} - \mathrm{Mean(Z, Background)}}{\mathrm{Standard\ deviation(Z, Background)}}$$

The Z-score was calculated from the q-values from the CuffDiff output [Chapter 3, Schabort et al. 2016]. Note that the enrichment statistic is independent from the number of targets in the gene set, as the background enrichment statistic was calculated to incorporate a mean and a standard deviation for each possible number of genes in a gene set [Patil and Nielsen 2005].

Results and Discussion

The scoring system used to construct the regulatory network and the resulting metrics are reported first, followed by optimisation of the network and extraction of differentially active transcription factors based on statistical enrichment.

Development of a scoring system based on multiple sources of evidence and likelihood ratios

In the motif scoring procedure, three parameters strongly affect the confidence or strength of the assignment of the motif: a) motif strength, which is the opposite of motif entropy, b) motif length, and c) the background nucleotide frequencies in the genome and its comparison with the motif of interest.

Figure 1 is a graphical overview of the 177 PPMs found in the JASPAR database in terms of the strength of the patterns. Figure 2 shows the distribution of PPM lengths of all 177 PPMs in the JASPAR database for fungi. A substantial fraction of these are short motifs of only five to nine base pairs in length. Figure 3 depicts the background frequency of each nucleotide base as a function of distance from the TLSS, using the draft genome of strain UFS-Y2791. It can be seen that the 200 bp upstream from the TLSS (signified by 800 bp – 1 000 bp in Figure 3) is highly biased, which encodes the 5' untranslated region on mRNA and the core promoter. Each of the 177 PPMs in the JASPAR database was scanned against upstream regulatory regions of *K. marxianus*, and after taking the distance-dependent background frequencies into account, putative interactions with minimum likelihood *Lm* of 10 were retained (see Methods).

Figure 1. Graphical overview of 177 motifs in the JASPAR database of PPMs for fungi. Nucleotide preference scores (probabilities) in PPMs are indicated as a dark colour, with black as 1 and white as zero. Rows in each PPM represent nucleotide scores for A, C, G and T, in that order from above. Rendering was performed in *Reactomica*.

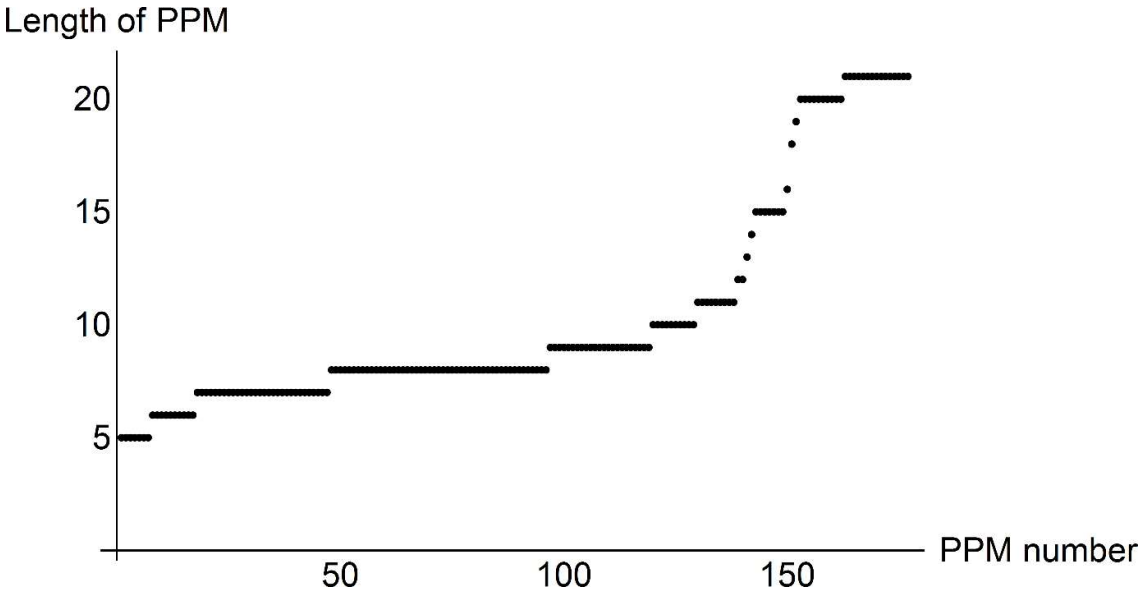


Figure 2. Distribution of motif lengths in the fungal JASPAR database.

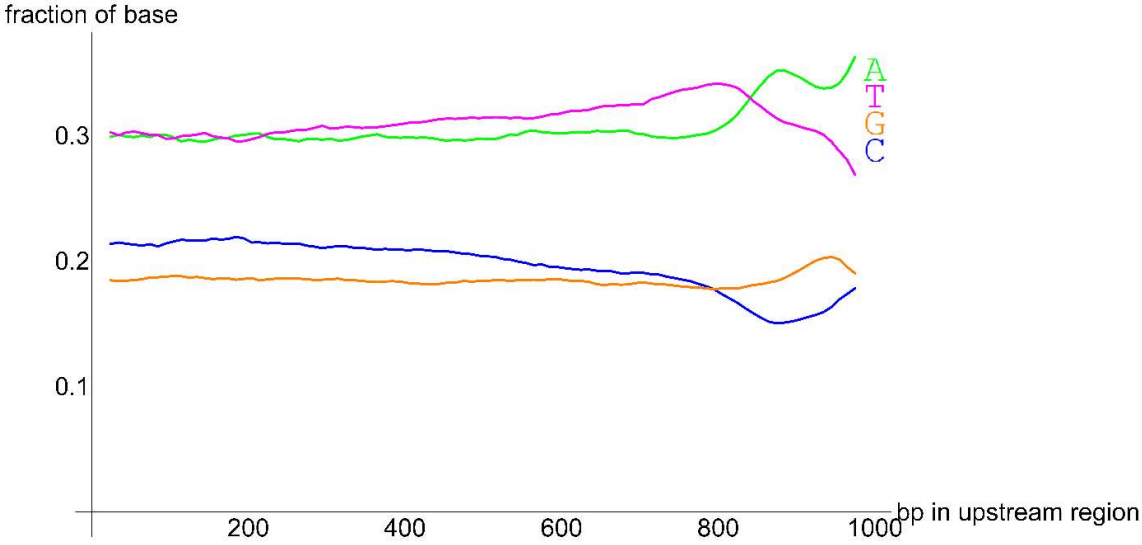


Figure 3. The background frequency of each base as a function of distance from the TLSS. In the figure, the position of the TLSS is taken as 1 000 bp, therefore 1 000 bp from the TLSS is indicative of position number 1.

Figure 4 shows the length distribution of intergenic regions that are upstream to genes. In the draft genome, 4 786 genes could be searched for motifs. A total of 1 416 762 motif matches were found with a likelihood ratio *Lm* at or above 10, some with extremely high likelihood ratios, as high as 5.34×10^{10} (Figure 5), indicating near-perfect matches to long motifs. The majority of motifs, however, had *Lm* values below 10 000. As many of these motifs occurred multiple times in the same regulatory region of a particular gene, the number of interactions based on *Lm* was 489 380.

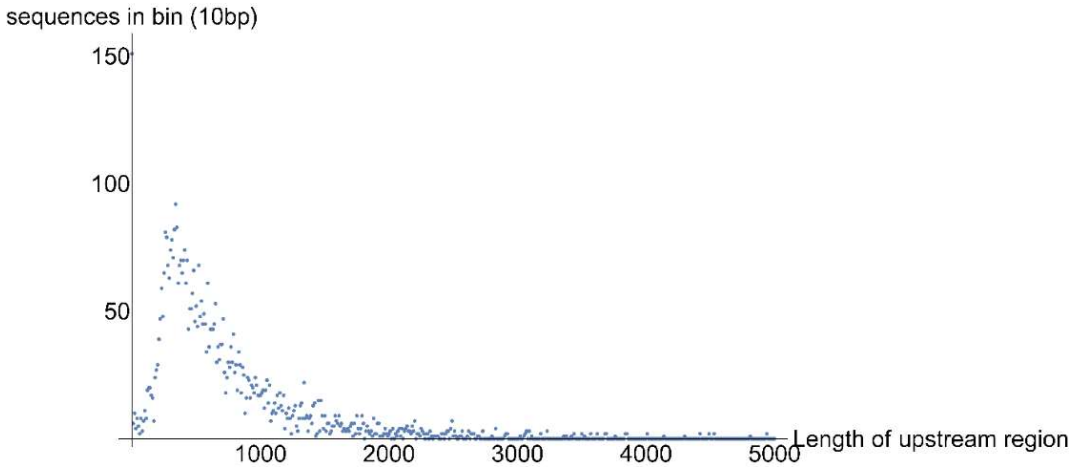


Figure 4. Length distribution of upstream regions in the *K. marxianus* UFS-2791 *de novo* draft genome assembly from Chapter 2.

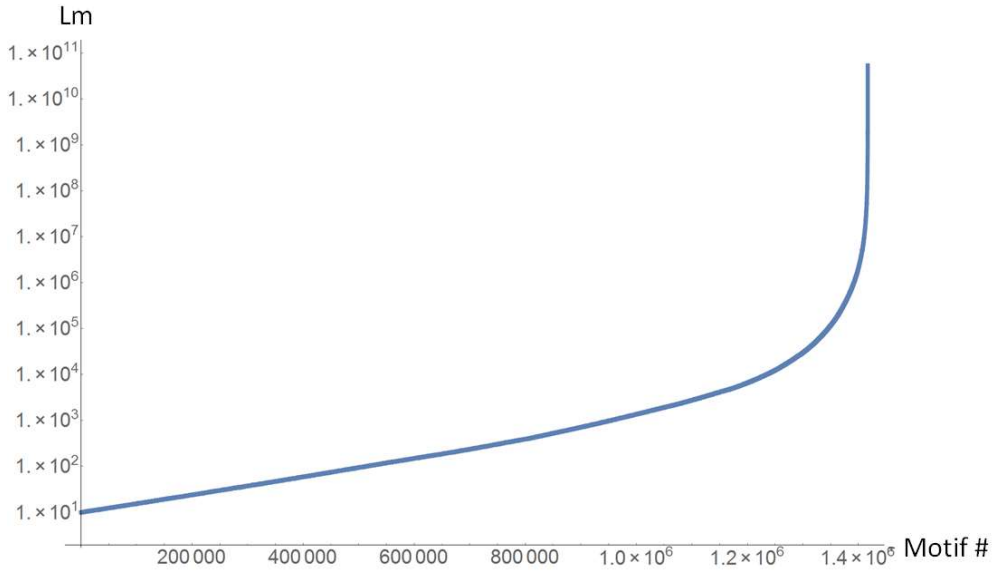


Figure 5. Distribution of the motif likelihood *Lm*.

The highest scoring motifs are likely to originate from the longer PPMs since these can potentially accumulate a higher likelihood score. While any of these motifs may be considered a sufficiently good motif match, especially the high-scoring ones, the vast majority would be false positives. Since there are approximately 33 838 interactions in the SGD, the number of 489 380 interactions assigned by *Lm* as above 10 was approximately 14.5-fold more than expected. The next goal was to combine *Lm* values with additional sources of evidence to eliminate the majority of these and to finally elucidate the true, functional DNA binding sites for TFs.

Figure 6 shows the distribution of *Lc* based on the conservation scores in the motif regions only, which has no normalisation with the conservation scores in the flanking regions, considered as background. It is evident that the majority of motifs were found in regions where three non-reference genomes aligned with the reference genome UFS-Y2791. A significant proportion of these motifs had identical nucleotides in in all four genomes, as evident from a uniform stretch with the conservation score of exactly 3. A small fraction only is in regions where more than three genomes align with the reference, of which only very few approach complete identity in the motif region. Also, a significant fraction of the motifs were in regions in which there was no alignment with any other genome, as indicated by a value of 0.

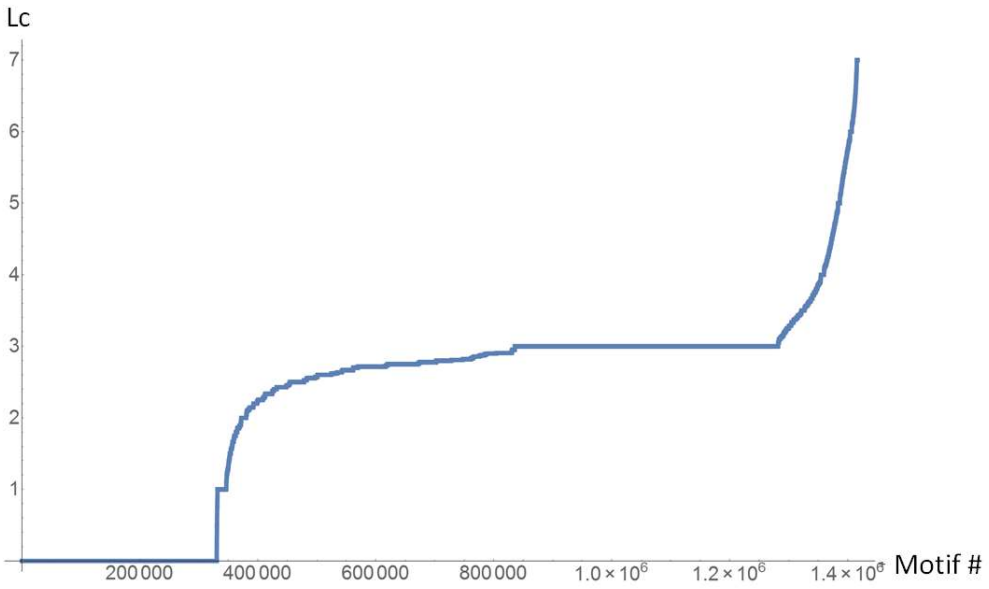


Figure 6. Distribution of conservation likelihood values *Lc*, taken as conservation scores corresponding the motif regions only, without normalisation with the background conservation.

The distribution of the *Lc* values based on the ratio of conservation scores within and around the motif is shown in Figure 7. It can be seen that a significant number of motifs proved to have a score of zero, due to the UFS-Y2791 genome not aligning with any other genome in the region that corresponded to the same motifs as noted above. Secondly, the majority of *Lc* values were very close to one, which would result in no change in the final likelihood. Thus, the background normalisation method automatically corrected for the number of alignments to the reference. Finally, a small fraction of motifs had likelihood ratios of up to 3.5, thus this classifier could increase the likelihood rank ratio *L* by about 3.5. Finally, a method was used in which the minimum value for *Lc* was set at 0.5 to avoid overly strict scoring by the conservation criterion, mostly due to inability of the multiple genome aligner to align divergent genomes. Figure 8 shows the distribution of *Lc* values from the latter

method. Figure 9 demonstrates the effect that each method of calculating *Lc* had on the final likelihood, *L*, calculated as the product of *Lm* and *Lc*. The distributions of final likelihoods, *L,* largely reflected that of *Lm,* but the first *Lc* method would eliminate approximately 25 % of motifs.

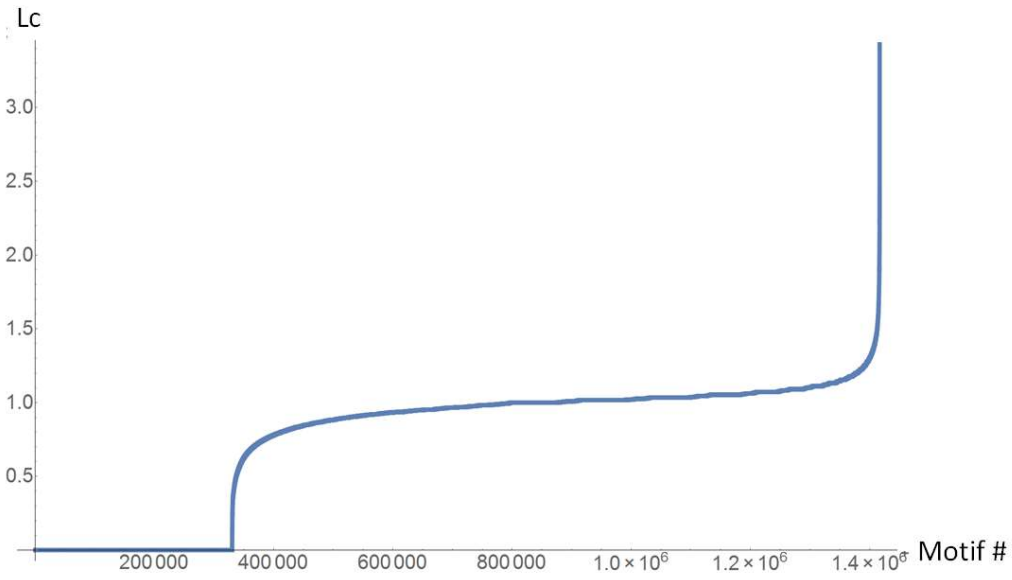


Figure 7. Distribution of conservation likelihood values *Lc,* calculated from the ratio of conservation scores inside motif regions divided by those from the flanking regions.

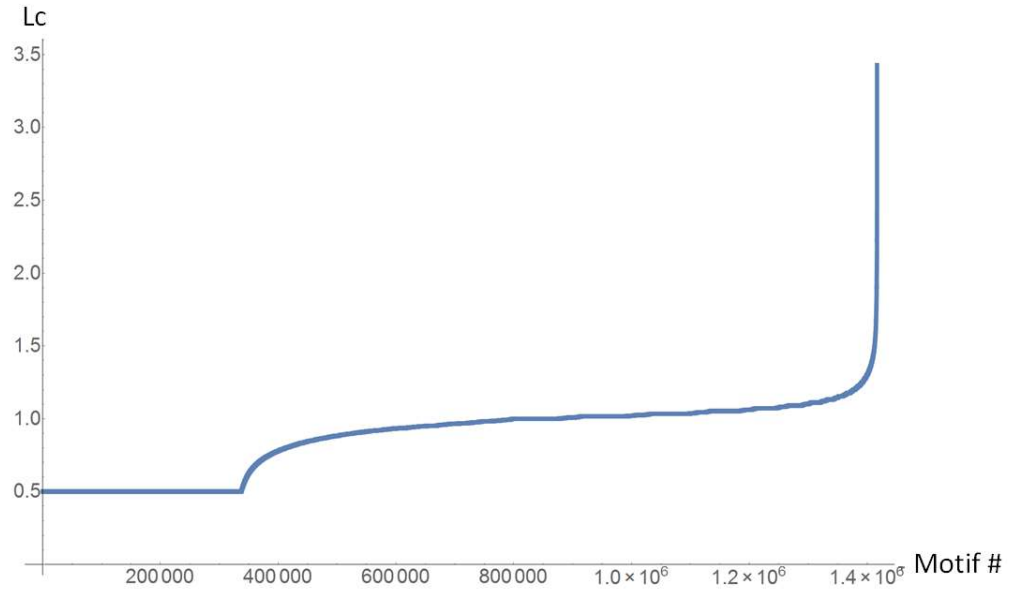


Figure 8. Distribution of conservation likelihood values *Lc,* calculated from the ratio of conservation scores inside motif regions, divided by those from the flanking regions, as well as a minimum value for *Lc* of 0.5.

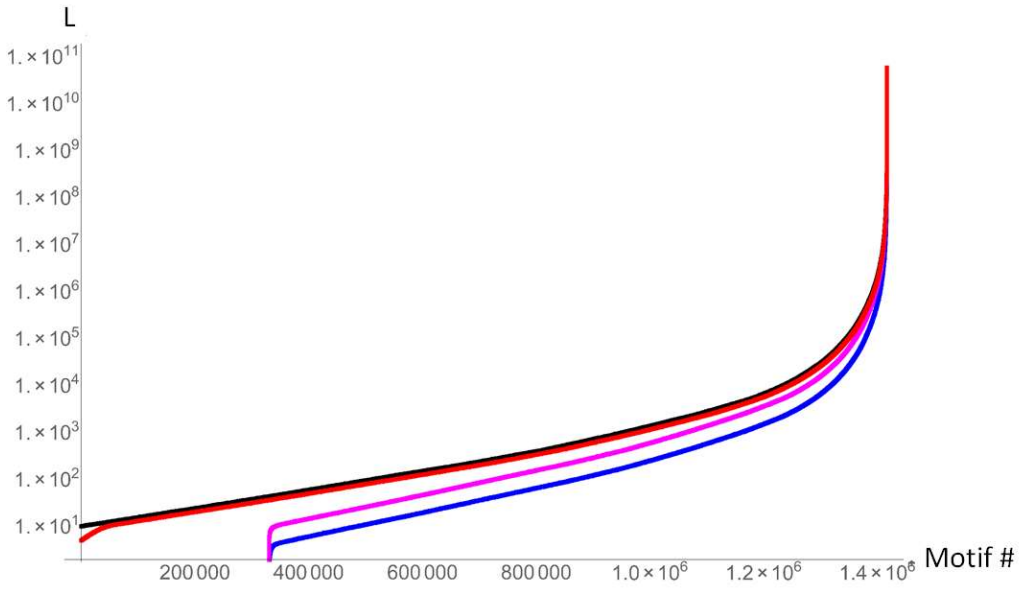


Figure 9. Distribution of final scores calculated as *Lm*×*Lc*. Black, *Lm* only. Blue, *Lm*×*Lc* with *Lc* calculated as motif conservation and no normalisation. Magenta, *Lm*×*Lc* with *Lc* as motif conservation normalized by conservation in the flanking regions. Red: *Lm*×*Lc* with *Lc* as motif conservation normalized by background, where the minimum value for *Lc* was 0.5.

The likelihood ratio profile for common GO terms, *Lg*, based on the Bayesian classifier and a Gold Standard dataset (see Materials and Methods), is shown in Figure 10. The raw data and calculations are shown in Table 1. Since the function was decreasing at the largest number of GO terms in common, possibly due to the small number of occurrences, the cumulative function was used instead as the measure for *Lg* (Figure 11).

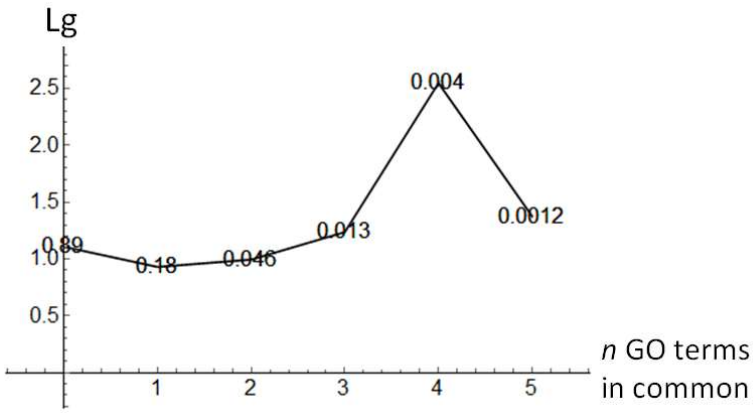


Figure 10. Likelihood ratio *Lg* as a function of *n* GO terms in common between a TF and a target gene. Numbers indicate the fraction of interactions in the training set that had this likelihood ratio.

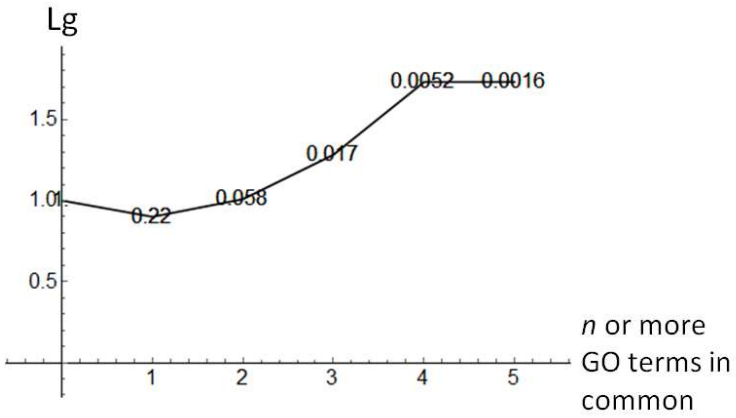


Figure 11. Likelihood ratio *Lg* as a function of at least *n* GO terms in common between a TF and a target gene. Numbers indicate the fraction of interactions in the training set that had this likelihood ratio.

Table 1. Calculation of *Lg* from GS positive and negative datasets.

n GO in common	n pairs GS pos	n in common	n pairs GS neg	GS pos	GS neg	n or more in GS pos	n or more in GS neg	sensitivity	FP rate	n GO in common for Lg	Lg
0	23637	0	21349	30183	28127	30183	28127	1.000	1.000	1	1
1	4798	1	5163	30183	28127	6546	6778	0.217	0.241	2	0.900
2	1236	2	1243	30183	28127	1748	1615	0.058	0.057	3	1.009
3	356	3	288	30183	28127	512	372	0.017	0.013	4	1.283
4	107	4	42	30183	28127	156	84	0.005	0.003	5	1.731
5	33	5	24	30183	28127	49	42	0.002	0.001	6	1.731 (1.087)
6	11	6	11							>6	1.731
7	1	9	7								
8	1										
9	1										
10	1										
26	1										

From Figure 12 it is evident that the vast majority of motif-target pairs in *K. marxianus* had zero or one GO term in common between the TF and target. The distribution of the likelihood rank ratio *L* in Figure 13 appeared identical to the original. For a small fraction, however, the likelihoods could be increased by up to 1.28 or 1.73-fold. Note that even though the effect on the overall distribution was not obvious, the order of likelihoods was altered, which would result in a different gene regulatory network, since only the best interactions are to be retained as interactions.

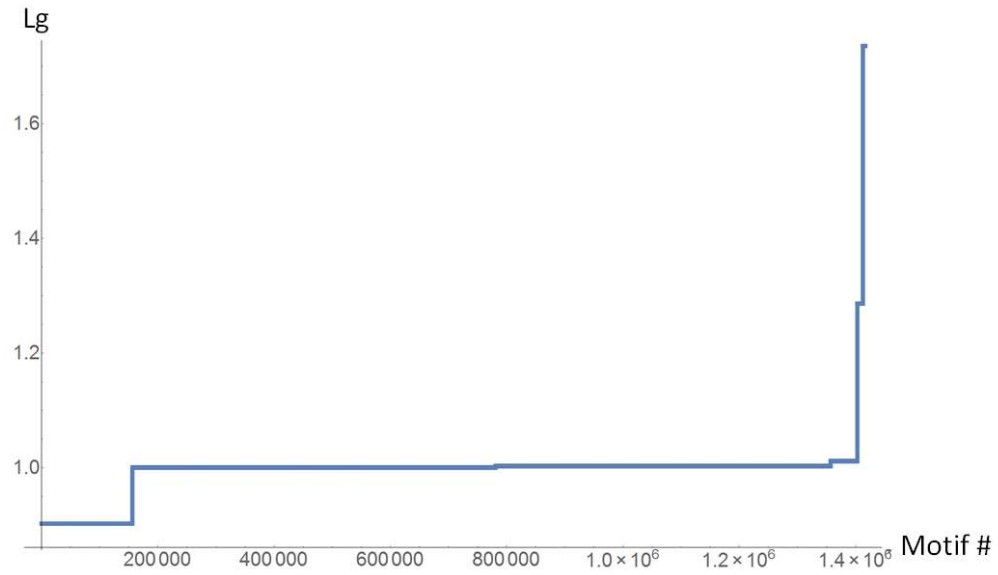


Figure 12. Likelihood ratios *Lg* for all motif-target pairs in *K. marxianus*.

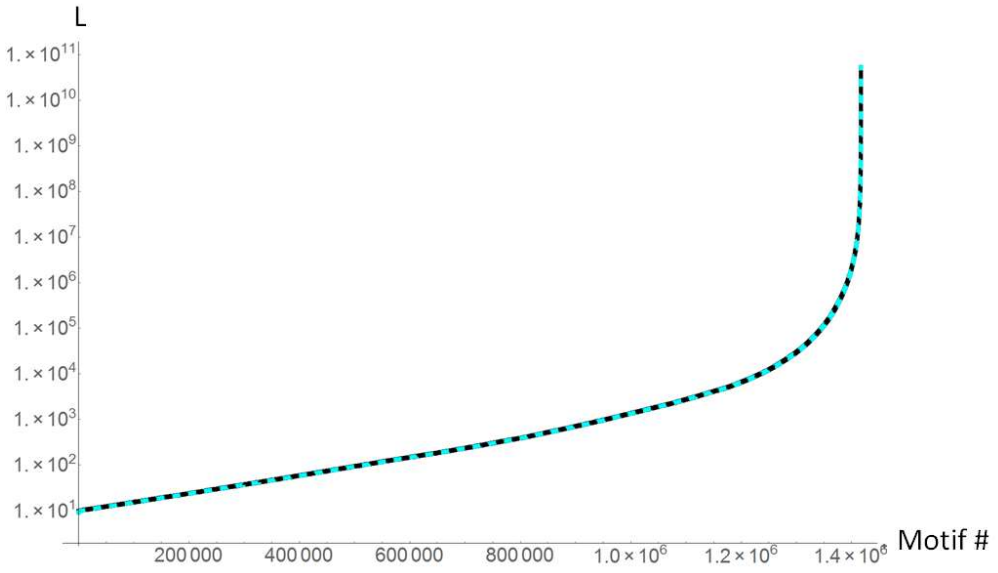


Figure 13. Effect of the likelihood ratio *Lg*. Black, *Lm*; cyan, *Lg*. The distribution was not altered, while for very few interactions, an improvement might have been achieved.

The likelihood based on a common interaction, *Li*, was assigned based on a custom scoring system. Out of the 13 799 interactions mapping between *K. marxianus* and *S. cerevisiae* based on homology, there was evidence for 4 706 interactions based on putative DNA binding sites with *Lm* values above 10, corresponding to 14 052 motifs. The number of motifs was substantially more than the number of interactions, since for many of these, multiple occurrences of the same motif were found in a given regulatory region. Figure 14 shows the distribution of the number of the same motif found for each

possible interaction from the complete list of motifs with *Lm* values above 10. Figure 15 shows that the likelihood based on common interactions between the two species, *Li*, was a very restrictive classifier, with only a small fraction of the motifs receiving improved scores. The data are shown in Table 2.

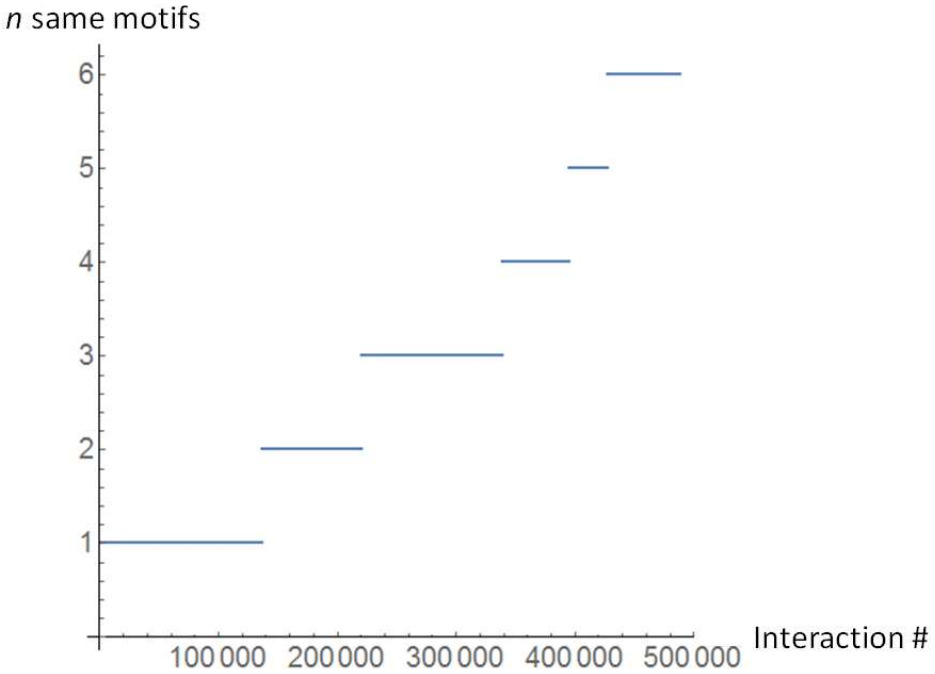


Figure 14. Distribution of the number of the same motifs found for a given TF-target interaction. The majority of interactions had multiple motif matches from the same motif.

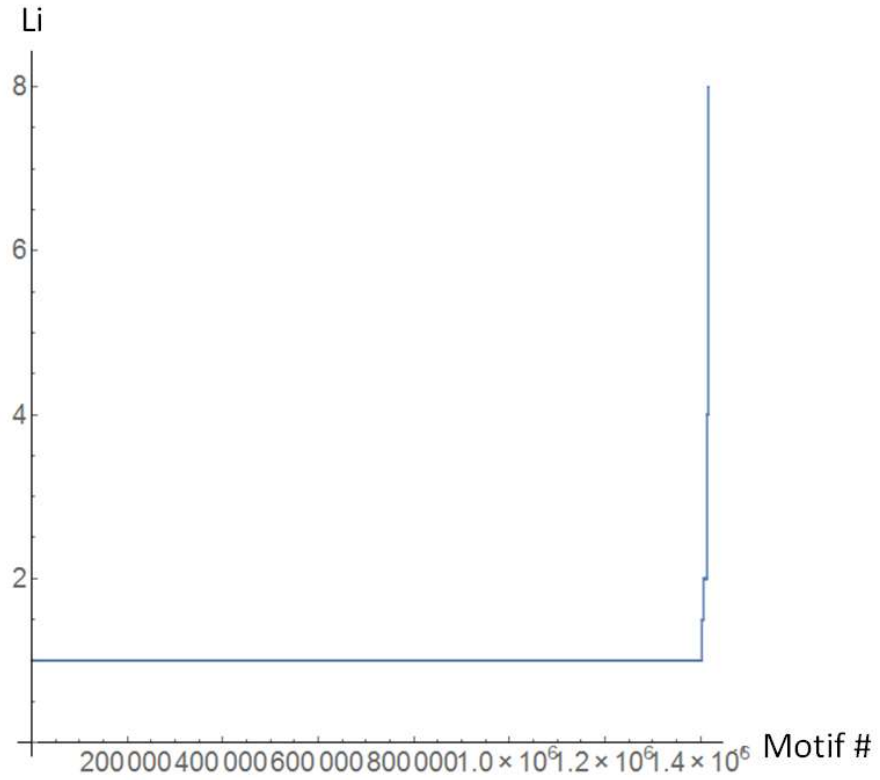


Figure 15. Likelihood based on a common interaction between *K. marxianus* and *S. cerevisiae*.

Table 2. Likelihood based on a common interaction between the two species *K. marxianus* and *S. cerevisiae*.

Li	Number of motifs
1	1 402 710
1.5	3 727
2	7 433
4	2 360
8	532

Methods for optimising the regulatory network

The true number of gene regulatory interactions is not known for any species. SGD contains 32 311 interactions for *S. cerevisiae*. Assuming approximately 6 000 genes for *S. cerevisiae*, this is approximately 5.4 interactions for every target gene. *K. marxianus* UFS-Y2791 has approximately 4 953 protein-encoding genes. At an assumed 5.4 TF-target interactions, 26 673 interactions would be present. Thus, the 26 673 putative interactions with the highest final likelihoods would construct an analogous network for *K. marxianus*. Using lower numbers should result in a network of higher confidence, however. Accordingly, a number of gene regulatory networks were constructed for each likelihood criterion assessed below, where the number of best scoring motifs-target interactions was

varied from 1 000 to 28 000. Note that these resulted in fewer TF-target gene interactions, since only one TF-target gene interaction was included among all motif-target interactions belonging to a target gene.

It was found that the likelihood ratio Lm which captures the motif strength, dominated the scoring system when calculating the final likelihood rank ratio, *L*, as *Lc*×*Lg*×*Li*. Several methods were subsequently tested in which the likelihoods *Lm*, *Lc*, *Lg* and *Li* were used in several combinations to find an improved balance between the parameters in terms of their contribution in shaping the distribution of rank ratio *L*. These are described in detail in Appendix 3. The best network among all methods was obtained by allowing 9 000 motifs, using the $Log_{10}(Lm)$×*Lc*×*Lg*×*Li* method, resulting in 5 443 interactions and 136 TFs (Table 3). A slightly higher scoring network could be generated with a by using 16 000 motifs, which would likely be less accurate due to the larger number of false positives.

Notably, Gcn4 and Gcr2 were revealed as the most significantly enriched in the 9000 motif model and this result was relatively robust to the choice of method. Gcn4 (general control protein 4) was significantly enriched with targets differentially regulated, mostly downward. Gcr2 (glycolysis regulator 2) was the second most significantly enriched TF, with a very significant enrichment score of 3.64 and differential targets almost exclusively down-regulated. Gcn4 is a known activator in amino acid and purine biosynthesis in *S. cerevisiae* [Hinnebusch and Fink 1983, Hope and Struhl 1985] and the RNA-seq data on *K. marxianus* [Chapter 3, Schabort et al. 2016] showed that amino acid and purine biosynthesis were down-regulated in the xylose growth medium. Gcr2 is well known as an activator of the expression of glycolytic enzymes [Uemura and Jigami 1992]. Although it binds DNA, it phosphorylates Gcr1 (glycolysis regulator 1), which also interacts with multiple glycolysis genes in *S. cerevisiae*.

Table 3. Enrichment statistics for an optimised gene regulatory network with 9 000 motifs allowed (5 443 interactions), using the likelihood function *Li* = $Log_{10}(Lm)$×*Lc*×*Lg*×*Li*. TFs with scores above 1.67 ($p < 0.05$) are shown.

name1	name2	Length[Targets]	Z(0)	nNo	nDown	nUp	motiflength
9547	GCN4	89	5.25	65	16	8	21
9549	GCR2	13	3.64	8	4	1	7
9645	SWI4	20	3.04	13	2	5	8
9543	GAL4	2	2.74	1	0	1	15
9629	SOK2	26	2.71	19	3	4	11
9669	YGR067C	9	2.67	5	2	2	14
9520	ASH1	6	2.55	3	0	3	10
9638	STP1	17	2.17	13	2	2	8

name1	name2	Length[Targets]	Z(0)	nNo	nDown	nUp	motiflength
9653	TYE7	30	2.13	22	7	1	7
9666	YDR520C	5	1.99	3	1	1	10
9631	SPT2	19	1.97	15	3	1	10
9656	UME6	13	1.91	9	1	3	13
9620	RTG3	103	1.89	83	14	6	20
9556	HAP1	24	1.87	18	3	3	8
9522	BAS1	21	1.85	16	2	3	21
9662	YAP6	46	1.85	35	3	8	20
9551	GLN3	3	1.83	2	1	0	5
9517	ARO80	39	1.79	31	3	5	21
9637	STE12	48	1.77	39	1	8	7
9550	GIS1	2	1.76	1	0	1	9
9667	YER130C	2	1.76	1	0	1	9
9527	CHA4	7	1.76	4	2	1	8
9679	YPR015C	19	1.67	15	0	4	20

Finally, to obtain a general overview of the network, it was rendered in *Reactomica*. Figure 16 shows the TFs interacting with their target genes. The TFs (triangles) were rendered using the *Clarity* colouring scheme for enrichment statistics, and used enrichment scores also for calculating the sizes of the triangles. A very useful aspect of this method is that it simultaneously (*a*) elucidated the most significantly enriched TFs (size), (*b*) indicated whether TFs were mostly associated with up-regulated genes (orange to red), down-regulated gene (blue), or with mixed differential expression (murky), and (*c*) revealed the number of interactions made by the TFs. Those TFs with few interactions are located on the periphery. It is immediately evident that Gcn4 was the most significantly enriched, but was associated with a mixture of up and down-regulated targets, whereas Gcr2 was also enriched but was strongly associated with down-regulated genes. The rendering of target genes was also suppressed by making them very small (if constitutively expressed) and semi-transparent to highlight the transcription factors. Moreover, the renderings contain dynamic content; on clicking them in *Reactomica*, a number of actions could be performed, while on hovering, the name of the protein, its differential expression or enrichment statistics are shown.

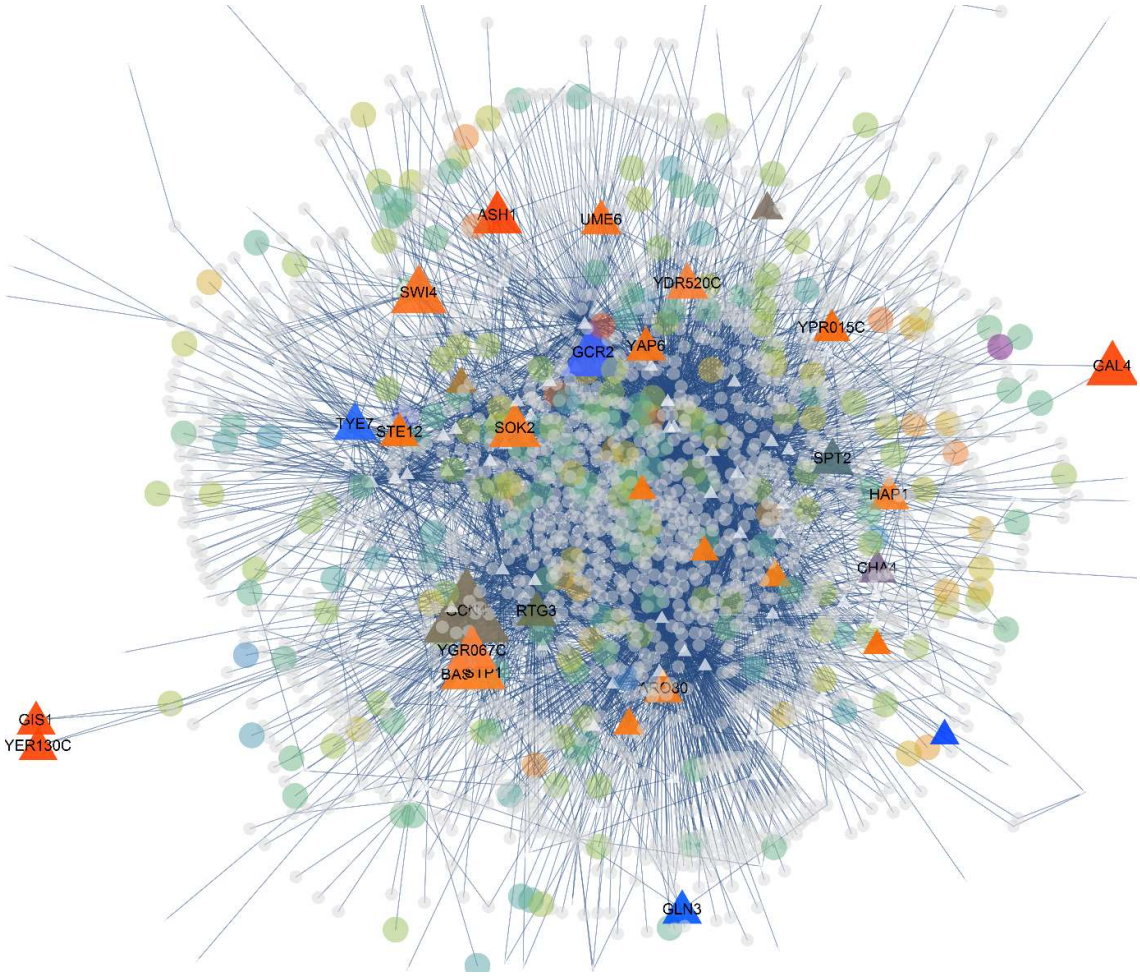


Figure 16. The gene regulatory network consisting of 5 443 interactions and 136 TFs using the *Clarity* colouring scheme for TFs. TFs were rendered as triangles, wherein the size is determined by the enrichment score. Warm colours (orange to red) indicated that a TF was mostly associated with up-regulated genes, cold colours (blue) indicated down-regulation of targets, while murky colours indicated mixed differential expression. Names were only included for TFs that were enriched with a score above 1.67 ($p < 0.05$).

Notably, when rendering the same network using the *Differential expression* colouring scheme for TFs (Figure 17), an interesting pattern was noticed. Most of the significantly enriched TFs, including Gcn4, were not differentially expressed themselves. Gcr2 was an exception, since its gene was down-regulated in the xylose medium.

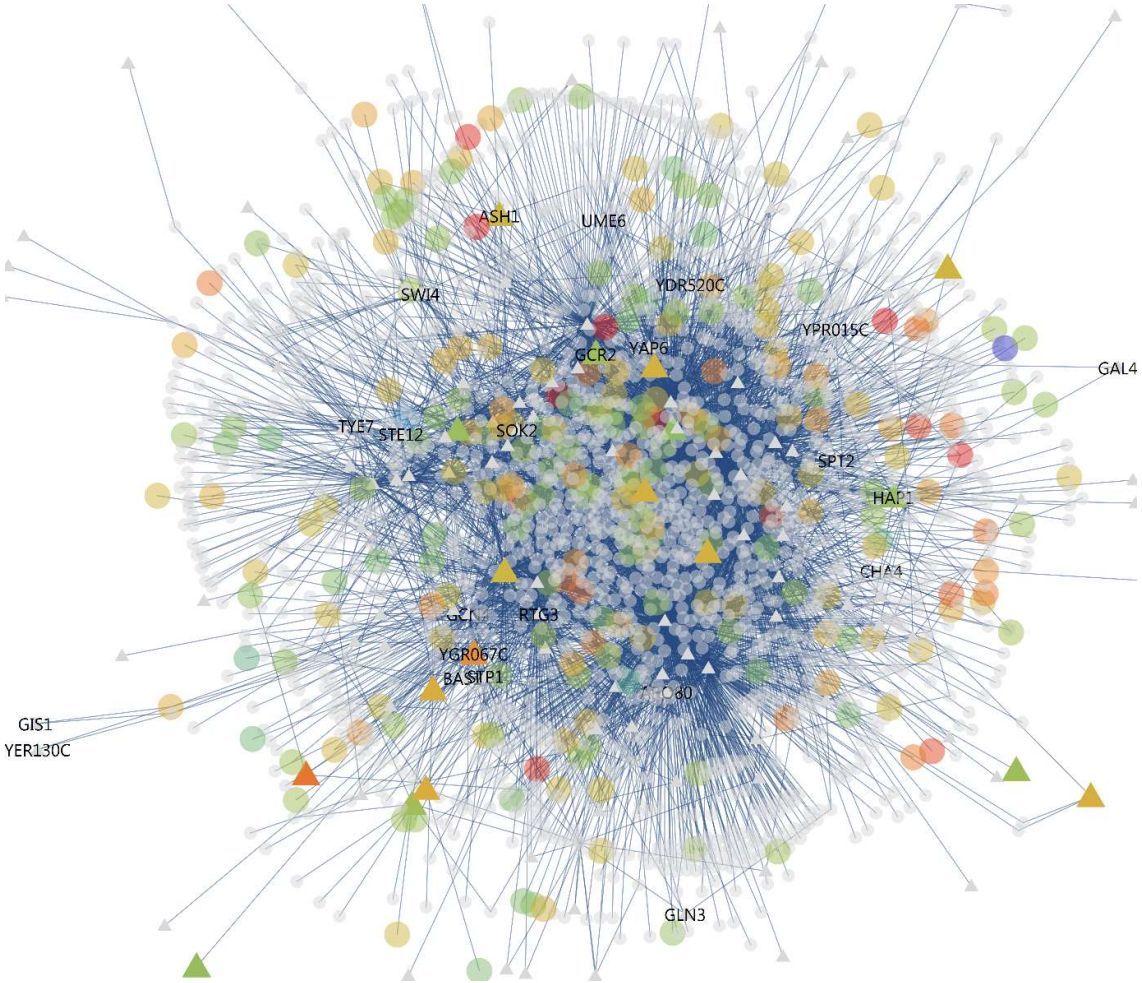


Figure 17. The Gene regulatory network consisting of 5 443 interactions and 136 TFs using the *Differential Expression* colouring scheme for TFs.

The subnetwork of Gcn4 was extracted to reveal its targets and their differential expression (Figure 18). It was clear that many of these were moderately down-regulated (green) with a few more strongly down-regulated (blue), while a few were moderately up-regulated (ochre) or more strongly up-regulated (brown to red).

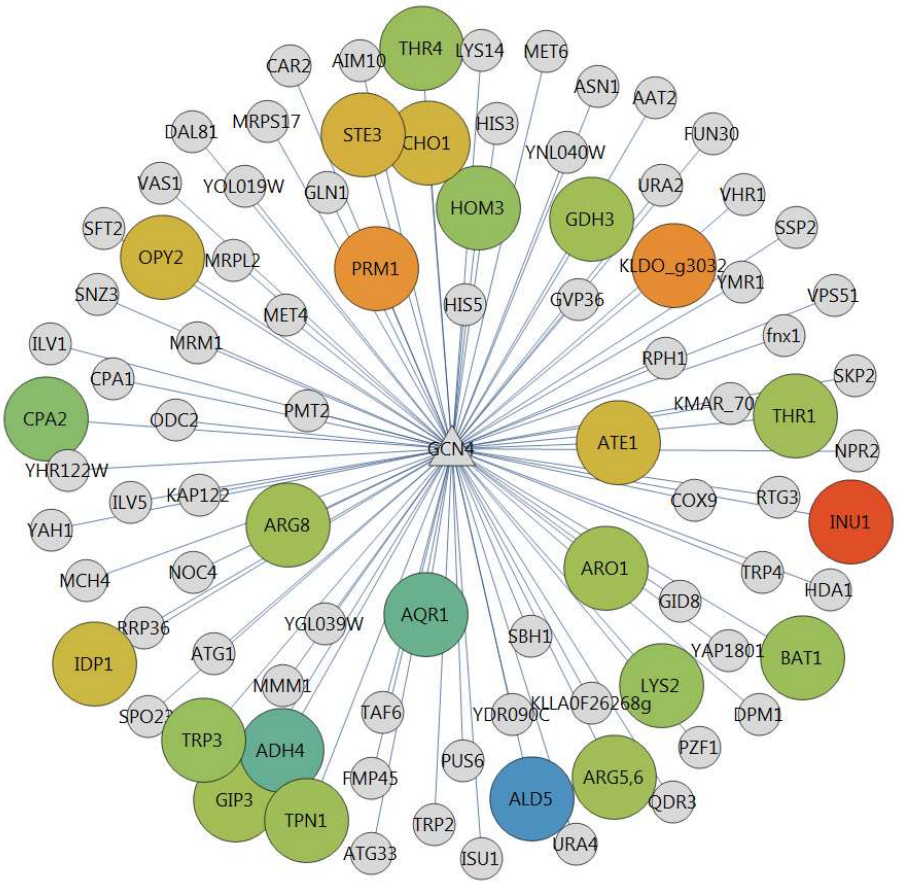


Figure 18. Targets of Gcn4. The target gene set which was significantly enriched and associated mostly with down-regulated targets, but also with a few up-regulated targets.

Conclusions

In this chapter, a likelihood based framework was developed for the construction of a regulatory network based on multiple sources of evidence. The calculations were shown for the motif match likelihood *Lm*, the likelihood for motif conservation among species *Lc*, the likelihood for GO terms in common between TF and target, *Lg*, and the likelihood for conservation of an interaction in the model species, *Li*. The likelihood classifier *Lg* was shown to find utility for a very small fraction of the number of interactions, and was not a very strong classifier as compared to the other classifiers *Lm*, *Lc* and *Li*. Besides, the classifier based on experimental evidence in the model species, *Li*, might be more suitable

as opposed to *Lg* in the case of *K. marxianus*, since the model species *S. cerevisiae* has been very well studied and the two species have many genes in common. However, if the closest model species were only distantly related to the species of interest, *Lg* might prove essential because GO terms apply to all proteins, irrespective of the species compared. Also, it may add some distinguishing power in the case of genes that are not conserved between the two species. Further, since *Lm* was dominant over other likelihoods, a balance had to be struck between these by suppressing the effect of Lm somewhat by using the Log_{10} of *Lm*. The most suitable network size, considering the enrichment statistics and, in particular, the distinguishing power for separating between enriched and non-enriched TFs, was found to be 9 000 motifs, resulting in 5 443 interactions and 136 TFs.

Many adaptations and improvements can still be made to the algorithms developed. For instance, the best length of the flanking regions for calculating the background conservation scores could be determined. Most important, it was evident that significantly enriched TFs could even be elucidated using only the supporting evidence *Lg*, *Lc*, and *Li*, eliminating the motif score *Lm* in the calculation of the final likelihood. This poses the question what the contribution of each of these forms of evidence might have been, and whether a single likelihood function should be applied to all TFs alike. Also, Adr1 and Mig1 were absent from the enriched list, even though the enumerative method of heptamer frequency comparisons suggested that Adr1, in particular, was differentially active [Chapter 4]. The motifs which the Adr1 and Mig1 bind are very short and degenerate, and would thus be suppressed by inclusion of the motif match likelihood *Lm*. However, even with only using additional sources of information (*Lc*×*Lg*×*Li*), these were not enriched. Assuming that the Adr1 and Mig1 targets were conserved between *K. marxianus* and *S. cerevisiae*, it may also be the case that the true targets have not been captured in the SGD; thus the interaction conservation likelihood *Li* would not have led to the inclusion of the target sets of these TFs. Conversely, these target sets may be different between the two species, the result of transcriptional rewiring during evolution. These questions are addressed in the next chapter. The method presented here shows flexibility in combining multiple sources of evidence, with many more possibilities. Moreover, a number of TFs were shown to be enriched, including Gcn4, Gcr2, Swi4, Sok2, YGR067C and Ash1, in that order. The regulation of the targets of Gcn4 and Gcr2 seems to be consistent with their known functions in activating amino acid biosynthesis and glycolysis, respectively. Notably, the genes encoding for strongly differentially active TFs such as Gcn4 were constitutively expressed, showing that these major regulators would be missed in the less mechanistic, reverse engineering approach which requires differential expression of not only the target genes, but also of the TF genes to infer interactions by correlation. Even though these results should be interpreted as preliminary, since the method is still not complete and the analysis based on

a draft genome, it is noteworthy that the first genome-scale gene regulatory network has been constructed for the species from a fragmented draft genome.

References

Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. Nat Genet. 2005;37: 382-390.

Bulyk 2006. DNA microarray technologies for measuring protein-DNA interactions. Curr Opin Biotechnol. 2006;17(4): 422-30.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED. *Saccharomyces* Genome Database: the genomics resource of budding yeast. Nucl Acids Res. 2011;1: 1–6.

Collins SR, Kemmeren P, Zhao X, Greenblatt JF, Spencer F, Holstege FCP, Weissman JS, Krogan NJ. Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. Mol Cell Proteomics. 2007;6: 439-450.

de Sousa AR, Penalva OP, Marcotte EM, Vogel C. Global signatures of protein and mRNA expression levels. Mol. Biosyst. 2009;5(12): 1512–1526. doi:10.1039/b908315d.

Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PloS ONE. 2010;5(6): e11147. doi:10.1371/journal.pone.0011147.

Duda RO, Hart PE, Stork DG. Pattern classification. New York: J. Wiley & Sons; 2001.

Gitter A, Siegfried Z, Klutstein M, Fornes O, Oliva B, Simon I, Bar-Joseph Z. Backup in gene regulatory networks explains differences between binding and knockout results. Mol Syst Biol. 2009;5: 276.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne J, Reynolds DB, Yoo J, Jennings EJ, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004;431(7004): 99–104.

Haury A, Mordelet F, Vera-Licona P, Vet J. TIGRESS: Trustful Inference of gene regulation using stability selection. BMC Syst. Biol. 2012;6(145).

Hinnebusch AG, Fink GR. Positive regulation in the general control of *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA. 1983;80: 5374-5378.

Hope IA, Struhl K. GCN4 protein, synthesized in vitro, binds HIS3 regulatory sequences: implications for general control of amino acid biosynthetic genes in yeast. Cell. 1985;43: 177-188.

Hu Z, Killion PJ, Lyer VR. Genetic reconstruction of a functional transcriptional regulatory network. Nat Genet. 2007;39(5): 683-687.

Jansen R, Yu H, Greenbaum D, Kluger Y, Krigan N, Chung S. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science. 2003;302(5644): 449-453.

Marbach D, Costello JC, Küffner R, Vega N, Prill RJ, Camacho DM, Allison KR, The DREAM5 Consortium, Kellis M, Collins JJ, Stolovitzky G. Wisdom of crowds for robust gene network inference. Nat Methods. 2013;9(8): 796-804. doi:10.1038/nmeth.2016.

Mathelier A, Zhao X, Zhang AW, Percy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H, Lim J, Shyr C, Tan G, Zhou M, Lenhard B, Sandelin A, Wasserman WW. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 2014:42(D1): D142-D147.

Nishida K, Frith MC, Nakai K. Pseudocounts for transcription factor binding sites. Nucleic Acids Res. 2008;37(3): 939-944.

Patil KR, Nielsen J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc Natl Acad Sci USA. 2005; 102(8): 2685–2689. PMID: 15710883.

Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan G. Causal protein-signaling networks derived from multiparameter single-cell data. Science. 2005;308: 523-529.

Schabort DWP, Letebele PK, Steyn L, Kilian SG, du Preez JC. Differential RNA-seq, multi-network analysis and metabolic regulation analysis of *Kluyveromyces marxianus* reveals a compartmentalised response to xylose. PLoS ONE. 2016;11(6): e0156242. doi:10.1371/journal.pone.0156242.

Sonderegger M, Jeppsson M, Hahn-Hägerdal B, Sauer U. Molecular basis for anaerobic growth of *Saccharomyces cerevisiae* on xylose, investigated by global gene expression and metabolic flux analysis. Appl Environ Microbiol. 2004;70(4): 2307–2317.

Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9): 1105-1111.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimental H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2013;7(3): 562–578. doi: 10.1038/nprot.2012.016.

Uemura H, Jigami Y. Role of GCR2 in transcriptional activation of yeast glycolytic genes. Mol Cell Biol. 1992;12(9): 3834-42.

Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput Biol. 2011;7(10): e1002240.

Wang L, Hou L, Qian M, Deng M. Integrating phosphorylation network with transcriptional network reveals novel functional relationships. PLoS ONE. 2012;7(3): e33160.

Wang K, Saito M, Basikiriska B, Alvarez MJ, Lim WK, Rajbhandari P. Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. Nat Biotechnol. 2009;27: 829-837.

Chapter 6

A gene regulatory network based on the complete genome of *Kluyveromyces marxianus*

Abstract

In Chapter 5, an integrative framework was described for constructing a genome-scale gene regulatory network based on multiple sources of information, using mechanistic evidence as far as possible, in a likelihood framework. The resulting network was based on a draft genome constructed *de novo* for the *Kluyveromyces marxianus* strain UFS-Y2791. Enrichment statistics revealed the transcription factor Gcn4, followed by Gcr2, as the main regulators. However, the short zinc finger motifs of Adr1 and Mig1, which have been elucidated in Chapter 4 using the enumerative method of heptamer frequencies, did not appear as enriched in the genome-scale network approach using likelihoods. Since the motifs recognised by Adr1 and Mig1 are short and degenerate, their inclusion in the network might be suppressed by the low scores they can generate in motif matching. In this chapter, various likelihood functions were tested and an improved method was developed. By comparing the results from various likelihood functions, the shortcomings in our knowledge could be highlighted. A complete genome was used for read-mapping of RNA-seq data and for constructing an improved genome-wide gene regulatory network for *K. marxianus*. Notably, the most significantly enriched regulators were not differentially expressed, but rather their differential activity seems to be regulated by post-translational modifications. This also demonstrates the advantage of using the mechanistic basis for gene regulatory networks used here over the reverse engineering approach.

Introduction

RNA-seq transcriptomics of the differential response of *Kluyveromyces marxianus* to glucose or xylose revealed a general pattern of down-regulation of genes involved in *de novo* amino acid synthesis, as well as up-regulation of many genes involved in the utilisation of alternative carbon sources [Chapter 3, Schabort et al. 2016]. In Chapter 4 the enumerative method of heptamer frequency comparison was used to reveal the transcription factors (TF) Adr1 and possibly Mig1 as the most likely regulators in the up-regulated response, which corresponded with data from the literature on *Saccharomyces cerevisiae* [Young et al. 2003]. In Chapter 5, a likelihood framework was developed to construct a genome-scale gene regulatory network combining the scores from motif scans, motif conservation among sister species, common Gene Ontology (GO) terms between TF and target, and experimental evidence for the same interaction in the *Saccharomyces* Genome Database (SGD). Using enrichment statistics, also termed reporter TFs [Patil and Nielsen 2005], Gcn4 and Gcr2 were revealed as the most important regulators in the differential response to glucose or xylose. This finding corresponded to the down-regulation of amino acid *de novo* synthesis pathways and glycolysis in which these two TFs were activators. However, there was no evidence in the enrichment statistics to support the involvement of Adr1 or Mig1. The likelihood framework effectively makes motifs compete for inclusion into the network based on evidence supporting their presence, and the longer, more selective motifs may outcompete the short motifs since they can potentially accumulate a larger score. This may explain the lack of finding Adr1 and Mig1 enrichment in the genome-scale reconstruction, since these have short degenerate motifs. The use of the Log_{10} of the motif score supported a better balance between the contribution of the motif likelihood, *Lm*, and other criteria. However, it has to be considered that the role of the motif scores for the short, degenerate motifs may be small compared to other sources of evidence and also that other methods of calculating the final likelihoods (termed functions) need to be explored.

In biological databases such as SGD, some effectors have been associated with hundreds of targets. A large fraction of these interactions were derived from knockout expression profiling studies. Possibly a significant fraction of these interactions might be the result of secondary effects, where the knockout of the regulator affected the expression of another, major regulator. To avoid the secondary effects, the type of evidence could be used to emphasise those interactions supported by direct physical evidence in addition to expression profiling in the likelihood framework. In this chapter, the first aim was to develop a function in which the type of experimental evidence in the model species is weighted. A simple scoring system is presented which is useful in the likelihood framework.

The second aim was to explore the effect that different functions for assigning the final likelihood for an interaction would have on the deductions made from TF enrichment statistics. For long, restrictive DNA binding motifs, the motif score alone might provide a better functions for assigning the targets as opposed to combining it with experimental data from the model species, which might both be incomplete, and possibly irrelevant due to evolutionary differences. The opposite might be true for short, degenerate motifs where the binding sites may be hardly detectable and would need strong support from evidence in the model species. To explore these scenarios, seven additional functions for calculating the likelihoods were designed. One of these was to construct the genome-scale network using only the gene regulatory targets from SGD, mapped to the homologs in *K. marxianus*, regardless of the presence of a DNA binding site. Moreover, by comparing the results from enrichment statistics between this and the other methods (the enumerative method and several likelihood functions), insight can be gained into the shortcomings in our knowledge of the DNA binding motifs in the species of interest or even in the model species, as well as possible transcriptional rewiring between the species.

It was also considered important to calculate enrichment using an additional criterion together with the Z-score method. The hypergeometric distribution can be used for a variety of purposes and was used in a previous chapter to calculate the probability of finding a certain number or more of differentially expressed targets in a gene set by chance, given the background (Chapter 4). A useful feature of the hypergeometric distribution is that it is simpler to calculate compared to the Z-score method, and hence facilitates calculating a probability value for either up or down regulation of a gene set, thus revealing also the direction of regulation of a gene set.

As was described in Addendum 1, it was found that transcripts from RNA-seq data could be read-mapped to the complete genome of DMKU3-1042. The final aim in this chapter was to apply the improved likelihood framework in the context of this complete genome and to elucidate the differentially active TFs that formed the mechanistic basis behind the transcriptional response.

Materials and Methods

RNA-seq data

RNA-seq data were generated in previous work for the strain UFS-Y2791 [Chapter 3, Schabort et al. 2016]. This data was read-mapped to the recently published complete genome of a different stain,

DMKU3-1042 from Lertwattanassakul et al. [2015] using TopHat [Trapnell et al. 2009]. Differential expression was calculated using CuffDiff [Trapnell et al. 2013].

Motif scans

Motif scans were performed for all regulatory regions in *K. marxianus* DMKU3-1042, defined as the 1000 bp upstream of a gene, which were not allowed to overlap with a neighbouring gene (See Chapter 5). Motifs were assigned in both directions. A motif likelihood *Lm* was calculated for a motif match by calculating the ratio of the probability of the sub-sequence matching the model, divided by the probability of a sub-sequence matching the background model of nucleotide frequencies in the upstream regions. Each motif likelihood score *Lm* was calculated as the product of independent occurrences as below, where *m* is the probability from the positional probability matrix (PPM) matching to the character at position *i*, and *b* is the background frequency of the character at position *i*.

$$L\mathrm{m} = \frac{\prod_i^n m\,[i]}{\prod_i^n b\,[i]}$$

In this likelihood formulation, the background frequency of each nucleotide A, C, G and T is taken into account to give an estimate of the statistical significance of a motif match. The background nucleotide frequencies were calculated in a sliding window of 30 base pairs in all upstream regions at various distances from the translation start site (TLSS). Effectively, the motif score was normalised to the background frequencies at the relevant distance from the TLSS.

Construction of a gene regulatory network

All data processing and visualisation were performed in algorithms developed for *Reactomica* [Chapter 3, Schabort et al. 2016] implemented in the Wolfram language. Various methods of network construction were implemented according to the likelihood framework described in Chapter 5, and similar to previous applications [Jansen et al. 2003, Collins et al. 2007, Chapter 5], but containing an expanded set of functions. The *n* top scoring motifs based on final likelihood, *L*, were allowed to be included as TF-target interactions. Sources of evidence included the motif score, *Lm*, the conservation of a motif among seven *Kluyveromyces* isolates, *Lc*, and the experimental evidence for an interaction in SGD, *Li*. The Gene Ontology score, *Lg* [Chapter 5], was considered inferior to the experimental evidence score *Li*, in this case, and omitted from the procedure.

The multiple genome alignment of seven isolates of *Kluyveromyces* using progressiveMauve [Darling et al. 2010] is described in Addendum 2, and the use of multiple genome alignments to calculate

conservation scores *Lc* for putative motifs is described in Addendum 3. The origin of motif score, *Lm*, is described above. The motif score, *Lm*, was implemented as $Log_{10}(Lm)$ in the calculation of final likelihoods to suppress the dominating effect of *Lm* and allow shorter and less precise (higher entropy) motifs to compete for inclusion in the network. For the experimental evidence score *Li*, transcriptional regulator-target sets were obtained from the YeastMINE interface of SGD of all gene regulatory interactions, where the targets were protein-encoding genes. The genes representing transcriptional regulators and targets in *S. cerevisiae* were mapped to *K. marxianus* genes annotated by Lertwattanassakul et al. [2015] in a two-step process. To improve the richness of the annotation, the unique identifier of each *K. marxianus* protein was used to retrieve the UniProt annotations, and the primary name of the gene was then used to match to those in the *S. cerevisiae* regulatory target sets. A scoring system was designed for incorporating the various types of data supporting an interaction in *S. cerevisiae* to calculate *Li* (described in the Results section).

Eight likelihood functions were compared to incorporate the three sources of evidence in various forms, as shown below.

A) $Log_{10}(Lm)$*Lc*Li	(*Lc* >= 0.5)
B) $Log_{10}(Lm)$+Lc+Li	(*Lc* >= 0.5)
C) *If*[*Li*==0, $Log_{10}(Lm)\times Lc$, *else* $Log_{10}(Lm)\times Lc\times Li$]	(*Lc* >= 0.5)
D) $Log_{10}(Lm)$	
E) $Log_{10}(Lm)$*Lc	(*Lc* >= 0.5)
F) Lc	(*Lc* >= 0.5)
G) Li	
H) SGD	

Function B was to sum the scores instead of their multiplication, which allowed inclusion of interactions even though evidence was not available for the model species. Function C was to apply a qualifier (*if* statement): if there is no experimental evidence for an interaction (*Li* = 0), assume a score of 1 for *Li* and perform multiplication. For all criteria that used the conservation score *Lc*, it was assumed to be 0.5 in cases where *Lc* was below 0.5. This was necessary to avoid exclusion of motifs that occurred in regions that could not be aligned to sister species. The progressiveMauve multiple genome aligner used in this study has been optimised for speed and, accordingly, divergent orthologous regions might be not be aligned. In addition, the functions $Log_{10}(Lm)$, $Log_{10}(Lm)\times Lc$, *Lc* and *Li* were also tested. Each of the functions were used to calculate the final likelihood score of a DNA

binding motif-target interaction, after which the motif table was sorted according to the final likelihood. The top *n* motifs were subsequently taken to construct a network of TF-target interactions. In cases where more than one occurrence of the same motif was found in the same regulatory region, the best scoring motif was retained. In effect, all motifs competed against one another for inclusion into the regulatory network, based on the evidence supporting their inclusion. Note that the scores from the *Li* function differ from what was calculated by considering only experimental evidence from SGD (the SGD method) in that the presence of at least one top-scoring motif (within the top *n* motifs) was required in the *Li* function, whereas with the SGD method an interaction was included, regardless of the presence of a motif in *K. marxianus*. Non-TF regulators were omitted from networks constructed using the seven likelihood functions, as the presence of a predicted DNA binding motif was required (with *Lm* above 100).

Gene set enrichment statistics

To calculate enrichment statistics, two criteria were used. The method of simulated background distribution using randomly picked gene sets (the Z-score method) were used [Patil and Nielsen 2005, Schabort et al. 2016] where the enrichment score was calculated as below.

$$\mathrm{S} = \frac{\mathrm{Z(total, Test)} - \mathrm{Mean(Z, Background)}}{\mathrm{Standard\ deviation(Z, Background)}}$$

The Z-score was calculated from the q-values from the CuffDiff output [Chapter 3, Schabort et al. 2016]. Secondly, the hypergeometric distribution was used to estimate the probability of finding the same number or more of genes up or down-regulated in each gene set in a background with 4 093 protein-encoding genes, with 323 up-regulated and 245 down-regulated. To correct for multiple comparisons, the p-values from the hypergeometric distribution were multiplied by the number of effectors (184) to obtain q-values. For simplicity, the enrichment score was calculated as $-\mathrm{Log}_{10}(q)$, where the p-value was the output from the hypergeometric test for either differential expression, up-regulation, or down-regulation.

The enumerative method of motif discovery and Occam's razor motifs

To complement the enrichment approach to elucidating differentially active TFs, the enumerative method of heptamer frequency comparisons was performed for the down-regulated set of genes, as was done for the up-regulated set [see Chapter 4]. The heptamers were mapped to PPMs and a threshold score of 0.7 was used to limit matches between heptamers and PPMs. The method of constructing a DNA binding motif for a single dominating TF as a simplest explanation approach to motif discovery was described in Chapter 4. A combined motif was constructed by aligning these

heptamers, filling in equal counts of unobserved bases, calculating an artificial counts matrix, and finally calculating a PPM.

Construction of an optimised likelihood network

The most likely target set for each TF was taken as the set of targets with the most significant enrichment score for either up or down-regulation, using the hypergeometric test, among all scoring functions, excluding the one based on SGD data alone. The final network consisted of only the top 38 TFs that were significantly enriched in at least one such test if two criteria were met: at least five targets were required and the q-value had to be lower than 0.005 (enrichment score of 2.301).

Results and Discussion

Emphasising experimental evidence in the model species

Table 1 shows the enrichment statistics using only biological prior information from SGD, without the requirement for a potential DNA binding site in the regulatory region of a target gene (Log_{10}(*Lm*)) or conservation among species (*Lc*). Gcn4 and Gcr2 were the most significantly enriched TFs (low q-value), considering any of the hypergeometric tests (differential, up or down regulation). The gene sets of Pdr3, Aca1 and Yap3 were too small to provide a meaningful test for consideration. Generally, a very good classification between up-regulated and down-regulated gene sets could be revealed for the significantly affected target gene sets.

Table 1. Enrichment statistics using only biological prior information of all effectors of transcription in *S. cerevisiae*, regardless of the presence of a top-scoring motif in *K. marxianus* (method SGD). The results were sorted to the minimum of the q-values for either the differential expression, up regulation, or down regulation of a target gene set, as calculated using the hypergeometric distribution after correction for multiple comparisons. Only the top 30 effectors are shown. FC refers to the fold change observed for the TF gene in RNA-seq data from Schabort et al. [2016]. The full table is shown in Addendum 4.

name1	name2	K	Z(0)	nNo	nDown	nUp	qDiff	qUp	qDown	direction	ID	FC
9597	PDR3	1	1.5	0	1	0	0	11.51	0	NA		
	ACA1	2	3.6	0	0	2	0	0	17.05	NA		
9660	YAP3	2	3.6	0	1	1	0	0.72	0.41	NA	gene1726	1
9547	GCN4	169	**11.0**	113	39	17	**1.4E-13**	3.27	**4.9E-16**	down	gene3451	1
9549	GCR2	40	**6.0**	24	14	2	**5.8E-05**	84.92	**2.3E-08**	down	gene2613	0.4
	TUP1	184	**6.8**	134	30	20	**2.4E-08**	0.92	**5.9E-08**	down	gene1081	1
	SPT3	266	**7.2**	203	37	26	**7.1E-08**	1.46	**7.6E-08**	down	gene1396	1
9656	UME6	295	**8.1**	228	25	42	**1.4E-07**	**1.6E-05**	0.27	up		
	HFI1	160	**8.2**	116	26	18	**1.9E-07**	0.90	**8.7E-07**	down	gene1998	1
	SPT20	184	**7.1**	139	29	16	**7.1E-06**	12.30	**2.7E-07**	down	gene1462	1
	BUR6	252	**6.9**	195	17	40	**3.0E-06**	**1.3E-06**	9.63	up	gene3898	1
9578	MET32	205	**6.9**	158	23	24	**3.2E-05**	0.16	**5.6E-03**	down	gene4085	1
	SUA7	401	**3.7**	333	17	51	0.01	**3.2E-05**	116.99	up	gene3433	1
	SPT10	423	**6.2**	344	31	48	**6.1E-05**	**2.1E-03**	0.93	up	gene3729	1
9513	AFT1	68	**5.7**	46	11	11	**7.2E-05**	0.16	**1.2E-02**	down	gene643	1

	CDC73	29	**5.9**	16	8	5	**9.7E-05**	1.47	**8.4E-04**	down	gene635	1
	SIN4	200	**4.1**	164	26	10	0.17	131.89	**1.1E-04**	down	gene3684	1
9541	FKH2	106	**6.8**	77	16	13	**1.1E-04**	1.08	**1.5E-03**	down	gene4226	1
9620	RTG3	38	**4.1**	25	10	3	**4.7E-03**	38.97	**1.6E-04**	down	gene4377	1
9522	BAS1	54	**4.8**	39	12	3	0.03	80.97	**1.6E-04**	down	gene1050	1
9622	SFP1	1306	2.8	1148	94	64	12.2	182.14	**2.4E-04**	down	gene2869	1
	SIN3	73	**5.8**	51	12	10	**3.0E-04**	0.96	0.01	diff	gene262	1
9637	STE12	98	**3.8**	76	3	19	0.05	**3.3E-04**	127.24	up	gene3811	1
9579	MET4	177	**6.8**	137	21	19	**3.5E-04**	1.26	**4.8E-03**	down	gene4937	1
9602	PUT3	33	**5.3**	21	2	10	**4.0E-03**	**5.0E-04**	37.68	up	gene3852	1
9563	HSF1	231	**4.8**	184	24	23	**1.3E-03**	1.68	0.01	diff		
9548	GCR1	414	**3.1**	350	39	25	0.27	96.38	**1.5E-03**	down	gene759	0.2
9545	GAT3	5	**4.6**	1	2	2	**2.9E-03**	0.41	0.18	diff		
	SPT7	69	**4.7**	50	12	7	0.01	11.89	**3.0E-03**	down	gene4603	1
	HMS1	14	**4.4**	7	4	3	0.01	1.68	0.06	NA	gene3793	1

Those involved with down-regulated gene sets, in order of the enrichment statistics, included Gcn4, Gcr2, Tup1, Spt3, Hfi1, Spt20, Met32, Aft1, Cdc73, Sin4, Fkh2, Rtg3, Ba1, Sfp1, Met4, Gcr1 and Spt7. Those involved with up-regulated gene sets were Ume6, Bur8, Sua7, Spt10, Ste12 and Put3. Only for Sin3, Hsf1 and Gat3 was the pattern unclear. A benefit of this method (SGD) is that it reveals regulators that do not necessarily have DNA binding sites associated, like Tup1.

Notably, out of all enriched regulators, only the genes for Gcr1 and Gcr2 were differentially expressed at transcription level. These were both down-regulated in the xylose medium where glycolytic genes were all down-regulated [Chapter 3, Schabort et al. 2016], which was consistent with their role as activators of glycolysis. The activator Adr1 and the repressor Mig1, which were revealed in a previous chapter [Chapter 4] using the enumerative heptamer frequency method as being important in the response, were not enriched.

Emphasising direct physical evidence for interactions

Many of the gene regulatory interactions captured in the SGD originated from transcriptional profiling. A fraction of these interactions may have originated from secondary effects which were the result of a regulator that regulates another major regulator. Direct physical evidence may be a strong indicator of true interactions. A simple scoring system was designed to assign a relative strength to each of a variety of types of experimental evidence contained in SGD, weighting the scores higher towards direct binding interactions such as chromatin immunoprecipitation (ChIP). The scoring system along with evidence codes is given in Table 2. Transcriptional profiling is also a useful, complementary method, and to combine the various sources of information, a simple summation of all forms of evidence of an interaction was used. Figure 1 shows the distribution of scores of the experimental evidence in SGD. Finally, a gene regulatory network was calculated as well as enrichment values for the top *n* among all interactions. For the top 5 855 interactions, strong evidence of direct physical interactions or multiple independent transcriptional profiling types were found, while for the majority,

evidence was found based on transcriptional profiling. Table 3 shows the numbers of interactions containing a certain score *Li* or higher, which were used as thresholds for assigning gene regulatory networks.

Table 2. The scoring system used to weight the various forms of experimental evidence supporting a regulatory interaction from SGD. The scores were weighted stronger towards evidence based on direct physical interaction with the regulatory regions as opposed to transcriptional profiling. For a final likelihood, *Li*, the scores for various sources were summed. A value of zero was applied for no experimental evidence.

annotation type	experiment type	number of occurrences	score	evidence code
binding enriched	genome-wide gene expression regulator binding enrichment	15486	2	1
expression activated	genome-wide gene expression regulator mutant expression profile	5719	2	2
binding enriched with conserved binding site	genome-wide gene expression regulator binding enrichment with conserved binding site	3782	4	3
	microarray RNA expression	2900	1.5	4
expression repressed	genome-wide gene expression regulator mutant expression profile	2538	2	5
expression repressed	microarray RNA expression	812	1.5	6
expression repressed	ethanol/glucose limitation	500	1.5	7
expression activated	microarray RNA expression	489	1.5	8
binding enriched	chromatin immunoprecipitation-chip assay	417	8	9
bound	chromatin immunoprecipitation-chip assay	404	8	10
	chromatin immunoprecipitation-chip assay	340	8	11
expression activated	ethanol/glucose limitation	197	1.5	12
bound	chromatin immunoprecipitation-seq assay	154	8	13
activated	microarray expression profiling	88	1.5	14
binding enriched	chromatin immunoprecipitation assay	10	8	15
	ethanol/glucose limitation	1	1.5	16
binding enriched/expression repressed	chromatin immunoprecipitation assay	1	8	17

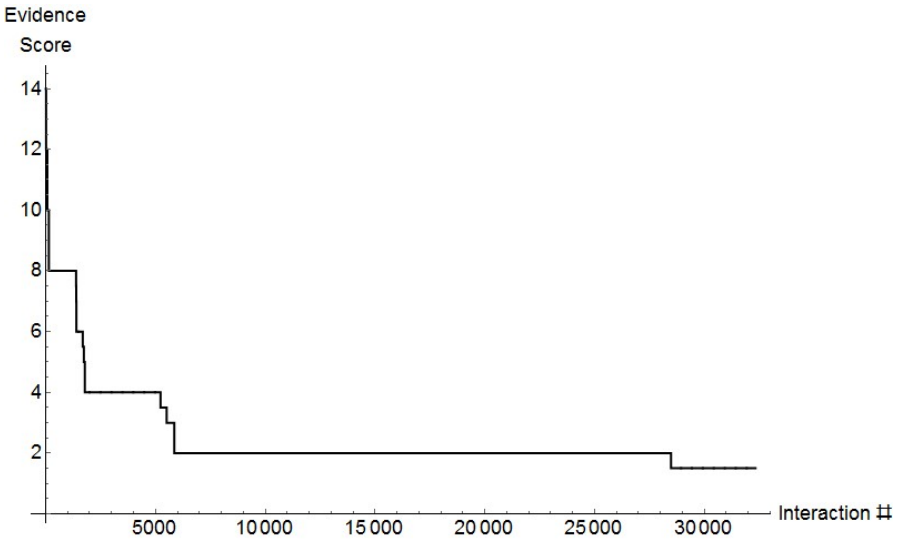


Figure 1. Scoring profile based on experimental evidence in SGD. The evidence score of each type of experiment supporting the interaction was summed for the final scores *Li*.

Table 3. The scoring of interactions along with the cumulative distribution function.

score	number	CDF
14	16	16
12	53	69
10	75	144
8	1235	1379
7.5	9	1388
7	5	1393
6	295	1688
5.5	46	1734
5	49	1783
4	3448	5231
3.5	275	5506
3	349	5855
2	22632	28487
1.5	3851	32338

Each of the values in the cumulative distribution function were used as a cut-off for generating a gene regulatory network and to test for the enrichment of transcription regulators. Figures 2-5 show the result of using the four different enrichment statistics (the Z-score method and three hypergeometric tests). The overall discovery pattern corresponded very well between the method based on the Z-score, which incorporates the strength of the differential expression of each genes, and the hypergeometric method of testing for differential expression, which is based on discrete statistics. It seems that by including scores of 8 or above only (1 379 of the top interactions), the most significantly enriched regulators were discovered, even though the set of effectors was small. This suggested both that the high-scoring ChIP experiments revealed the true interactions and that perhaps there was bias in the database towards experiments that focussed on the main regulators in the glucose de-repression response. Notably, the enrichment scores for the down-regulated sets were much more significant as compared to the up-regulated sets. Further, while there was a general increase in enrichment scores with an increase in the number of interactions, there was a drop and recovery in the score of the up-regulated gene statistics.

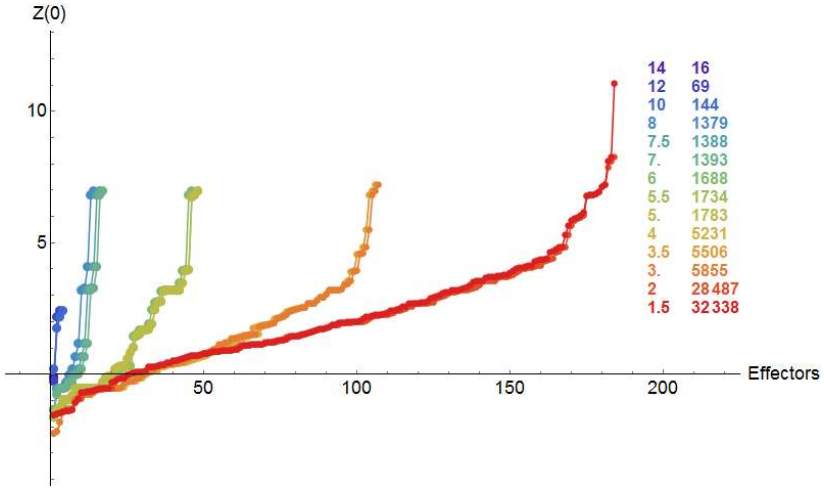


Figure 2. Enrichment scores based on Z-scores as a function of the threshold score for experimental evidence in *S. cerevisiae* (*Li*).

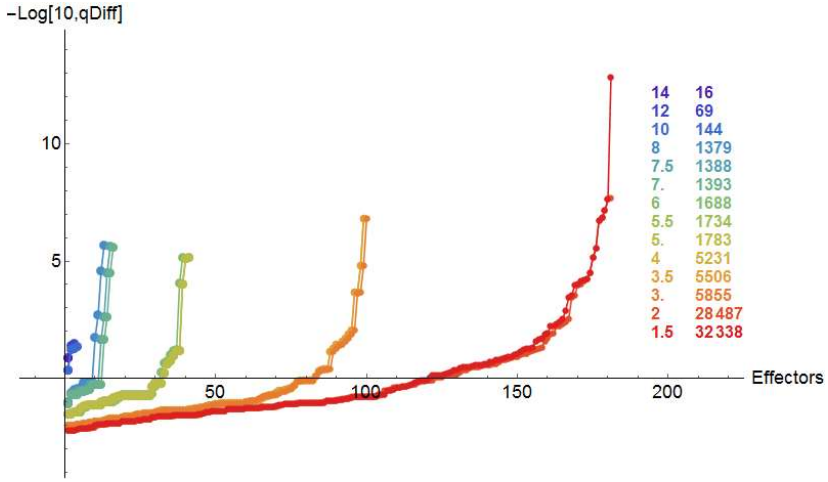


Figure 3. Enrichment scores based on hypergeometric p-values for differential expression, adjusted for multiple comparisons, as a function of the threshold score for experimental evidence in *S. cerevisiae* (*Li*).

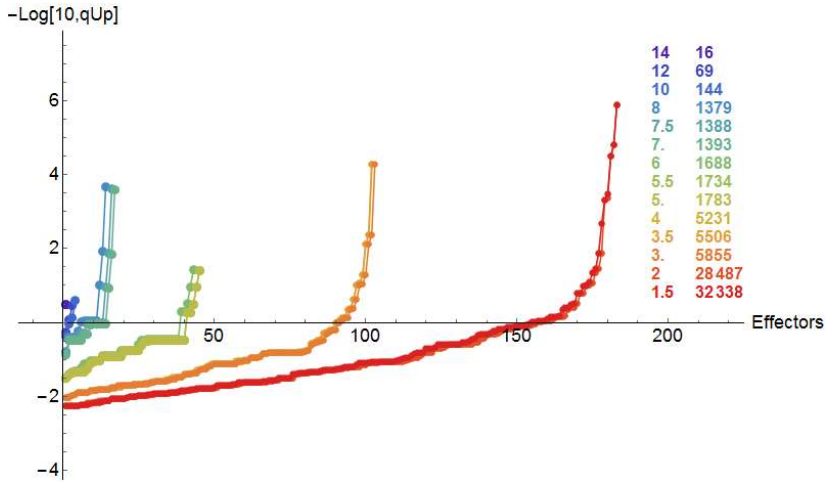


Figure 4. Enrichment scores based on hypergeometric p-values for up-regulation, adjusted for multiple comparisons, as a function of the threshold score for experimental evidence in *S. cerevisiae* (*Li*).

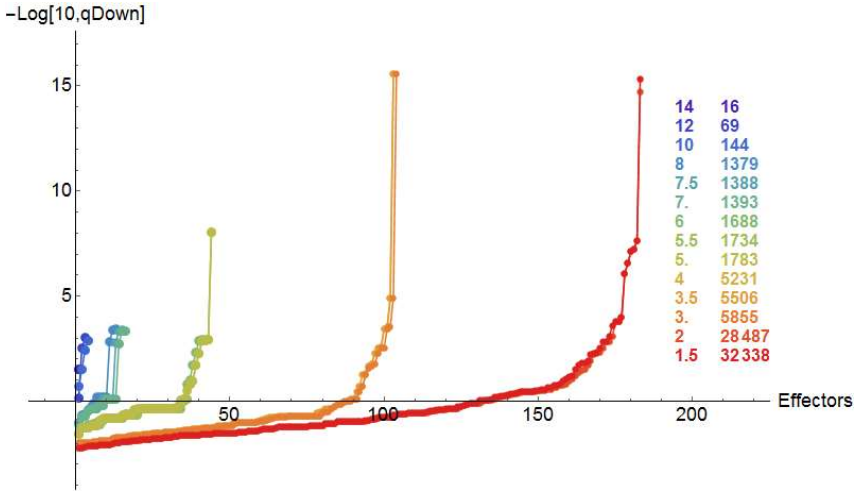


Figure 5. Enrichment scores based on hypergeometric p-values for down-regulation, adjusted for multiple comparisons, as a function of the threshold score for experimental evidence in *S. cerevisiae* (*Li*).

What the contribution of each type of evidence might be was also investigated (Figures 6 and 7). To this end, the final score *Li* for each interaction was taken as the maximum evidence score among different types of experiments supporting the interaction, instead of the sum (Figure 6). The inclusion of the transcriptomic experiment in which a conserved binding site was also found in *S. cerevisiae* ("genome-wide gene expression regulator binding enrichment with conserved binding site", evidence code 3 with score 4) seems to have provided many additional effectors with intermediate enrichment, although the most significantly enriched effectors were included already with the high-scoring ChIP data (Figure 7). Some of the microarray data types provided redundant information, as there was seemingly no difference between the scores of 1.5 and 2 in terms of the overall enrichment profile

(Figure 7). In addition, the enrichment scores are shown using the hypergeometric test for differential expression (Figure 8), up-regulation (Figure 9) and down-regulation (Figure 10).

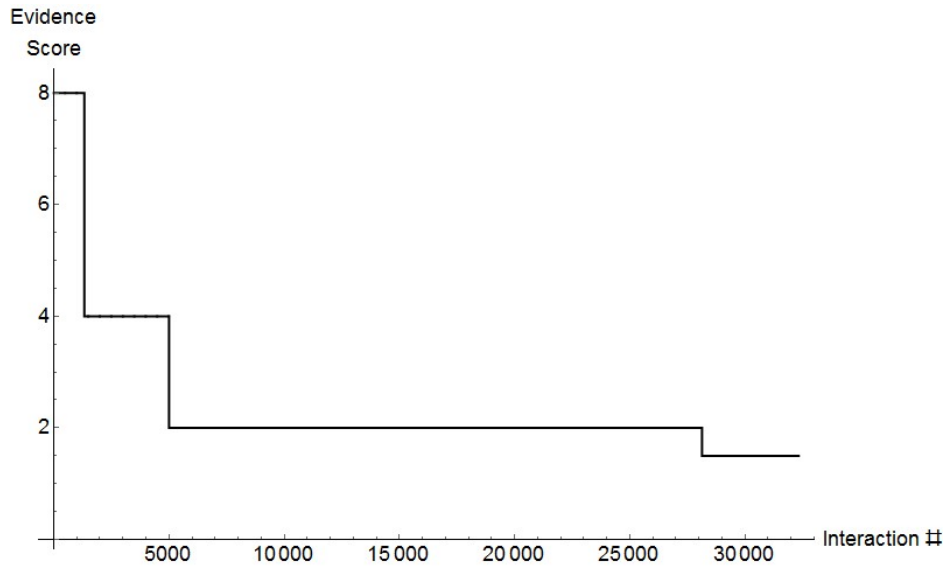


Figure 6. The scoring profile based on experimental evidence in SGD using maximum evidence scores. To calculate the final score for each interaction, the maximum evidence score among different types of experiment supporting the interaction was taken to calculate *Li*.

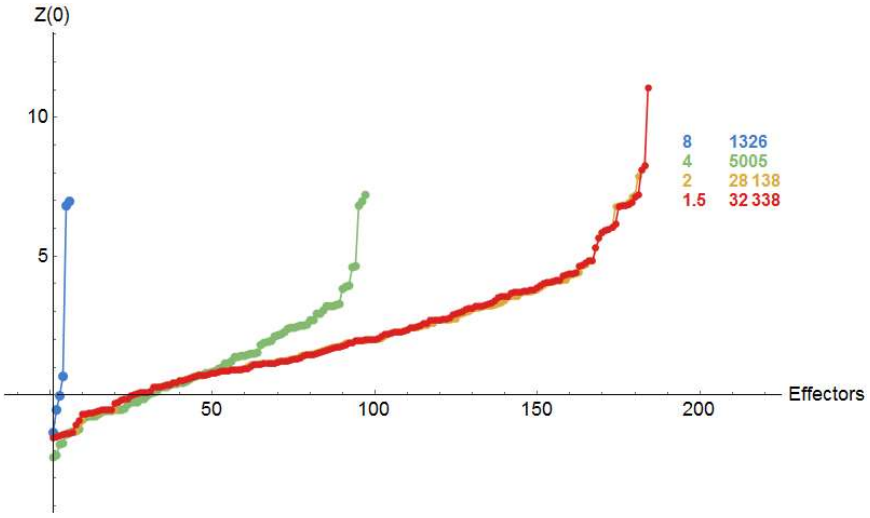


Figure 7. Enrichment scores based on Z-score as a function of the threshold for the confidence in experimental evidence in *S. cerevisiae* using maximal evidence scores for *Li*.

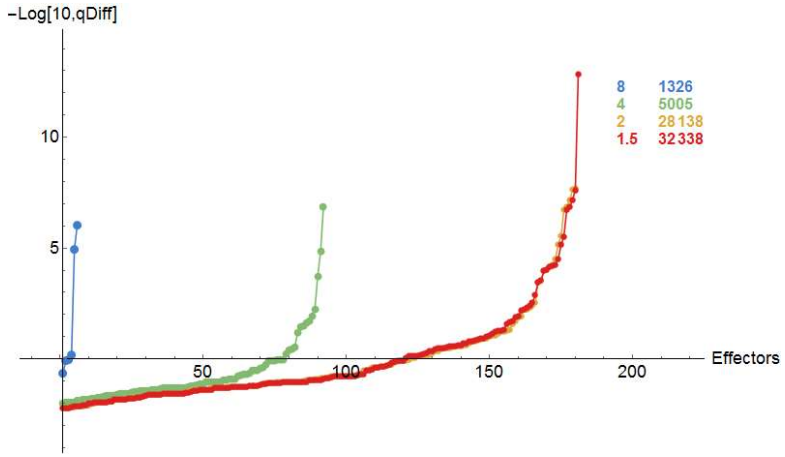


Figure 8. Enrichment scores based on hypergeometric p-values for differential expression as a function of the threshold for the confidence in experimental evidence in *S. cerevisiae* using maximal evidence scores for *Li*.

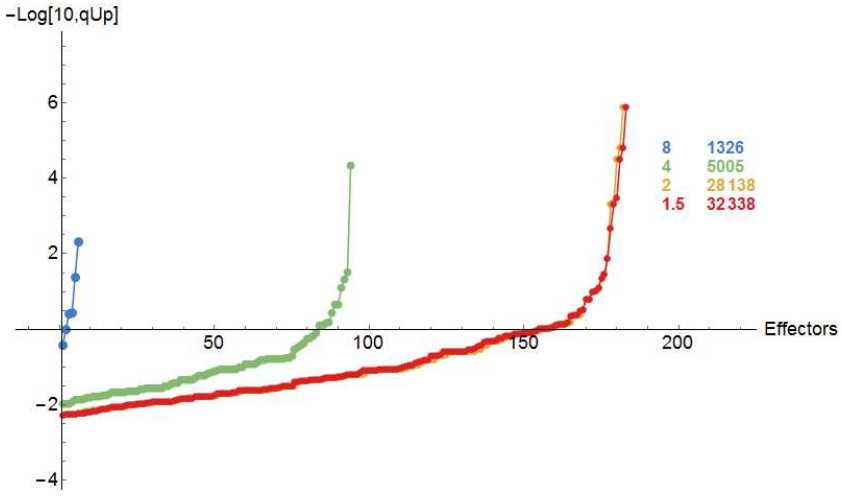


Figure 9. Enrichment scores based on hypergeometric p-values for up-regulation as a function of the threshold for the confidence in experimental evidence in *S. cerevisiae* using maximal evidence scores for *Li*.

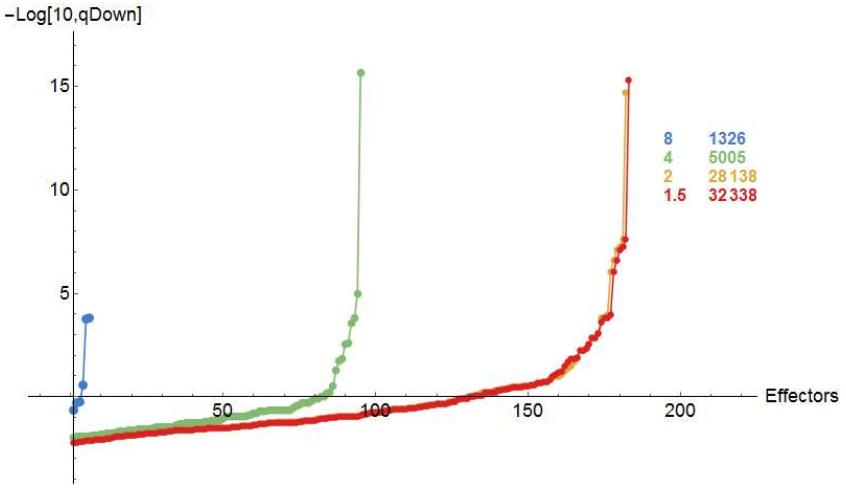


Figure 10. Enrichment scores based hypergeometric p-values for down-regulation as a function of the threshold for the confidence in experimental evidence in *S. cerevisiae* using maximal evidence scores for *Li*.

Evidently, the enrichment score of the top effectors did not seem to decline with an increased number of interactions, while at the same time the number of effectors were increased. This suggested that the different experiments provided evidence for different effectors, that the data were generally applicable to the different species, and that various forms of data strengthened the confidence in assignments. A good separation between highly enriched and non-enriched regulators could be considered a sign of improved accuracy of the network. The hypergeometric test appeared to better elucidate those regulators that were important, compared to the Z-score method. Thus, the best function seemed to be to sum the scores in the experimental evidence and for the purpose of elucidating active regulators of gene expression, all the evidence could be combined using the evidence code developed, if only one threshold for the evidence score *Li* had to be chosen.

The enrichment score increases of individual effectors can be seen in Figure 11. Significance was reached for most of the important effectors only when transcriptome profiling data were included. For Bur6, Sua7, Cdc73 and Spt7, the discovery of their apparent importance in *K. marxianus* was only based on the transcriptome profiling data of *S. cerevisiae*. Only for one regulator, Met32, the scores fell from significant to insignificant with the final inclusion of all transcriptome profiling data. The pattern of decreasing significance with additional transcriptome profiling was shared by Met4 only. Ume6 was initially significantly enriched. Subsequently, its score fell below significance, only to reach a higher level of significance with all the data included.

In summary, the scoring system worked very well, and all data could be included for the best general network. However, choosing the most significantly enriched target set for each regulator would avoid

missing some significantly enriched regulators, for which some of the transcription profiling data may have resulted from secondary effects. An added benefit in using the SGD data alone is that the effectors are not only transcription factors that bind directly to DNA, but also include other modifiers of chromatin, such as kinases and acetylases, which associate with certain genes and affect their expression.

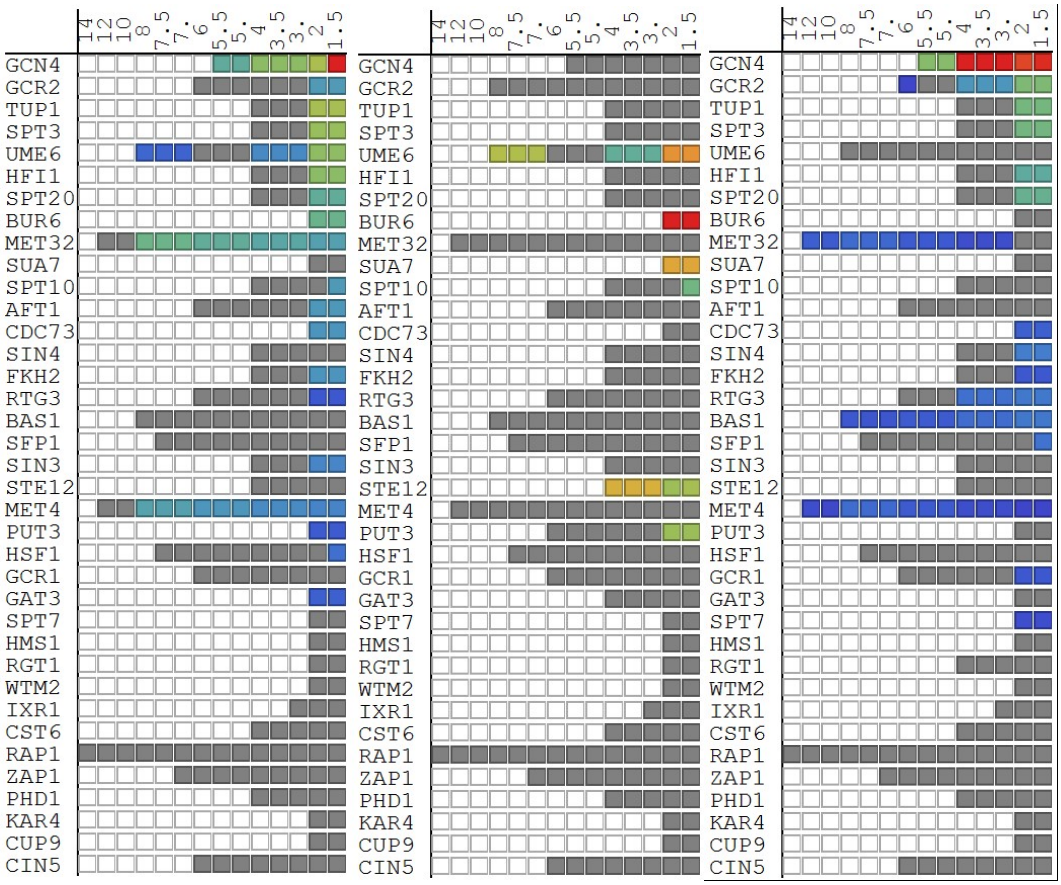


Figure 11. Enrichment statistics of effectors, calculated with the hypergeometric distribution as a function of the total evidence score for an interaction. Hypergeometric p-values were corrected for the number of comparisons by multiplication with 184 to result in q-values. Grey blocks indicate that the regulator had at least one target. Coloured blocks indicate that the q-value was at or below 0.005 (enrichment statistic –Log(q) >= 2.301). Left: statistic for differential expression; Middle: statistic for up-regulation; Right: statistic for down-regulation. Warmest colours were chosen for the minimum q-value in each of the three test sets.

A gene regulatory network based on the complete genome of DMKU3-1042

The experimental data on interactions in the model species proved valuable to suggest active transcription factors in *K. marxianus*, regardless of the presence of DNA binding sites in this species. The presence of binding sites in *K. marxianus* could now be combined with the experimental evidence of interactions in *S. cerevisiae*, as well as motif conservation among *Kluyveromyces* species, in the likelihood framework.

The nucleotide composition in the complete genome of strain DMKU3-1042 changes as a function of distance from the TLSS. Figure 12 depicts the background frequency of each nucleotide base as a function of distance from the TLSS using the complete genome for strain DMKU3-1042 [Lertwattanassakul et al. 2015]. The pattern is almost identical to that found with strain UFS-2791 [Chapter 5].

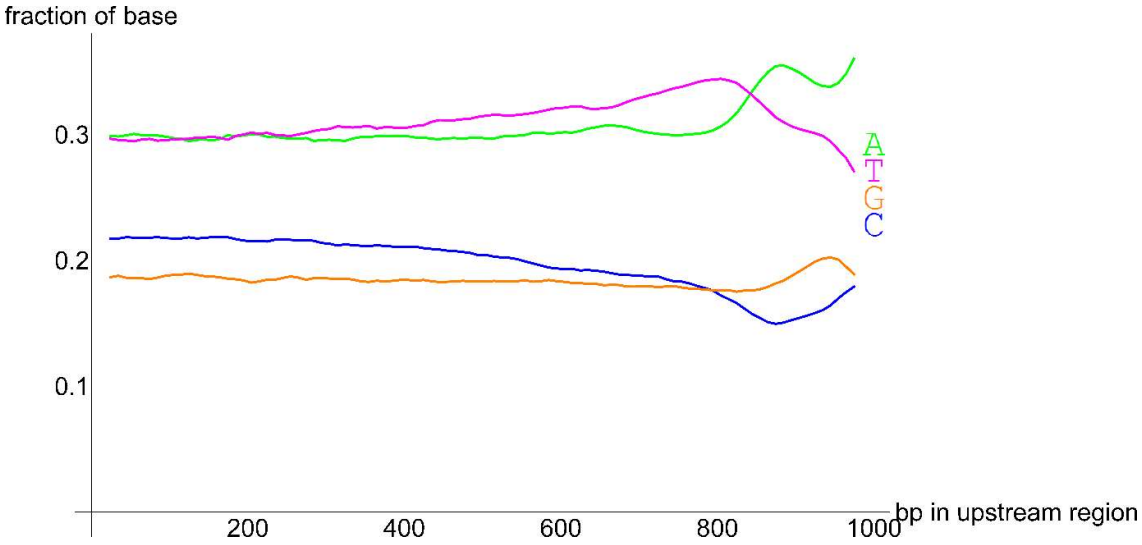


Figure 12. The background frequency of each base as a function of distance from the TLSS. The position of the TLSS was taken as 1 000 bp, thus 1 000 bp from the TLSS is indicated by position number 1.

The distributions of scores for the three basic forms of data used in the likelihood framework are shown in Figures 13 and 14. Very large scores were calculated for the long, informative motifs. Using Log_{10} of the motif likelihood score, *Lm*, suppressed the dominant effect of the motif score alone. The large number of motifs found with *Lm* value above 100 (892 106) made it necessary to utilise motif conservation and experimental evidence from the model species to a large extent. The distribution of motif conservation scores, *Lc*, showed a large number of motifs with a score of approximately 1,

indicating that the motif occurred in a well conserved region [see Chapter 5 and Addenda 2 and 3 for explanation of the method]. The score was well above 1 for only for a small fraction of motifs, which corresponded to motifs that were better conserved compared to the immediately neighbouring regions. For 24.7% of motifs, no alignment was obtained to any of the other six *Kluyveromyces* genomes (Figure 13). The distribution of the experimental evidence coupled to the gene corresponding to each motif, *Li*, is also shown in Figure 14. For the vast majority of motifs there was no supporting experimental evidence (Figure 13), hence providing a strict limitation to the inclusion of an interaction, if *Li* values were to be multiplied with $Log_{10}(Lm)$ or *Lc*.

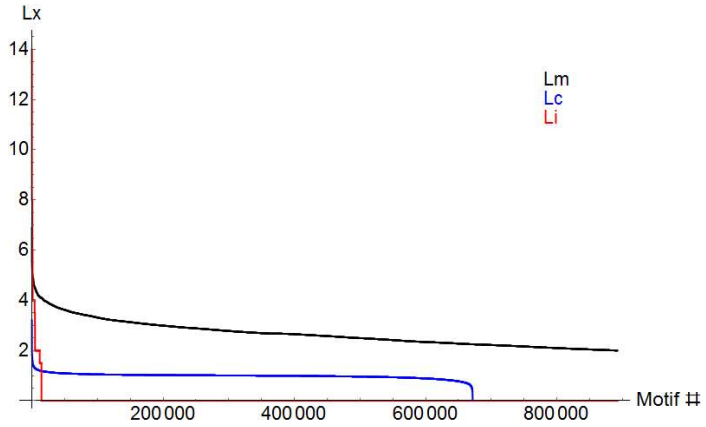


Figure 13. Distribution of all scores for motif likelihood ($Log_{10}(Lm)$), motif conservation among *Kluyveromyces species* (*Lc*) and experimental evidence coupled to the gene corresponding to each motif (*Li*).

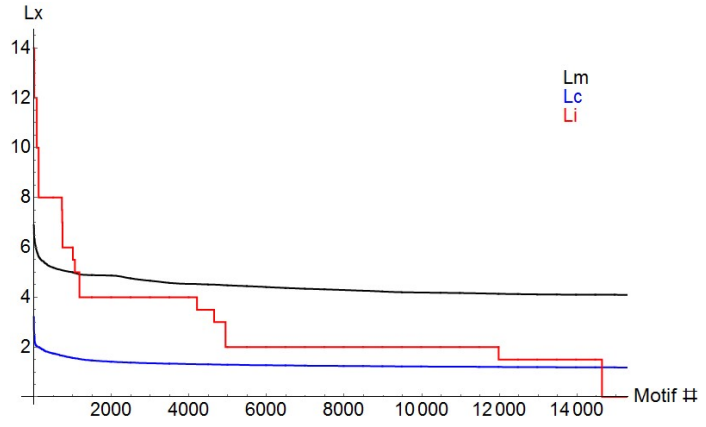


Figure 14. Distribution of the top 14 800 scores for motif likelihood ($Log_{10}(Lm)$), motif conservation among *Kluyveromyces species* (*Lc*) and experimental evidence coupled to the gene corresponding to each motif (*Li*).

The distribution of the final likelihood, *L*, is shown in Figure 15, as calculated using seven different likelihood functions. When using the function $Log_{10}(Lm)\times Lc\times Li$, 14 642 motifs were included and thus the vast majority of possible interactions were discarded. However, since this did not allow the discovery of any of the potentially important transcription factors discoverable by motifs scans when no model species evidence was available, other functions were also investigated. Function C gave a final likelihood distribution in-between that of A and B (Figure 15).

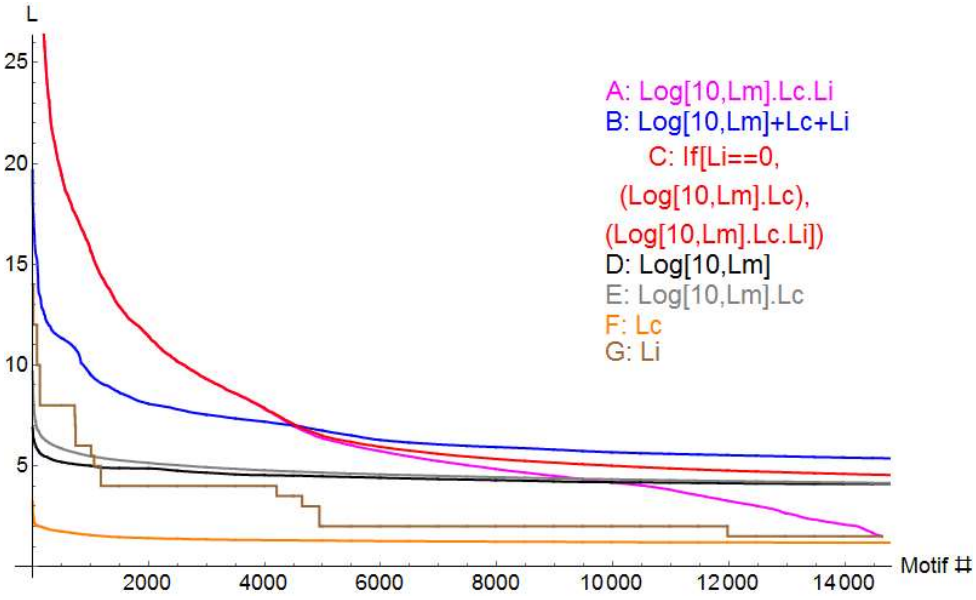


Figure 15. Distributions of final likelihood scores based on seven likelihood functions.

The number *n* of interactions was varied to construct various networks and the enrichment statistic was calculated for each target set of a TF. The number of interactions *n* was varied from 500 to 14 500. The latter value is approximately the number of regulatory interactions for which there was evidence in *S. cerevisiae*. Figures 16-19 show the enrichment scores obtained for all TFs when using the function A ($Log_{10}(Lm)\times Lc\times Li$). An additional innovation in this representation was the inclusion of the name of each transcription factor on the plot, as well as the relative number of targets indicated as grey disks. In this format, the large target gene sets are situated on the near-horizontal line of non-significance, whereas the enriched transcription factors rise out from this area. The increased separation of significant from insignificant sets was evident for the hypergeometric method (compare Figures 16 and 17). Gcn4 was again revealed as the most important regulator of down-regulated targets, which was followed by Met32 and Arg81 (Figure 19). Ste12 was significant for the up-regulated targets (Figure 18). Notably, the down-regulated set was more significant than the up-regulated set.

As the amount of data caused a cluttering of data points, a better approach to visualisation was to use the block graphics shown in Figures 20 and 21. The same features were visible as in the list plots (Figures 16-19), but additional features became visible. This format is very powerful since it gives a clearer picture of how enrichment of each regulator changes as a function of the parameter varied; in this case, the number of high-scoring motifs allowed for inclusion into the network. This block format was generated for all functions (see Addendum 4, Figures 1-12), and a summary is given in a later section.

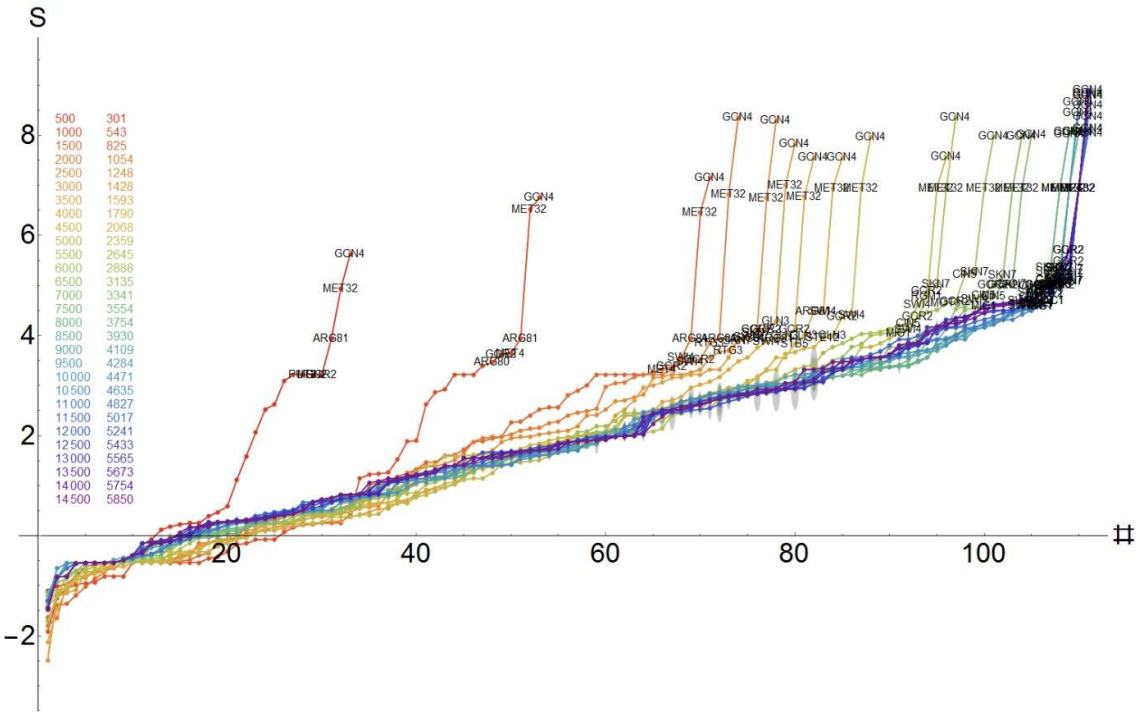


Figure 16. Enrichment scores based on Z(0) values as a function of the number of motifs included in the likelihood framework using $Log_{10}(Lm)\times Lc\times Li$ for calculation of final likelihoods. Grey disks indicate the relative number of targets of each effector.

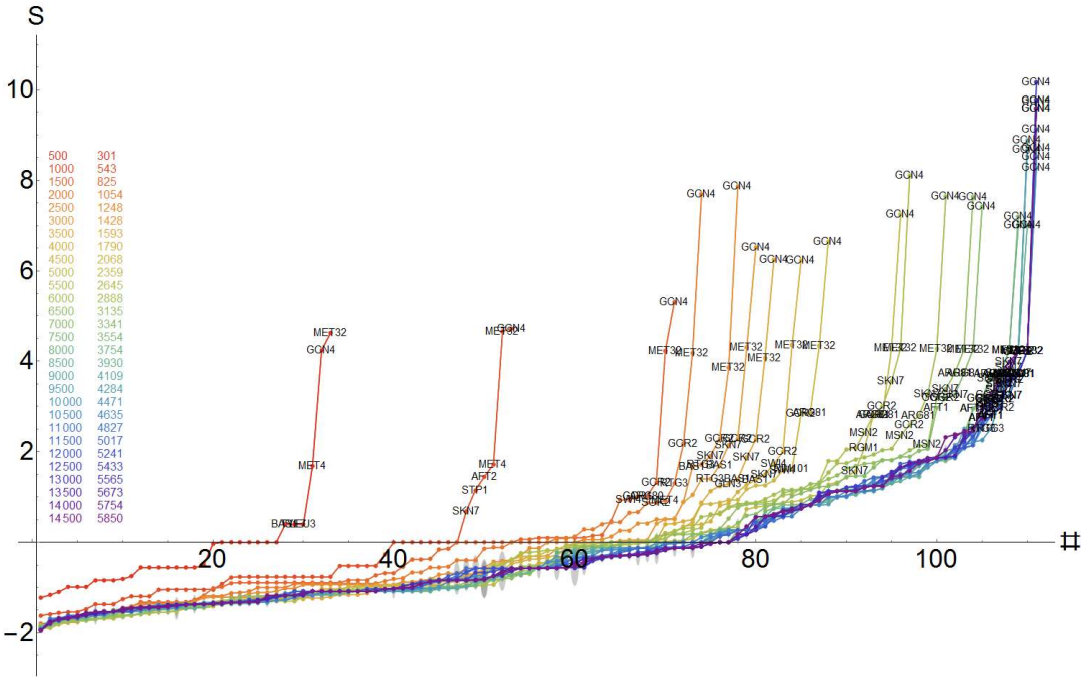


Figure 17. Enrichment scores based on the q value for differential expression, using the hypergeometric distribution, as a function of the number of motifs included in the likelihood framework using $\mathrm{Log}_{10}(Lm)\times Lc\times Li$ for calculation of final likelihoods.

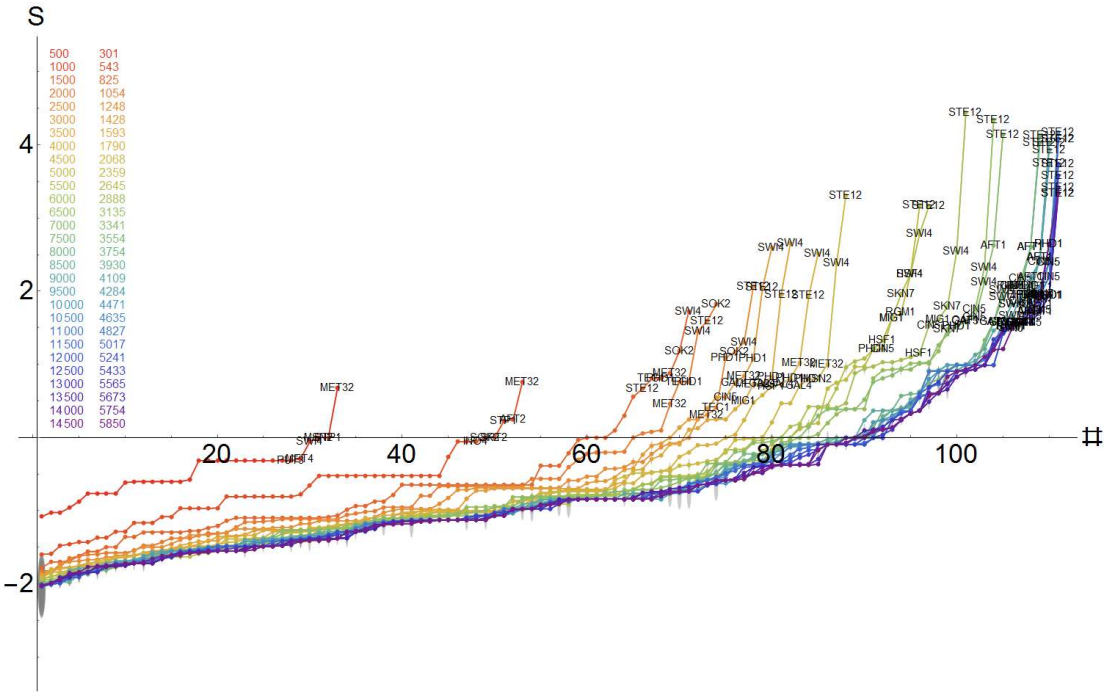


Figure 18. Enrichment scores based on the q value for up-regulation, using the hypergeometric distribution, as a function of the number of motifs included in the likelihood framework using $\mathrm{Log}_{10}(Lm)\times Lc\times Li$ for calculation of final likelihoods.

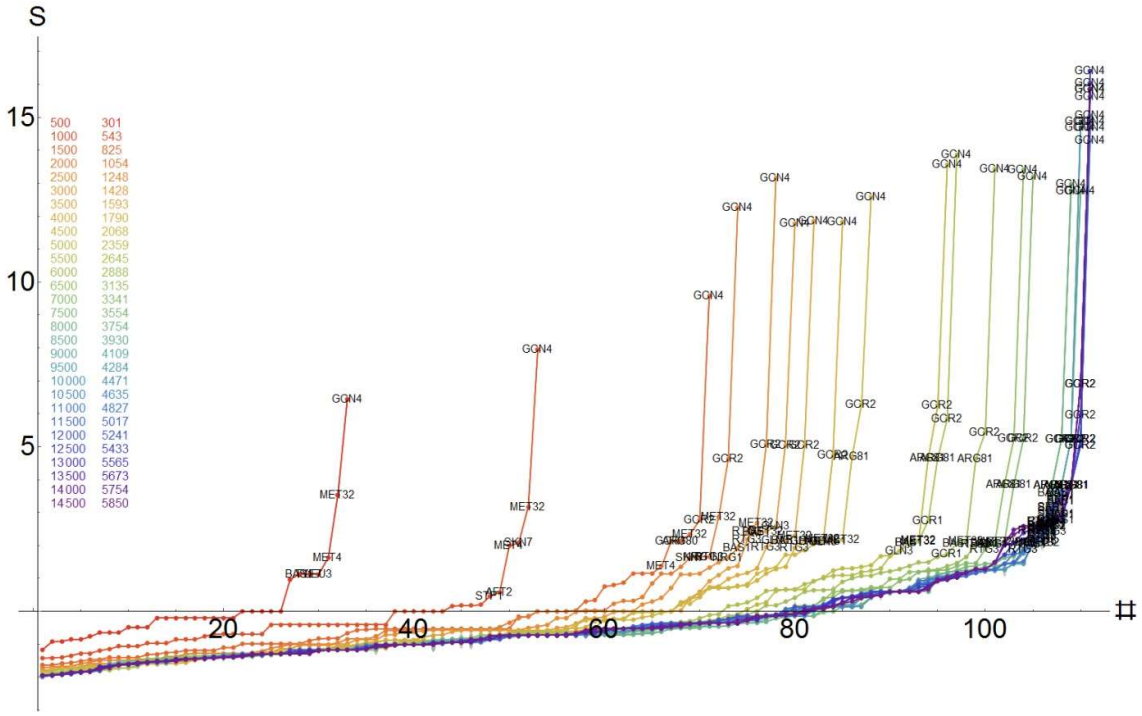


Figure 19. Enrichment scores based on the q value for down-regulation, using the hypergeometric distribution, as a function of the number of motifs included in the likelihood framework using $Log_{10}(Lm)\times Lc\times Li$ for calculation of final likelihoods.

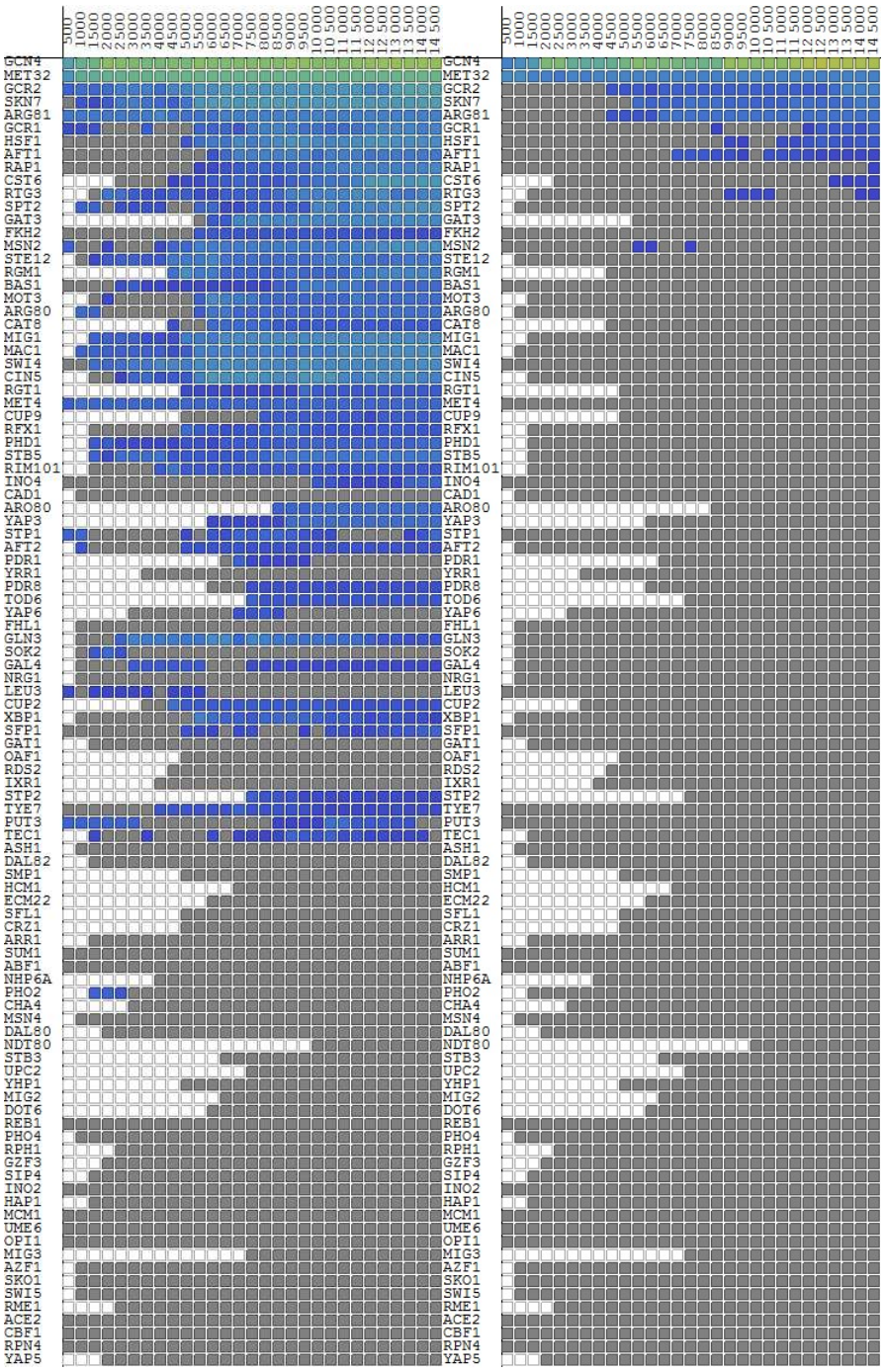


Figure 20. Enrichment statistics calculated for differential expression in gene sets using the Z-score method (left) or the hypergeometric distribution (right) for networks constructed using the function $Log_{10}(Lm)\times Lc\times Li$ for the final likelihood. Hypergeometric p-values were corrected for the number of comparisons by multiplication with 184. Grey blocks indicate that the effector had at least one target. Coloured blocks indicate that the q-value was at or below 0.005 for the hypergeometric tests (enrichment score >= 2.301) or that the enrichment statistic for the Z-score method was at or above 1.64 (p <= 0.05). The warmest colour (red) was chosen as 16, which was the highest enrichment score among all tests (Gcn4, down-regulation). White blocks indicate no target genes for a TF.

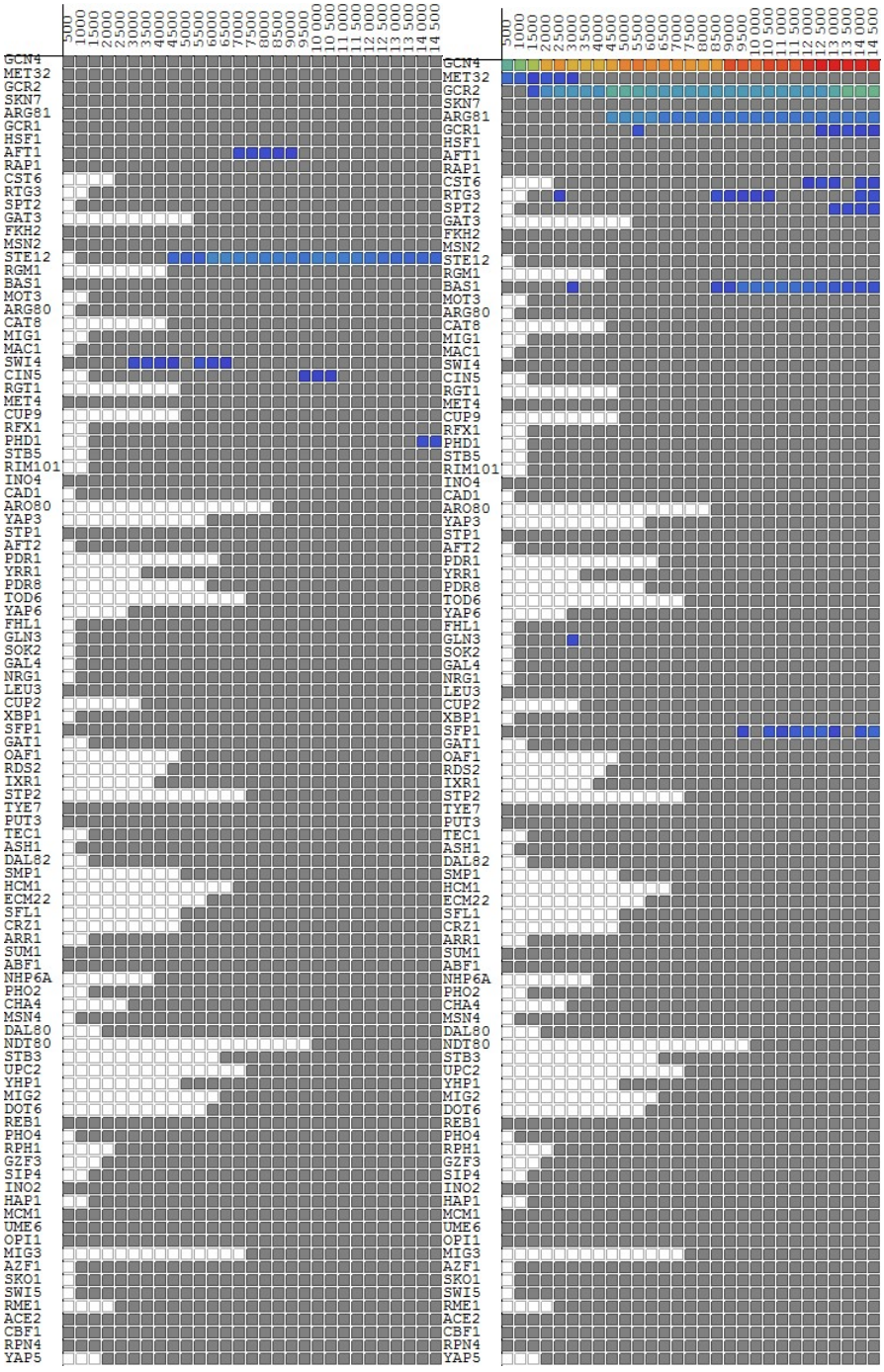


Figure 21. Enrichment statistics calculated for up-regulation (left) and down-regulation (right) using the hypergeometric distribution for networks constructed using the function $Log_{10}(Lm)\times Lc\times Li$ for the final likelihood.

The enumerative heptamer frequency method as independent measure to elucidate active transcription factors

After correcting for multiple comparisons, five heptamers were found to be significantly over-represented in the regulatory regions, using the enumerative method. The results are summarised in Table 4. A k-mer network was also drawn to cluster TFs based on their similarity with the same hemptamers (Figure 22).

Table 4. Over-represented heptamers in the upstream regulatory regions of 245 down-regulated genes, with a q-value smaller or equal to 0.05 after correcting p-values for multiple comparisons using the total number of heptamers (16 384) as multiplier. The minimum fractional match score between heptamer and PPM was set to 0.7.

JASPAR Matrix ID	TF	Heptamer	p	q	motif score/ potential	pattern
9540	FKH1	AAAAAAA	3.82E-23	6.3E-19	0.781987	[CT][AG]AA[AT]TGTAAACAAA[AG][AG][AG][GT][AG]
9634	STB3	AAAAAAA	3.82E-23	6.3E-19	0.751023	G[GT][CT][CT]AAA[AT]TTTTTCACT[CT][AT][GT][GT]
9521	AZF1	AAAAAAA	3.82E-23	6.3E-19	0.745638	AAAAAGAAA
9587	NDT80	AAAAAAA	3.82E-23	6.3E-19	0.722486	[AT][CT][CT][CT][AC][GT]G[AC]CACAAAA[AT]C[CG][AC][AT][AC]
9547	GCN4	TGACTCA	1.04E-07	0.0017	0.903521	[CG][AG][AT][AG][AG][GT]ATGAGTCAT[AC][CT][AT][CT][AC][AT]
9516	ARG81	TGACTCA	1.04E-07	0.0017	0.881325	[AG]TGACTC[ACT]
9522	BAS1	TGACTCA	1.04E-07	0.0017	0.819378	[GT]C[AT][CT][AG]GCC[AC]GAGTCA[AG][AG][AT][CT][AG][AG]
9515	ARG80	TGACTCA	1.04E-07	0.0017	0.78727	AGAC[GT]C
9663	YAP7	TGACTCA	1.04E-07	0.0017	0.736474	ATTAGTAAGCA
9532	CUP9	TGACTCA	1.04E-07	0.0017	0.716854	TGACACAT[AT]
9540	FKH1	TTTTTTT	4.05E-07	0.0066	0.781987	[CT][AG]AA[AT]TGTAAACAAA[AG][AG][AG][GT][AG]
9634	STB3	TTTTTTT	4.05E-07	0.0066	0.751023	G[GT][CT][CT]AAA[AT]TTTTTCACT[CT][AT][GT][GT]
9521	AZF1	TTTTTTT	4.05E-07	0.0066	0.745638	AAAAAGAAA
9587	NDT80	TTTTTTT	4.05E-07	0.0066	0.722486	[AT][CT][CT][CT][AC][GT]G[AC]CACAAAA[AT]C[CG][AC][AT][AC]
9585	MSN2	CCCCCCC	2.32E-06	0.0380	0.864646	AGGGG
9534	DAL81	CCCCCCC	2.32E-06	0.0380	0.857143	AAAAGCCGCGGGCGGGATT
9586	MSN4	CCCCCCC	2.32E-06	0.0380	0.830323	AGGGG
9610	RGM1	CCCCCCC	2.32E-06	0.0380	0.754	AGGGG
9675	YML081W	CCCCCCC	2.32E-06	0.0380	0.748571	[AC]CCCC[GT]C[AT][CT]
9669	YGR067C	CCCCCCC	2.32E-06	0.0380	0.745347	[AG]CCCC[AG]C[AT][CT][CT][AGT][GT][CGT][AG]
9582	MIG2	CCCCCCC	2.32E-06	0.0380	0.724971	CCCCGC[ACG]
9514	AFT2	GTGTGTG	2.39E-06	0.0391	0.735714	[CG]ACACCC[CG]
9685	ZMS1	CCCCCCC	2.32E-06	0.0380	0.714388	T[AT]CCCCGC[AT]
9581	MIG1	CCCCCCC	2.32E-06	0.0380	0.70368	[AC]CCCC[AG]C
9680	YPR022C	CCCCCCC	2.32E-06	0.0380	0.701556	CCCCAC[CG]

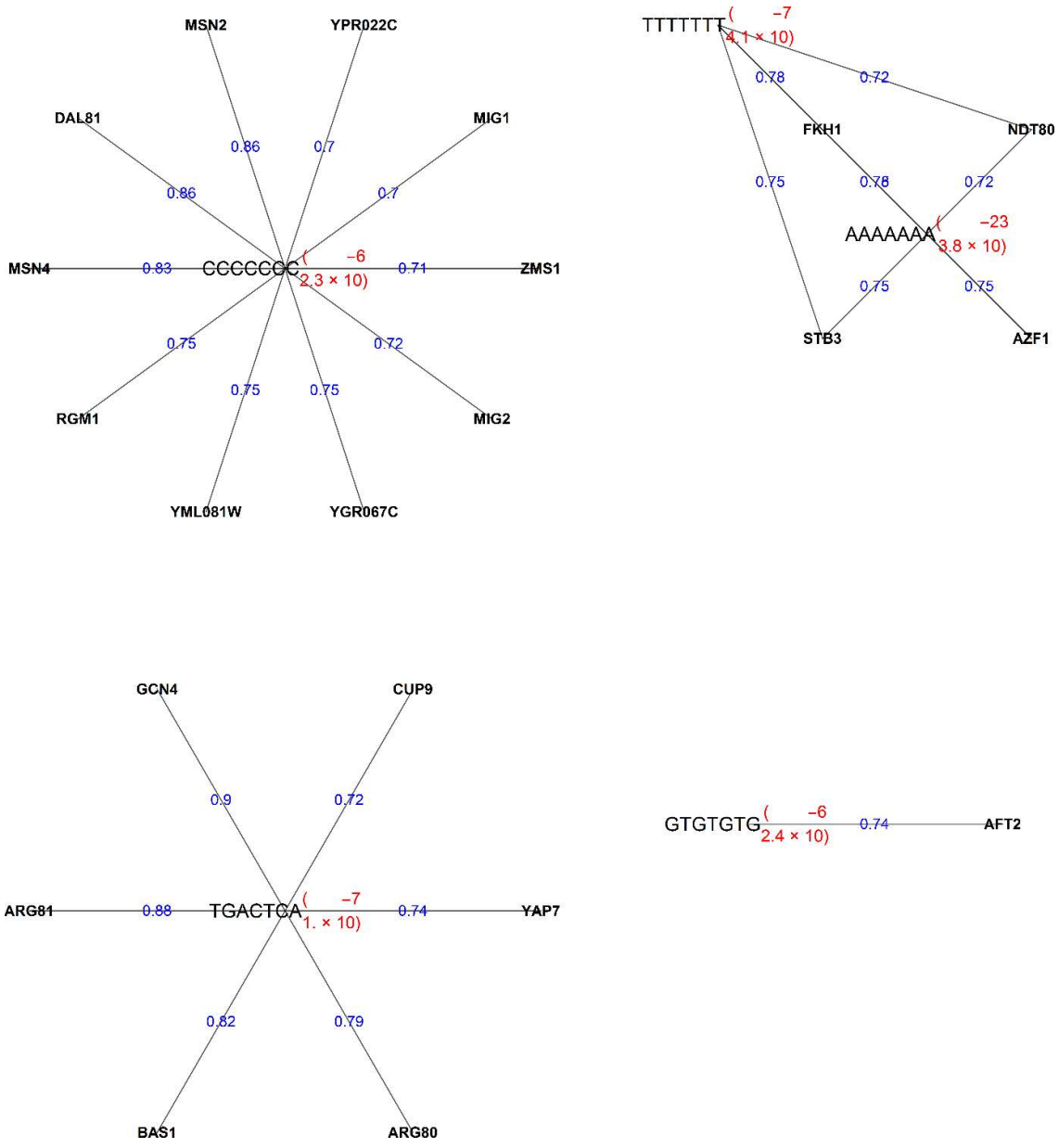


Figure 22. Heptamer network of over-represented heptamers in upstream regulatory regions of 245 down-regulated genes, with a q-value smaller or equal to 0.05 after correction for multiple comparisons using the total number of heptamers (16 384) as multiplier. The minimum fractional match score between heptamer and PPM was set to 0.7.

Notably, poly A, poly T and poly C were in this list. Also, poly G was ninth in terms of over-representation, and TATATAT was at number 22. Fkh1, Stb3, Azf1 and Ndt80 mapped to poly A and poly T. Polynucleotide stretches as well as the TATA-like sequences render a region of DNA likely to have an open structure, devoid of nucleosomes [Radman-Livaja et al. 2010]. Some of these stretches

have also been seen in the up-regulated set as over-represented [Chapter 4]. The over-representation of these may thus suggest open DNA in all differentially expressed genes and not necessarily that the transcription factors Fkh1, Stb3, Azf1 and Ndt80 were differentially active. TATA-box elements also indicate environmentally responsive genes [Radman-Livaja et al. 2010]. This patterning will form the basis of future studies. Whatever the case may be, the interpretation of these sequences should be done with caution. None of the transcription factors bind to stretches of exclusively polynucleotides. Their motifs always include some bias towards alternative bases. Consequently, this suggests that those transcription factors that map to these heptamers deserve some caution. Nevertheless, it is remarkable that again Gcn4 stands out as a regulator of down-regulated genes. It had the closest match among all transcription factors to the heptamer TGACTCA. On close inspection of the top 30 heptamers (Table 5), several heptamers were found to match closely to TGACTCA. The Occam's razor approach was used again, as was performed for up-regulated genes in Chapter 4. The combined motif is shown in Figure 23.

Table 5. The top 30 heptamers from the enumerative approach using the regulatory regions from down-regulated genes. Heptamers with a strong similarity are indicated in boldface.

K-mer	p	q
AAAAAAA	3.82E-23	6.27E-19
TGACTCA	1.04E-07	0.001697
TTTTTTT	4.05E-07	0.006644
CCCCCCC	2.32E-06	0.037997
GTGTGTG	2.39E-06	0.039096
TGAGTCA	9.09E-06	0.148981
GAAGCCC	1.12E-05	0.183445
TGTGTGT	1.16E-05	0.189646
GGGGGGG	1.72E-05	0.281152
TGACACA	1.8E-05	0.294789
GCCCGAT	1.93E-05	0.316475
GAGCACC	2.79E-05	0.457846
CCCCTGG	3.3E-05	0.539999
GTGAGTG	3.72E-05	0.609601
GTTGCCC	3.87E-05	0.634797
CGCACCG	4.54E-05	0.743857
CAGGTCA	4.82E-05	0.790382
AGGGGCT	5.15E-05	0.843853
TCGTTTA	5.79E-05	0.948575
TGCTGAC	6.04E-05	0.988867
CCTGGTC	7.17E-05	1.175253
ATATATA	7.25E-05	1.188533
GACTCAT	7.49E-05	1.226862
ACTCCGG	8.22E-05	1.347221
TTGGGCT	9.29E-05	1.521616
GAGGGGT	0.000117	1.912039
GCCCAAG	0.000127	2.076725
TTATGCG	0.000128	2.100279
GCACGCA	0.00014	2.297738
GCTGACA	0.000146	2.395133

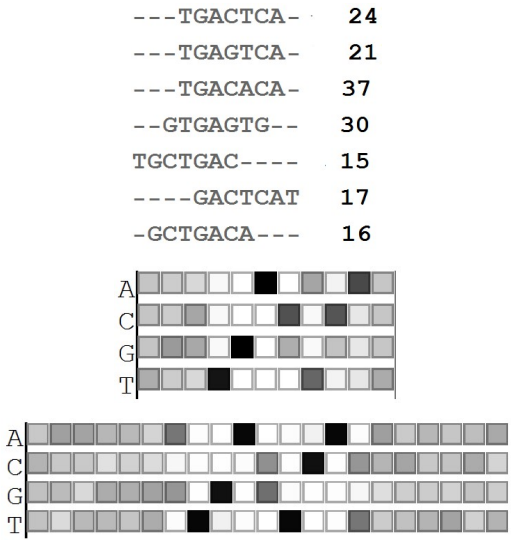


Figure 23. Occam's razor approach to elucidate a common motif in regulatory regions of down-regulated genes, using top heptamers from the enumerative method. Top: alignment of seven heptamers from the top 30. Middle: combined motif obtained by the Occam's razor approach. Bottom: the PPM for Gcn4 from the JASPAR database.

The activator of genes involved with alternative carbon source utilisation Adr1, and the repressor Mig1, were not among the most significantly enriched TFs in the likelihood functions. However, when the strength of the motif match ($Log_{10}(Lm)$) was emphasised, and with the support of motif conservation among *Kluyveromyces* species (*Lc*) in the function $Log_{10}(Lm) \times Lc$, Adr1 was seventh in terms of the enrichment (Addendum 4, Figures 5 and 6). When using only the motif score $Log_{10}(Lm)$, the score dropped. This suggested that the experimental evidence captured in the SGD for at least this motif was incomplete, that that the short motif on its own was insufficient to accumulate a large enough score to end up in the top 14 500 motifs, and that the conservation criterion played a very important role in the likelihood framework. It was also noted that two other motifs in the top ten in the results from the $Log_{10}(Lm) \times Lc$ function shared a resemblance with the Adr1 motif (Figure 24). These were Zms1, the most significantly enriched TF among all conditions (see results below) and also matching to the top heptamer in the up-regulated gene set [see Chapter 4], and significantly enriched Aft2. All three these TFs bind a stretch of four cytosines or guanines.

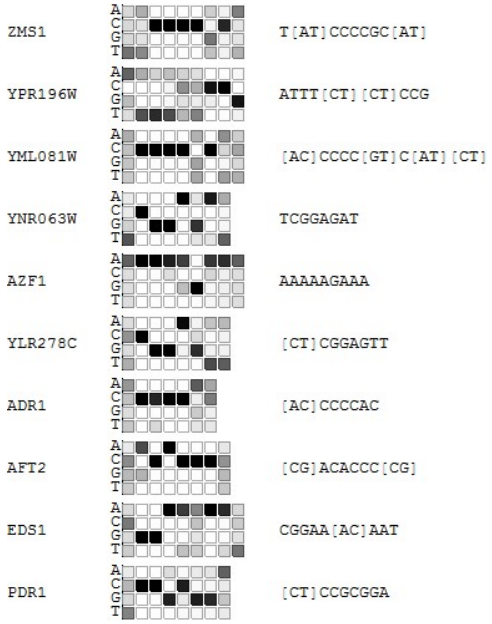


Figure 24. The top ten enriched motifs using the $Log_{10}(Lm)\times Lc$ function with a network constructed from 14 500 motifs. A resemblance was found between Zms1, Adr1 and Aft2.

Combining multiple sources of evidence for optimised regulatory networks

The results from several enrichment tests were summarised in a comparative table of nine separate functions (Table 6). The eight likelihood functions comprised various methods to include the importance of the motif score ($Log_{10}(Lm)$), the conservation of a motif among several *Kluyveromyces* species (*Lc*) and the strength of the experimental evidence for an interaction in SGD (*Li*), applying the evidence strength scoring system (Table 2). The method based only on data from SGD regardless of motif presence was also included (method SGD), as well as the heptamer frequency method (method *I*). For each of the likelihood methods, the maximal enrichment score was selected among those from several networks, where the interactions were included based on whether their motifs were among the top *n* motifs (Figures 20-21 and Addendum 4, Figures 1-12). Notably, each method revealed different regulators that might be implied in the differential response. Discovery of selected TFs is discussed below (see Addendum 4 for others).

Table 6. Multiple sources of evidence used to reveal differentially active transcription factors. For genes only annotated in the UFS-Y2791 strain, the differential expression fold change was included from the read-mapping of RNA-seq data to the UFS-Y2791 draft genome. For all cases where the motif strength (*Lm*) was indicated in the final likelihood, the Log_{10} of *Lm* was used. SGD refers to the regulatory target sets in SGD, where the targets could be mapped to homologs in *K. marxianus* DMKU3-1042 via the name of the gene, and regardless of whether the gene for the regulator was present in *K. marxianus*.

Up-regulated target gene sets based on hypergeometric test

Reg	Lm*Lc*Li	Lm*Lc|| Lm*Lc*Li	Lm+Lc+Li	Lm	Lm*Lc	Lc	Li	SGD	I	gene	FC	motif
ZMS1		4.2(7)			5.7(36)				y			9685
STE12	4.5(59)	4.8(75)	4.9(56)				4.0(64)	3.5(98)		gene3811	1	9637
UME6								4.8(295)	y	UFS(g2782.t1)	1	9656
CRZ1						3.8(56)				gene4817	1	9529
MCM1				3.8(47)					y	gene2718	1	9575
NRG1				3.5(93)						gene2521	1	9591
PUT3								3.3(33)		gene3852	1	9602
OPI1					3.1(6)					UFS(g1678.t1)	1	9593
AZF1				3.0(568)					y	gene3930	1	9521
AFT1	2.6(27)	2.9(30)								gene643	1	9513
SWI4	2.8(39)	2.6(34)								gene5080	1	9645
PHD1	2.7(33)						2.7(33)					9599
YAP7				2.6(48)						UFS(g2174.t1)	1	9663
YNR063W					2.48(54)				y			9676
CIN5	2.4(42)	2.5(40)	2.33(42)							UFS(g3405.t1)	1	9528
RSC30				2.4(30)					y			9619
RGM1						2.4(154)			y			9610
TEA1						2.4(48)				gene3148	1	9649
BUR6								5.9(252)		gene3898	1	NA
SUA7								4.5(401)		gene3433	1	NA
SPT10								2.7(423)		gene3729	1	NA

Down-regulated target gene sets based on hypergeometric test

Reg	Lm*Lc*Li	Lm*Lc|| Lm*Lc*Li	Lm+Lc+Li	Lm	Lm*Lc	Lc	Li	SGD	I	gene	FC	motif
GCN4	16.4(73)	13.8(56)	15.2(56)		2.4(20)		16.2(68)	15.3(169)	y	gene3451	1	9547
GCR2	6.9(33)	8.0(23)	8.2(27)				6.9(33)	7.6(40)		gene2613	0.40	9549
AZF1				6.5(743)	7.0(536)				y	gene3930	1	9521
SFP1	3.4(885)		2.8(527)				5.1(593)	3.6(1306)		gene2869	1	9622
ARG81	4.7(5)	4.6(5)	4.7(5)				4.6(5)		y	gene1497	1	9516
MET32	3.5(96)	3.5(96)	3.6(124)				4.4(118)	2.2(205)		gene4085	1	9578
BAS1	3.6(15)	2.4(10)	2.8(9)				3.4(16)	3.8(54)	y	gene1050	1	9522
RGT1					3.5(33)					gene4011	1	9611
RTG3	2.8(19)	2.4(10)					3.3(22)	3.8(38)		gene4377	1	9620
SUT2			2.9(47)	2.7(114)						UFS(g3260.t1)	3.4	9644
LEU3		2.9(38)	2.9(39)	2.6(10)	2.4(35)					gene4135	1	9568
GCR1	2.8(60)	2.9(57)	2.8(5)				2.8(182)	2.8(414)		gene759	0.21	9548
YLR278C					2.9(65)					UFS(g1445.t1)	1	9674
CUP2				2.8(63)								9531
FKH2								2.8(106)		gene4226	1	9541
CST6	2.7(112)						2.5(128)			gene1355	1	9530
MAC1							2.6(5)			gene2791	1	9570
GLN3	2.6(21)	2.6(21)	2.4(22)				2.6(21)			UFS(g4493.t1)	1	9551
SPT2	2.6(20)						2.6(20)			(UFS)g3039.t1	1	9631
MET4							2.6(23)	2.3(177)		gene4937	1	9579
CBF1					2.5(428)					(UFS)g4946.t1	1	9525
YPR022C		2.4(9)			2.4(9)					UFS)g1761.t1	1	9680
TYE7				2.4(308)						gene3913	1	9653
TUP1								7.2(184)		gene1081	1	NA
SPT3								7.1(266)		gene1396	1	NA
SPT20								6.6(184)		gene1462	1	NA
HFI1								6.1(160)		gene1998	1	NA
SIN4								4(200)		gene3684	1	NA
CDC73								3.1(29)		gene635	1	NA
SPT7								2.5(69)		gene4603	1	NA

Regulators associated with down-regulated target genes

Gcn4

Strong enrichment was observed for Gcn4. All data contributed to the assignment of interactions. Motif strength alone could not elucidate the gene set, but additional conservation criteria allowed discovery of a small target set. The best enrichment was achieved by combining all data as $Log_{10}(Lm)\times Lc\times Li$. Using the SGD set with the requirement for a motif in the top 14 500 (*Li*) decreased the gene set to less than half that of the SGD set, but by inclusion of conservation criteria and motif strength, five additional targets were found. The accuracy of the PPM deserves attention since the target set was substantially smaller than the SGD set, although it is possible that substantial rewiring may have taken place. Gcn4 was also found as the most likely TF for regulation of down-regulated genes using the k-mer networking approach. The Occam's razor motif may provide an improved motif for Gcn4. Its involvement is in line with its known function as a major activator of genes for *de novo* amino acid biosynthesis, which were down-regulated in the xylose medium. The GCN4 gene was constitutively expressed, however, suggesting involvement of post-translation modifications in the differential response.

Gcr2

Strong enrichment was observed for Gcr2. By allowing motifs to be discovered by the functions $Log_{10}(Lm)+Lc+Li$ and $If[Li==0, Log_{10}(Lm)\times Lc, Log_{10}(Lm)\times Lc\times Li]$, which both decouple the final likelihood from the requirement for experimental evidence in SGD, but may take advantage of it, a slight improvement was made to the enrichment score while keeping fewer targets. At the same time, using only the motif strength $Log_{10}(Lm)$, or by requiring motif conservation $Log_{10}(Lm)\times Lc$, resulted in no significant enrichment. When a motif (with an *Lm* score of above 100) was required to allow inclusion of an interaction from SGD (function *Li*), seven of the 40 targets were removed. This result suggested that the SGD does capture a substantial number of the conserved interactions, but that some of these were not conserved between *S. cerevisiae* and *K. marxianus*. Disconnecting the requirement for SGD evidence ($Log_{10}(Lm)+Lc+Li$ and $If[Li==0, Log_{10}(Lm)\times Lc$, else $Log_{10}(Lm)\times Lc\times Li]$) allowed six or ten genes, respectively, to be removed by competition by other high-scoring motifs, but improved the enrichment. Thus, the motif seemed to be well described and the incorrect motifs were removed. The conservation score *Lc* apparently did not contribute much to the scoring. In addition, the Gcr2 gene was 2.5-fold down-regulated, supporting its role along with Gcr1 as positive regulators of the down-regulated glycolytic genes.

Arg81

Enrichment of Arg81 using the SGD target set was insignificant. The Arg81 motif is similar to that of Gcn4 and out of the five targets for Arg81, three were shared by Gcn4 (see below). Enrichment of Arg81 may thus be partly explained by the differential activity of Gcn4.

Gcr1

Gcr1 showed significant enrichment in both the SGD target set as well as in criteria that combined all sources of data. The positive role of the conservation criterion was evident. When a motif was required in addition to SGD data, less than half of the sites were found, and with a similar enrichment score. This result suggested that either the motif was sub-optimal and possibly too conservative, or that many of the targets in the SGD set resulted from secondary effects.

Tup1

Strong enrichment was found for Tup1. Only SGD data could be used, since Tup1 does not have an associated DNA binding motif. Tup1 is recruited as a suppressor of gene expression by DNA binding proteins, including Mig1 [Treitel et al. 1995]. Occurrence in the down-regulated set was consistent with its function as a suppressor of gene regulation.

Regulators associated with up-regulated target genes

ZMS1

The 36 targets of Zms1 were discovered by motif strength and conservation among the *Kluyveromyces* species. Conservation was required, but incorporation of SGD data was detrimental to the enrichment score, suggesting that these were unique to *K. marxianus,* or that the targets in *S. cerevisiae* were not correctly documented. Also, it mapped to the most over-represented heptamer, implying that the enrichment might have been the result of a different zinc finger type of TF, such as Adr1 or Mig1.

Ste12

The Ste12 gene set was enriched using the SGD method. Requirement for the motifs (*Li*) made an improvement, and motif strength was more important than sequence conservation. Yet, motif strength on its own failed to discover the enriched set. The results suggested that the Ste12 gene set has been conserved, but these regions could not be aligned well by the genome aligner.

Phd1

Phd1 was shown to be enriched only when motifs were required (*Li*) or by using all data (Log_{10}(*Lm*)×*Lc*×*Li*), and not by the SGD data alone. Increased emphasis on the motif score did not allow significance, since the motif was not very precise, and did not compete well with other motifs in this method. Phd1 activates genes for pseudohyphal growth, and some of these genes were up-regulated in the xylose medium. No PHD1 gene was found, however.

Creating an optimised enrichment network

The optimised target gene sets for each of the transcription factors identified were extracted and a genome-scale gene regulatory network was constructed (Figure 25). Only TFs were included, and only those that were associated with DNA-based evidence in *K. marxianus* (functions containing Log_{10}(*Lm*), *Lc* and *Li*), regardless of the SGD method. The network comprises 3 545 interactions involving 38 TFs and 2145 protein-encoding gene targets. Some TFs have overlapping target sets. Figure 26 shows the number of targets in common between each pair of TFs and the number of targets in each gene set.

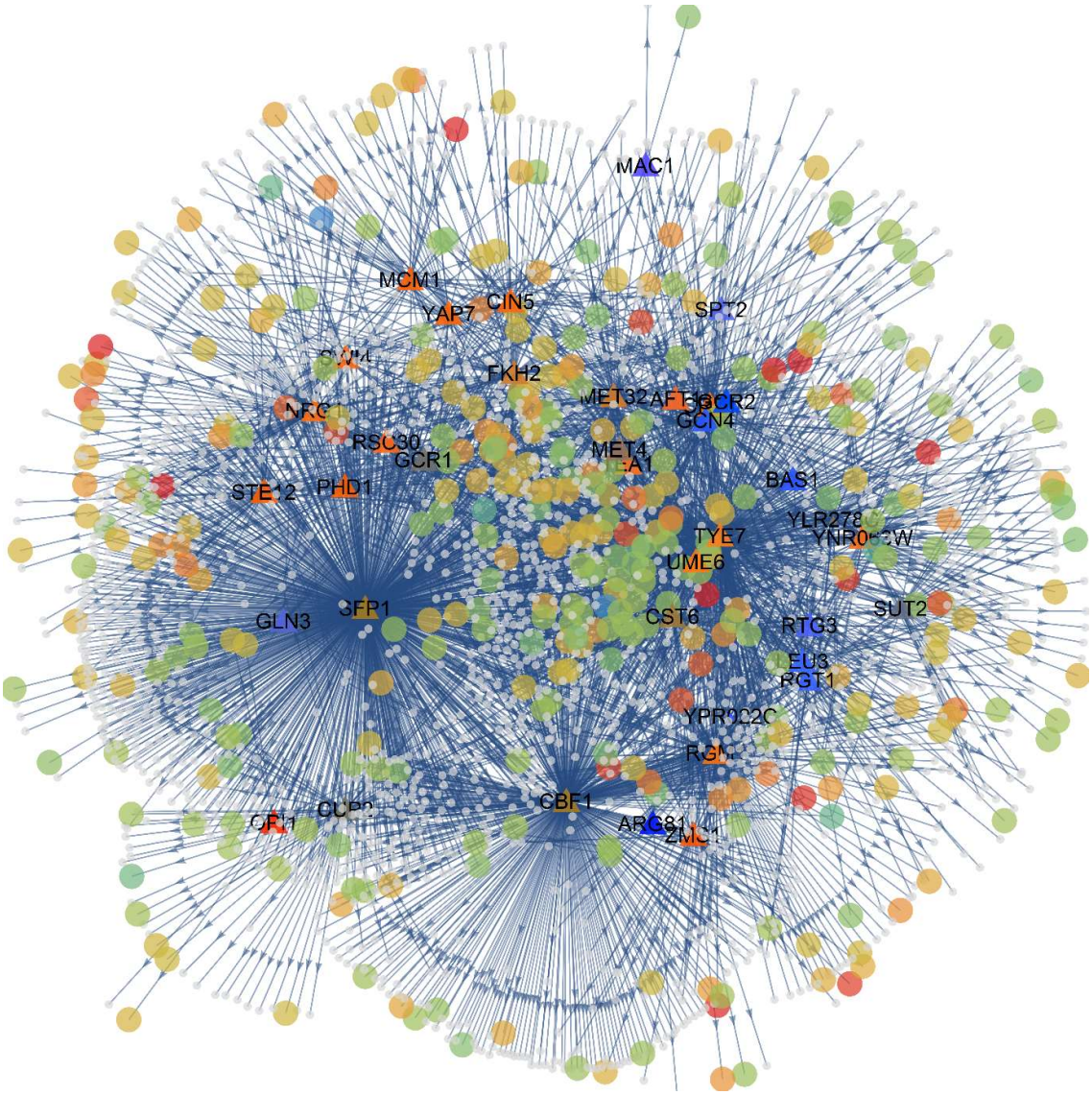


Figure 25. The network optimised for the best enrichment statistics. TFs are indicated in triangles and targets in circles. The calculation of the colours of TFs was based on their enrichment statistics using the Z-score approach and the 'clarity' scheme in *Reactomica*: warm colours (red) indicate a target gene set that was generally up-regulated; cold colours (blue) indicate a target gene set that was generally down-regulated; murkiness indicates a target gene set that was approximately equally up and down-regulated. For the targets (circles), large circles indicate differential expression, with warm colours indicating up-regulation and cold colours indicating down-regulation. Small, light grey circles indicate targets that were constitutively expressed.

	ZMS1	STE12	CRZ1	MCM1	NRG1	OPI1	AFT1	SWI4	PHD1	YAP7	YNR063W	CIN5	RSC30	RGM1	TEA1	GCN4	GCR2	SFP1	ARG81	MET32	BAS1	RGT1	RTG3	SUT2	LEU3	GCR1	YLR278C	CUP2	CST6	MAC1	GLN3	SPT2	MET4	CBF1	YPR022C	TYE7	UME6	FKH2
ZMS1	36	0	7	0	4	0	0	0	1	0	5	0	0	7	4	0	0	5	0	2	1	0	0	1	0	1	2	1	1	0	1	0	1	8	2	6	7	1
STE12	0	56	1	5	0	0	3	7	6	1	0	1	0	2	1	0	0	23	0	4	0	1	0	0	0	1	0	1	6	0	0	0	1	2	0	4	6	8
CRZ1	7	1	56	1	5	0	3	0	1	4	5	2	2	14	8	0	1	7	0	5	1	1	2	3	7	4	3	0	4	1	1	1	1	12	0	13	7	3
MCM1	0	5	1	47	3	0	0	5	4	0	3	1	0	1	1	1	1	10	0	1	0	0	0	0	0	1	0	1	3	0	1	0	0	1	0	9	4	2
NRG1	4	0	5	3	93	0	0	0	1	2	4	1	0	9	2	3	1	10	0	7	1	0	0	0	0	2	0	0	0	0	5	1	1	13	0	14	6	6
OPI1	0	0	0	0	0	6	0	0	0	0	0	0	0	2	2	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	0	0	1	0
AFT1	0	3	3	0	0	0	30	0	1	1	0	3	0	4	1	0	2	10	0	6	1	2	3	0	3	3	1	0	4	1	0	1	3	5	0	5	9	7
SWI4	0	7	0	5	0	0	0	39	8	1	0	1	1	0	2	5	0	11	0	1	0	0	0	0	1	2	0	3	4	1	0	1	1	2	0	4	6	4
PHD1	1	6	1	4	1	0	1	8	33	0	2	5	0	0	1	4	0	9	0	4	0	1	1	0	0	0	0	1	2	0	1	0	1	3	0	0	10	4
YAP7	0	1	4	0	2	0	1	1	0	48	0	0	3	1	2	1	1	8	0	7	0	0	0	2	2	3	0	1	2	0	1	0	4	9	0	12	5	3
YNR063W	5	0	5	3	4	0	0	0	2	0	54	1	0	5	10	2	2	9	0	2	1	5	1	3	0	2	8	0	1	0	1	1	1	7	0	5	4	1
CIN5	0	1	2	1	1	0	3	1	5	0	1	40	0	1	3	2	1	11	0	3	0	0	0	0	0	2	3	0	1	1	1	3	0	7	0	3	8	3
RSC30	0	0	2	0	0	0	0	1	0	3	0	0	30	1	0	0	0	1	0	2	0	4	0	0	2	0	4	1	2	0	0	0	0	6	0	5	7	1
RGM1	7	2	14	1	9	2	4	0	0	1	5	1	1	154	8	1	3	32	0	7	1	2	3	1	3	3	1	2	8	0	3	1	0	28	2	23	11	8
TEA1	4	1	8	1	2	2	1	2	1	2	10	3	0	8	48	0	3	10	0	3	3	2	0	6	0	5	6	3	1	1	1	2	0	18	2	12	5	3
GCN4	0	0	0	1	3	0	0	5	4	1	2	2	0	1	0	73	0	22	3	9	2	2	3	0	4	2	3	0	6	1	6	1	3	13	1	14	16	4
GCR2	0	0	1	1	1	0	2	0	0	1	2	1	0	3	3	0	27	11	0	4	1	1	3	2	2	7	1	0	3	1	0	3	0	6	1	6	4	5
SFP1	5	23	7	10	10	1	10	11	9	8	9	11	1	32	10	22	11	885	2	43	9	5	7	9	7	22	15	12	54	2	5	9	8	92	1	69	89	24
ARG81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	2	5	0	0	1	1	0	0	0	0	0	0	0	1	0	0	2	0	0	2	1
MET32	2	4	5	1	7	0	6	1	4	7	2	3	2	7	3	9	4	43	0	118	2	3	6	1	3	8	3	2	5	1	3	2	23	19	3	15	24	11
BAS1	1	0	1	0	1	0	1	0	0	0	1	0	0	1	3	2	1	9	0	2	15	0	0	1	1	1	1	0	3	0	0	1	1	2	1	2	1	1
RGT1	0	1	1	0	0	0	2	0	1	0	5	0	4	2	2	2	1	5	1	3	0	33	0	0	5	0	3	0	1	0	0	0	1	6	2	6	6	2
RTG3	0	0	2	0	0	0	3	0	1	0	1	0	0	3	0	3	3	7	1	6	0	0	22	0	1	0	2	0	3	0	0	0	2	2	0	2	7	4
SUT2	1	0	3	0	0	0	0	0	0	2	3	0	0	1	6	0	2	9	0	1	1	0	0	47	1	3	11	3	2	0	0	3	0	5	0	4	4	1
LEU3	0	0	7	0	0	0	3	1	0	2	0	0	2	3	0	4	2	7	0	3	1	5	1	1	38	0	6	3	0	0	0	0	0	8	0	11	4	2
GCR1	1	1	4	1	2	2	3	2	0	3	2	2	0	3	5	2	7	22	0	8	1	0	0	3	0	57	3	1	11	0	0	1	2	10	1	14	6	3
YLR278C	2	0	3	0	0	0	1	0	0	0	8	3	4	1	6	3	1	15	0	3	1	3	2	11	6	3	65	3	3	1	0	3	0	13	0	8	11	3
CUP2	1	1	0	1	0	0	0	3	1	1	0	0	1	2	3	0	0	12	0	2	0	0	0	3	3	1	3	63	2	0	0	0	1	5	0	6	4	1
CST6	1	6	4	3	0	0	4	4	2	2	1	1	2	8	1	6	3	54	0	5	3	1	3	2	0	11	3	2	112	0	0	0	2	16	1	11	20	9
MAC1	0	0	1	0	0	0	1	1	0	0	0	1	0	0	1	1	1	2	0	1	0	0	0	0	0	0	1	0	0	5	0	3	0	0	0	0	2	1
GLN3	1	0	1	1	5	0	0	0	1	1	1	1	0	3	1	6	0	5	1	3	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	1	1	2
SPT2	0	0	1	0	1	0	1	1	0	0	1	3	0	1	2	1	3	9	0	2	1	0	0	3	0	1	3	0	0	3	0	20	0	3	0	2	5	1
MET4	1	1	1	0	1	0	3	1	1	4	1	0	0	0	0	3	0	8	0	23	1	1	2	0	0	2	0	1	2	0	0	0	23	4	0	4	4	3
CBF1	8	2	12	1	13	2	5	2	3	9	7	7	6	28	18	13	6	92	2	19	2	6	2	5	8	10	13	5	16	0	0	3	4	428	9	107	35	19
YPR022C	2	0	0	0	0	0	0	0	0	0	0	0	0	2	2	1	1	1	0	3	1	2	0	0	0	1	0	0	1	0	0	0	0	9	9	5	1	0
TYE7	6	4	13	9	14	0	5	4	0	12	5	3	5	23	12	14	6	69	0	15	2	6	2	4	11	14	8	6	11	0	1	2	4	107	5	308	31	11
UME6	7	6	7	4	6	1	9	6	10	5	4	8	7	11	5	16	4	89	2	24	1	6	7	4	4	6	11	4	20	2	1	5	4	35	1	31	295	26
FKH2	1	8	3	2	6	0	7	4	4	3	1	3	1	8	3	4	5	24	1	11	1	2	4	1	2	3	3	1	9	1	2	1	3	19	0	11	26	106

Figure 26. Number of target genes in common between transcription factors. The number of targets in each gene set is found on the diagonal.

Subnets for various TFs are shown below (Figure 27-39). The down-regulated Gcn4 targets are mostly genes encoding for enzymes in amino acid biosynthesis (Figure 27). Differentially expressed targets for Gcr2 and Arg81 were exclusively down-regulated, while the majority of differentially expressed targets for Gcn4, Bas1, Rtg3 and Gcr1 were down-regulated approximately two to four-fold (Figures 27-32). Many of these targets are involved in amino acid or nucleotide biosynthesis, one-carbon metabolism and glycolysis. This pattern supports the observation of a specific growth rate in a defined xylose medium that was approximately 40% of that in the glucose medium [Chapter 3, Schabort et al. 2016]. A substantial fraction of the Met32 targets were up-regulated and it is less clear whether Met32 was associated with down-regulated or up-regulated targets (Figure 33). Among the generally up-regulated target gene sets, Zms1, Ste12 and Mcm1 had differentially expressed targets which were almost exclusively up-regulated (Figures 34-36). The differentially expressed targets of activator Phd1 were exclusively up-regulated (Figure 37). For Nrg1 and Aft1, a more mixed pattern of differential expression was observed (Figures 38 and 39).

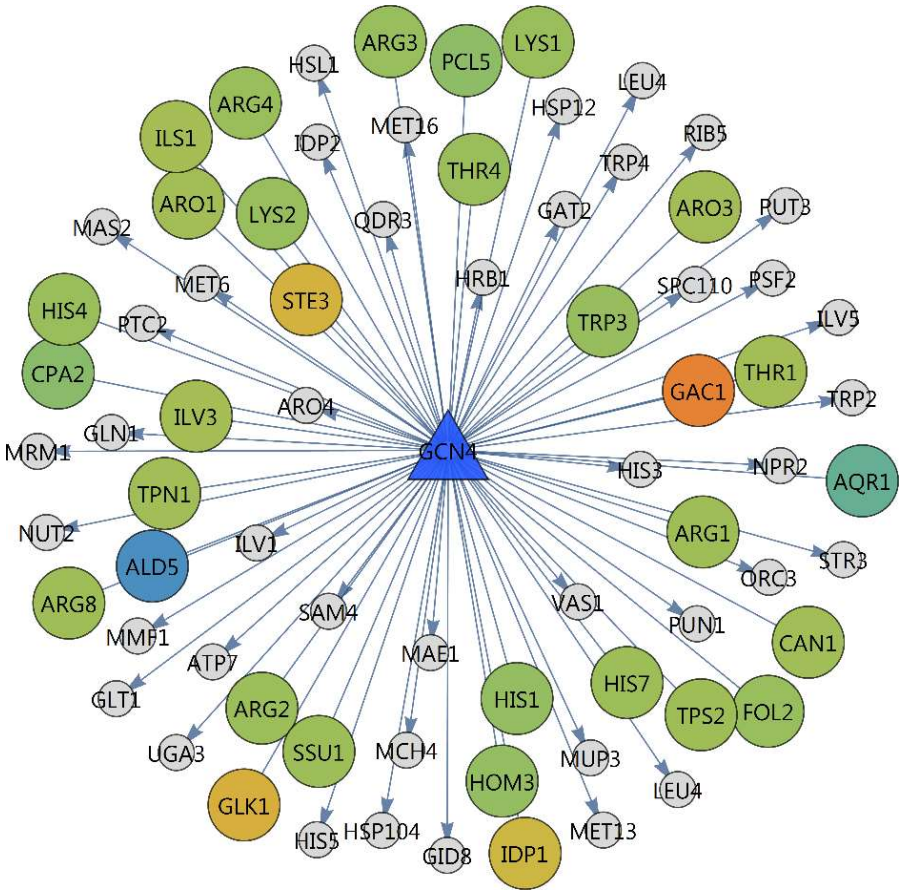


Figure 27. TF-target network with Gcn4 as TF. For the TF (triangle), warm colours (red) indicate a target gene set that was generally up-regulated; cold colours (blue) indicate a target gene set that was generally down-regulated; murkiness indicates a target gene set that was approximately equally up and down-regulated. For the targets (circles), large circles indicate differential expression, with warm colours indicating up-regulation and cold colours indicating down-regulation. Small, light grey circles indicate targets that were constitutively expressed. Differentially expressed target genes were mostly down-regulated.

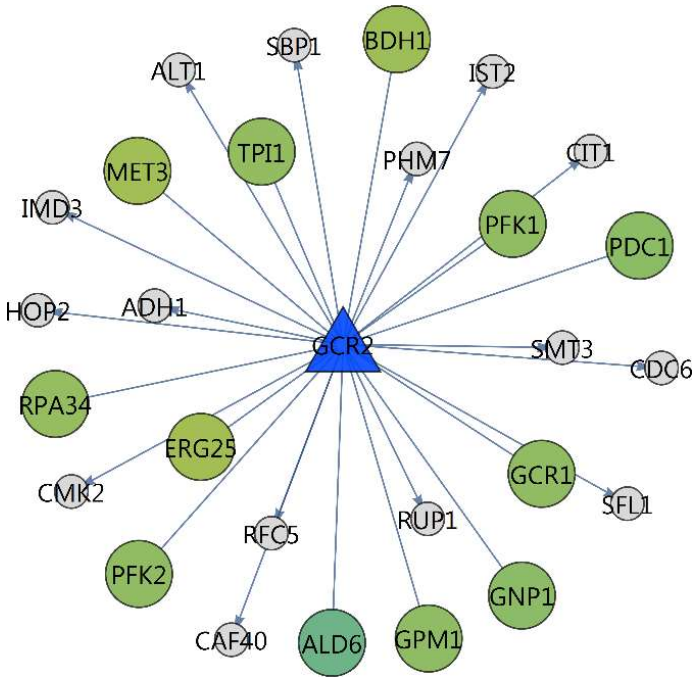


Figure 28. TF-target network with Gcr2 as TF. Differentially expressed target genes were mostly down-regulated. See Figure 27 legend for the meaning of colours and shapes.

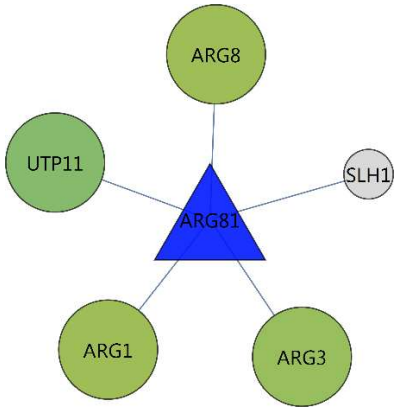


Figure 29. The TF-target network with Arg81 as TF. Differentially expressed target genes were mostly down-regulated. See Figure 27 legend for the meaning of colours and shapes.

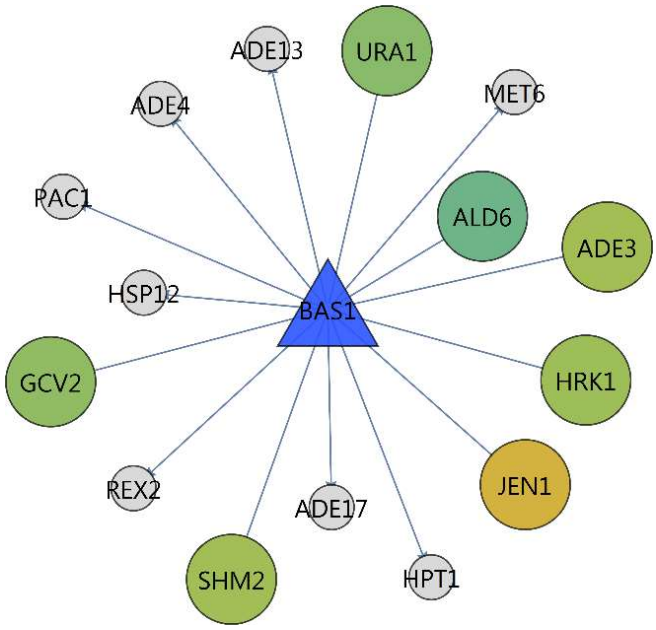


Figure 30. The TF-target network with Bas1 as TF. Differentially expressed target genes were mostly down-regulated. See Figure 27 legend for the meaning of colours and shapes.

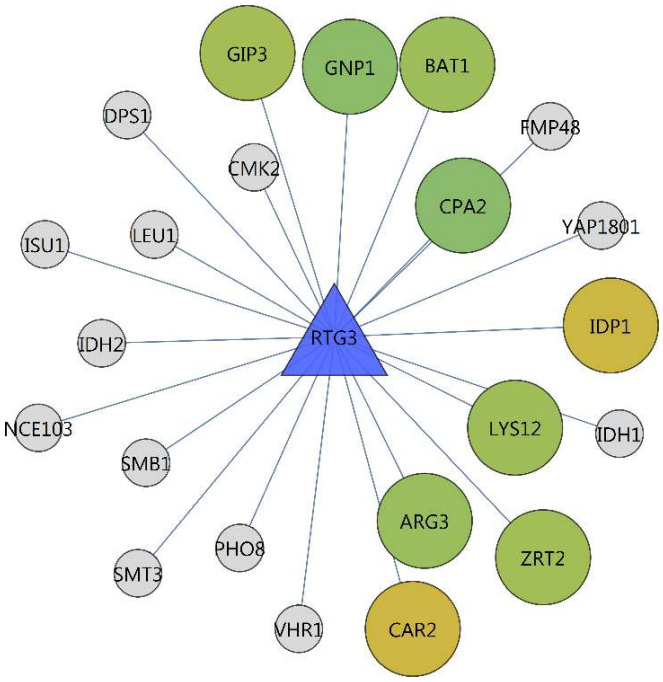


Figure 31. The TF-target network with Rtg3 as TF. Differentially expressed target genes were mostly down-regulated. See Figure 27 legend for the meaning of colours and shapes.

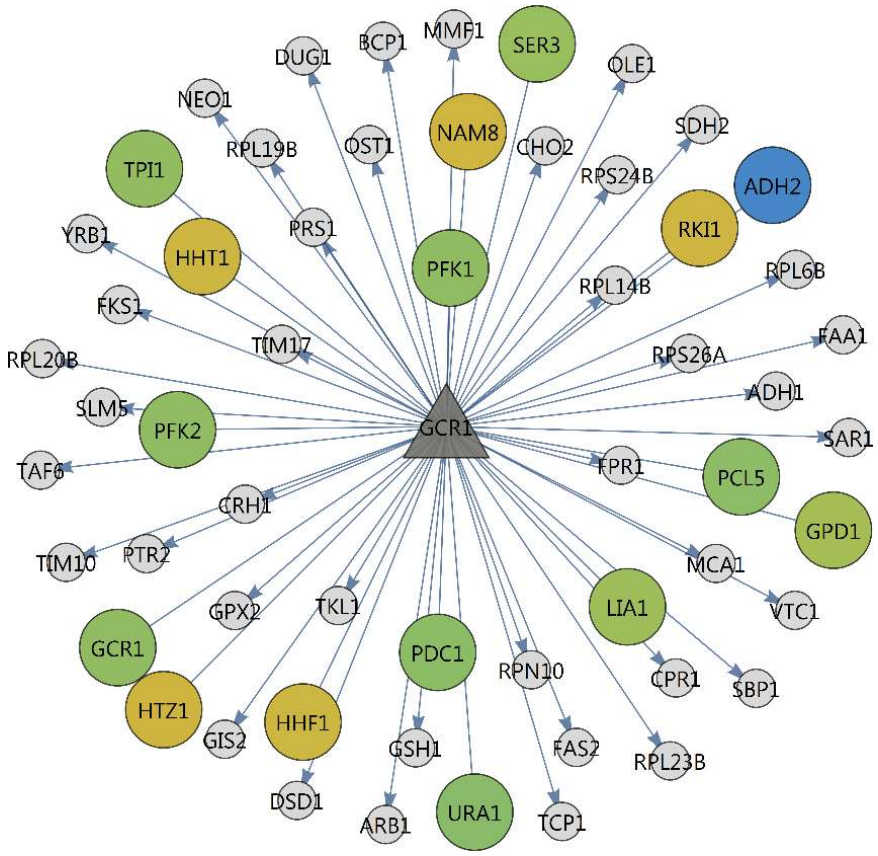


Figure 32. The TF-target network with Gcr1 as TF. Differentially expressed target genes consisted of both up and down-regulated genes, with the majority being down-regulated. See Figure 27 legend for the meaning of colours and shapes.

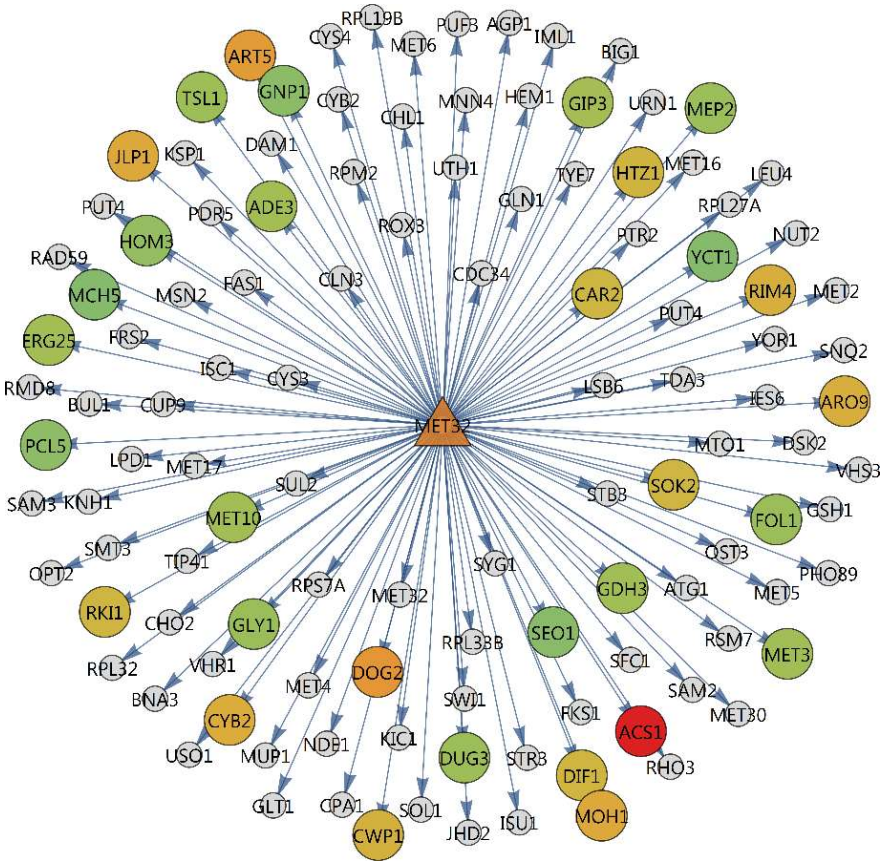


Figure 33. The TF-target network with Met32 as TF. Differentially expressed target genes consisted of both up and down-regulated genes. See Figure 27 legend for the meaning of colours and shapes.

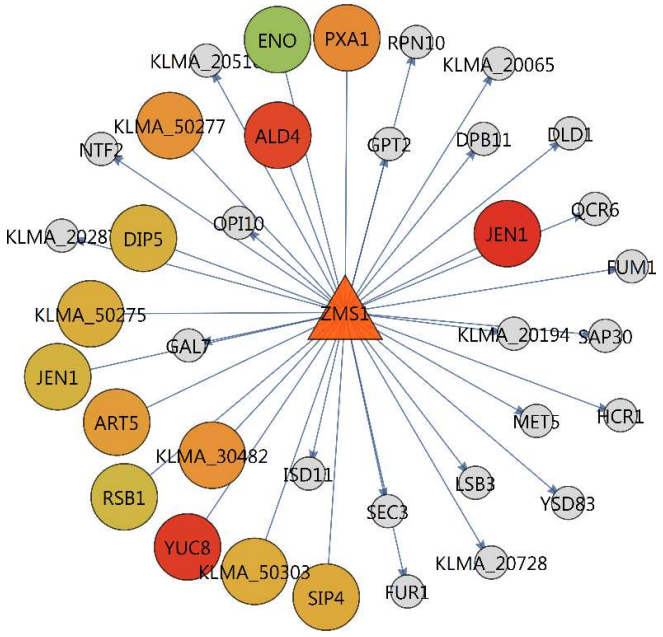


Figure 34. The TF-target network with Zms1 as TF. Differentially expressed target genes consisted of almost exclusively up-regulated genes. See Figure 27 legend for the meaning of colours and shapes.

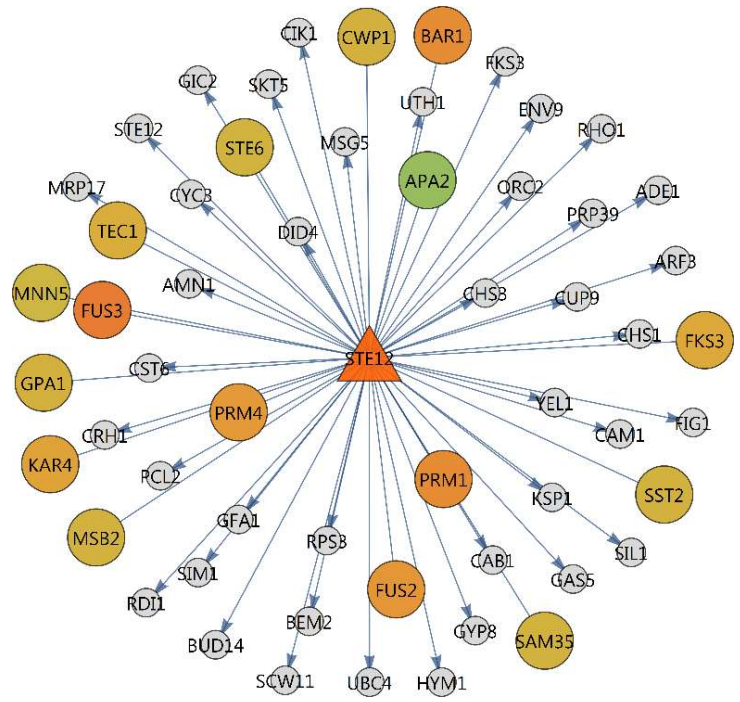


Figure 35. The TF-target network with Ste12 as TF. Differentially expressed target genes consisted of almost exclusively up-regulated genes. See Figure 27 legend for the meaning of colours and shapes.

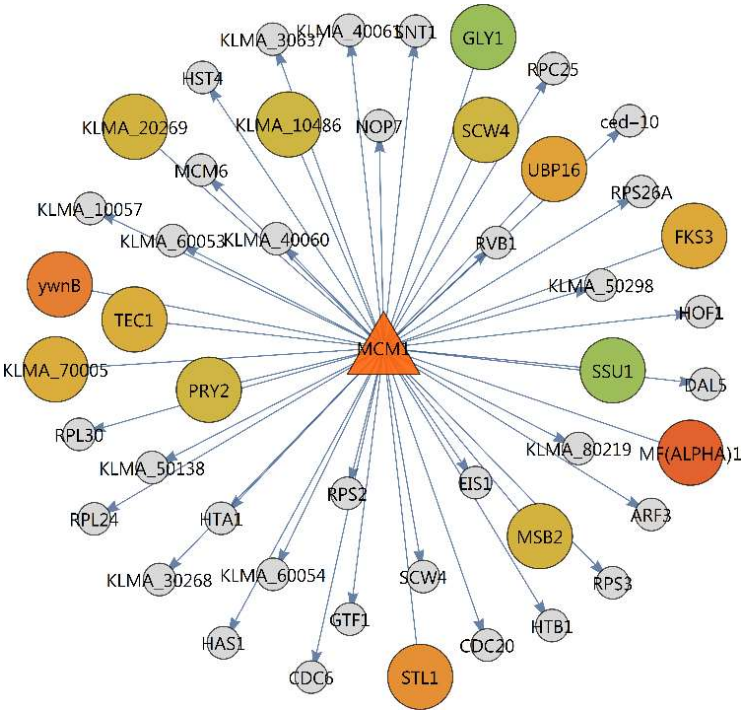


Figure 36. The TF-target network with Mcm1 as TF. Differentially expressed target genes consisted of almost exclusively up-regulated genes. See Figure 27 legend for the meaning of colours and shapes.

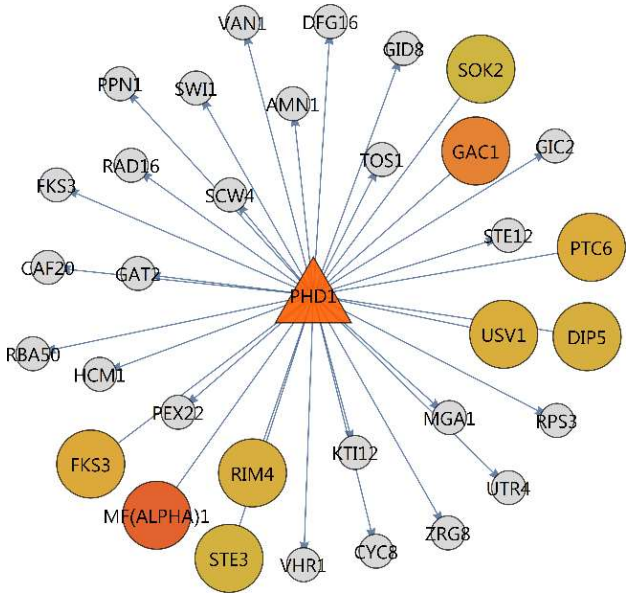


Figure 37. The TF-target network with Phd1 as TF. Differentially expressed target genes consisted of exclusively up-regulated genes. See Figure 27 legend for the meaning of colours and shapes.

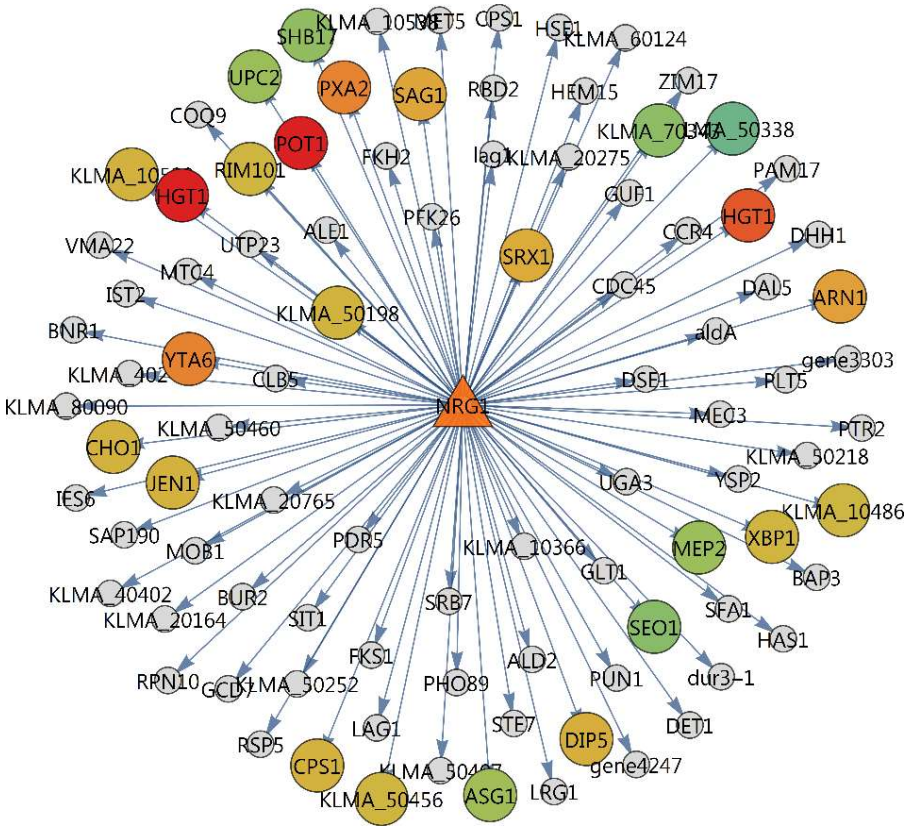


Figure 38. The TF-target network with Nrg1 as TF. Differentially expressed target genes consisted of equal proportions of up and down-regulated genes. See Figure 27 legend for the meaning of colours and shapes.

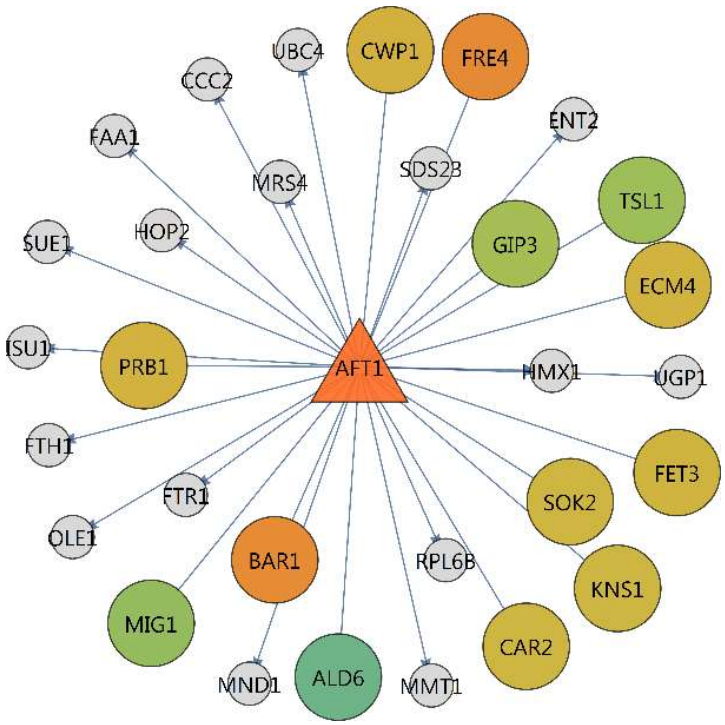


Figure 39. The TF-target network with Aft1 as TF. Differentially expressed target genes consisted of both up and down-regulated genes. See Figure 27 legend for the meaning of colours and shapes.

Mapping transcription factors to pathways

To functionally organise the gene regulatory network, target genes of the down-regulated gene sets were mapped to pathways. Figure 40 shows the TF-enzyme-pathway network. The regulators associated with down-regulated targets (Gcn4, Arg81, Bas1, Gcr2, Gcr1) were clearly associated with pathways that were, as a whole, significantly enriched and down-regulated in the xylose medium. Met32 seems to have a more mixed character. The up-regulated TF-target sets were omitted from the analysis for clarity.

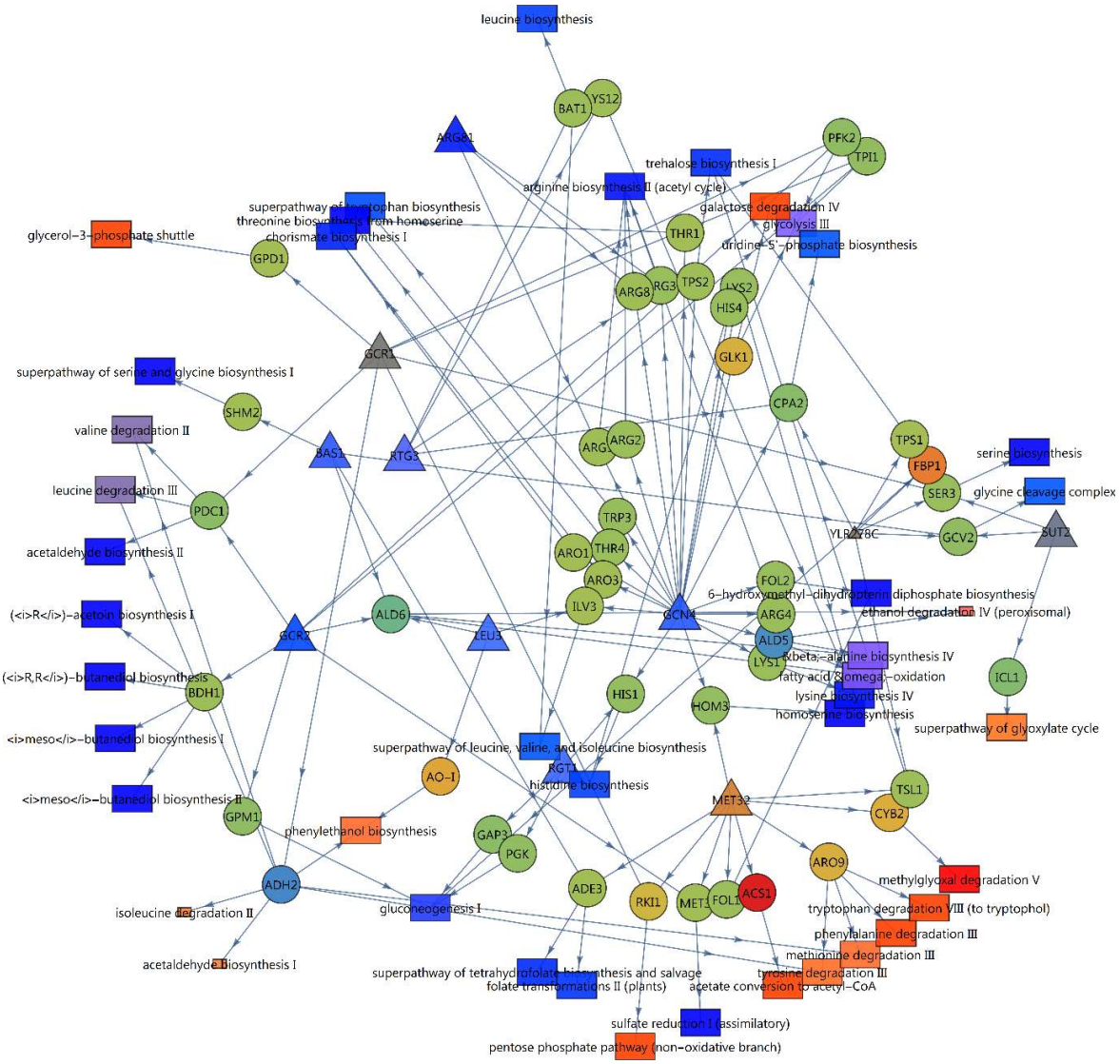


Figure 40. The transcription factor-enzyme-pathway network. The calculation of the colours of TFs and pathways was based on their enrichment statistics using the Z-score approach and the 'clarity' scheme in *Reactomica*: warm colours indicate gene set that was generally up-regulated; cold colours indicate a gene set that was generally down-regulated; murkiness indicates a gene set that was approximately equally up and down-regulated. Circles indicate genes encoding enzymes involved in the pathways to which they map. Warm colours indicate up-regulation and cold colours indicate down-regulation.

TFs regulating other TFs

To explore the higher gene regulatory programme among TFs, a TF-TF network was constructed (Figure 41). Among the enriched TFs, only Gcr1 and Gcr2 were themselves differentially expressed at transcript level.

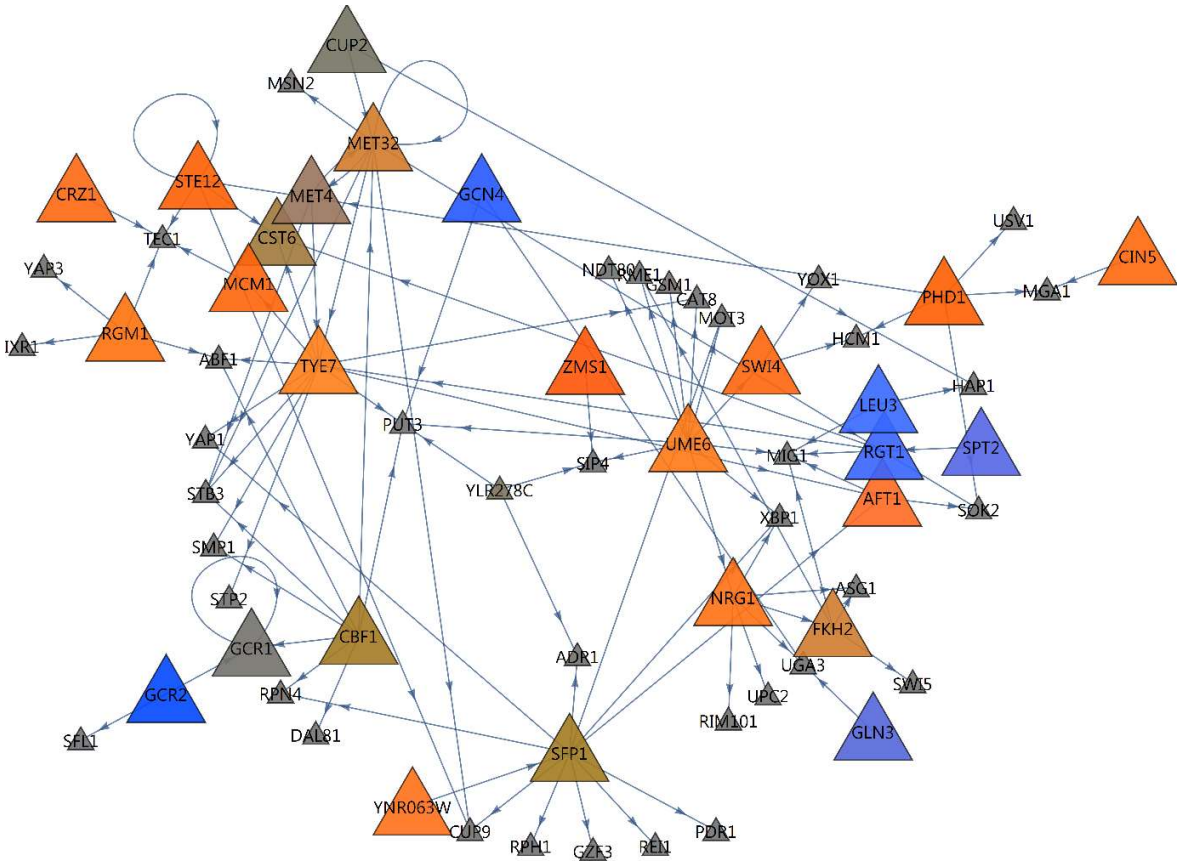


Figure 41. TF-TF network. Colours indicate enrichment statistics from the target set of each TF, using the 'clarity' scheme in *Reactomica*. Small grey triangles represent non-enriched TFs. Loops indicate auto-regulation by a TF.

Discussion

Construction of gene regulatory networks is greatly hampered by the fact that DNA binding motifs of transcription factors are very short and not very precise, resulting in many false positive predictions from motif scans alone. Such predictive work has to be complemented with experimental data on TF-target interactions. Multiple high-throughput methods such as transcriptome profiling have revealed many interactions in the model species *S. cerevisiae*, where some of these may have resulted from secondary transcriptional effects. Other methods such as ChIP provide direct physical evidence of TF-target interactions. In this study a framework to incorporate these multiple sources of evidence was demonstrated, in which the benefits of each method could be combined in the construction of a likelihood based gene regulatory network. This framework is flexible in incorporating the multiple sources of evidence using an evidence code based scoring system. As the focus is increasingly on non-model yeast species such as *K. marxianus* for the production of biofuels and recombinant proteins, a major challenge is to construct such networks for these non-model species to predict outcomes of

genetic engineering strategies. Experiments such as ChIP is labour intensive and costly, making it an impractical approach for the rapid construction of gene regulatory networks when there is not a large research community working on the same species. The likelihood framework demonstrated here effectively bridges the gap between datasets that could rapidly be generated for a non-model species (a complete or draft genome and RNA-seq) and the wealth of interaction data on a model species such as *S. cerevisiae*. Seven different likelihood based functions were used to construct networks, and these were compared to one which was purely based on the interaction data from SGD. For each network, the enrichment statistical methods using the Z-scores or the hypergeometric distribution were calculated for the target gene set of each regulator. The hypergeometric method was found to be much simpler in implementation, especially in distinguishing between generally up or down-regulated target sets.

Finding an increased enrichment statistic for a few, sufficiently large gene sets could be considered a sign of capturing more true interactions and fewer false interactions. Also, finding a clearer pattern of either up or down-regulation in a target gene set could similarly be seen as an improvement, since a TF would generally do either of the two activities, although some may function in both mechanisms. By choosing optimally enriched target sets, it was found that each of the seven likelihood functions had its merits, depending on the specific TF. A single scoring function may not be a suitable treatment for all TFs. The various scenarios found included (*a*) a target gene set that was transcriptionally rewired compared to the model species, (*b*) that the motif was sub-optimal and required improvement (relaxation by inclusion of pseudocounts or requiring more pattern specificity), (*c*) that the enriched motif might rather be a proxy for a different TF that has more experimental evidence supporting its activity, and (*d*) that the dataset in SGD might contain interactions originating from secondary effects. This may provide a new avenue of research in that these scenarios suggest the foci for further experimentation, or even automated machine learning approaches to improvement of the gene regulatory network. It was finally opted to construct the final network on the best enrichment statistics for a given TF among all functions that required a DNA motif to be present in the species of interest.

In addition, the study revealed key insights into the mechanistic basis of the differential genetic response of *K. marxianus* to glucose or xylose in a defined medium. It was revealed that primarily Gcn4, but also Arg81, Bas1 and Rtg3 were strongly involved in *de novo* amino acid synthesis, corresponding to the observation that their gene targets were mostly down-regulated. Although it is known that Bas1 has a similar DNA binding specificity to that of Gcn4 [Springer et al. 1996], only two

of the 15 targets in the network were common with the 73 targets of Gcn4. Thus, Bas1 activity cannot be explained by Gcn4 differential activity. Similarly, Gcr2 and Gcr1, for which the targets were also down-regulated, are known activators of glycolysis [Holland et al. 1987, Uemura et al. 1992]. Tup1 is also in the list of effectors with down-regulated targets, but is not described by a DNA binding motif like the transcription factors. Tup1 is a protein involved with repression of gene expression and recruited by transcription factors such as Mig1 [Treitel et al. 1995].

In the target sets that were mostly up-regulated, the master transcriptional regulator of pseudohyphal growth, Phd1, was revealed as differentially active (Figure 37). Ste12 clearly was also a regulator of up-regulated targets (Figure 35). These are both activators of the pseudohyphal growth response [Broach 2012] which was detected by gene set enrichment using Gene Ontology [Chapter 3, Schabort et al. 2016]. Mating factor alpha -1 (MF(alpha)1) may have been up-regulated by increased activity of Phd1 or by Mcm1 (Figure 37). Mcm1 is known to be involved with regulation of mating type specific genes [Elble and Tye 1991] and PHD1 is a master regulator in pseudohyphal growth activation [Gimeno and Fink 1994]. Phd1 also targets Ste12 transcription. Tec1, which is under the control of Ste12 in the network and also an activator of pseudohyphal growth genes [Broach 2012], was up-regulated 4.3-fold, whereas Ste12 was constitutively expressed. Tec1 cooperates with Ste12 to make the activity specific to pseudohyphal growth, distinguishing it from the role in mating [Roberts and Fink 1994, Madhani and Fink 1997]. Two kinases phosphorylate and activate Ste12, namely Fus3 [Elion et al. 1993] and Kss1 [Bardwell et al. 1998]. Kss1 activity is associated with pseudohyphal growth, whereas Fus3 is associated with conjugative/polarised growth during mating. Both these kinases form part of the pheromone signalling cascade involving Ste2, Ste3, Gpa1, Ste18, Ste5, Ste20, Ste11 and Ste7, the latter of which phosphorylates Fus3 and Kss1. The mating pheromone binds to Ste2 (on a-cells) or Ste3 (on α-cells) on the cell surface, sending the signal to the kinase cascade. Ste2 was 12.6-fold up-regulated, Ste3 3.1-fold, and the G-protein alpha subunit Gpa1 3.3-fold. In addition, the pheromone MF(alpha)1 was up-regulated 49-fold. Evidently, the absence of glucose repression made the pheromone signalling system both more active and responsive. Up-regulation of Tec1 suggested that the system promoted pseudohyphal growth, yet the 25-fold up-regulation of Fus3 suggested suppression of pseudohyphal growth, since Fus3 phosphorylates Tec1, leading to its ubiquitin-dependent degradation [Bao et al. 2004].

Thus, Phd1 and Mcm1 both have potential to be involved with increased pseudohyphal growth, complementary to the pheromone system. The kinases that regulate the activity of Phd1, Ste12 and

Mcm1 may play a significant role in morphology which is not immediately evident from the RNA-seq data analysis performed thus far.

A completely separate and complementary approach to the elucidation of differentially active transcription factors was also taken. The enumerative method of heptamer counting is unbiased to evidence of TF-target interactions in the model species and is based purely on the presence of over-represented heptamers in a set of regulatory regions (for up or down regulated gene sets) compared to the background (for all genes). While the enumerative method is well established, its fundamental limitation lies in the short nature of the heptamers revealed, which makes it difficult to reveal which TF was responsible for the differential activity. The Occam's razor approach that was developed in Chapter 4 was shown here to reveal the TF which was also found to be the most significantly enriched in the likelihood method, which was Gcn4. Notably, the Occam's razor approach to the enumerative method also revealed the activity of some zinc finger TF in the up-regulated gene sets, which most likely was Adr1 or Mig1, or a combination of the two, but these were not revealed by the likelihood method as among the most significantly enriched TFs. It is possible that the significantly enriched Zms1, which also is a zinc finger protein binding to a stretch of guanines or cytosines, may be a proxy for Adr1 or Mig1 and that any or both of these TFs may have obtained a slightly different binding specificity compared to the *S. cerevisiae* Adr1 and Mig1. The Occam's razor approach did, however, reveal a striking similarity to the Adr1 site. It was also noted that the Adr1 and Mig1 target sets in SGD appeared to miss some of the true targets of these important TFs, compared to studies such as done by Young et al. [2003] which revealed that Adr1 targeted most of the peroxisomal metabolic genes. This could be the reason why neither of these were enriched when using the SGD target sets. Additional support for their activity was found in that Adr1 was 37-fold up-regulated, which was the most significant among all TFs, while Mig1 was 4-fold down-regulated, consistent with their roles as activator and repressor of transcription, respectively. Notably, it was found that the enrichment score of Adr1 was substantially higher when using only motif strength and motif conservation among *Kluyveromyces* species, bringing Adr1 into the seventh position in terms of enrichment score. When using only the motif score $\text{Log}_{10}(Lm)$, the score dropped. This observation suggested that the experimental evidence captured in the SGD for at least this motif was incomplete, that that the short motif on its own was insufficient to accumulate a large enough score to end up in the top 14 500 motifs, and that the conservation criterion played a very important role in the likelihood framework. It was also noted that two other motifs in the top ten in the results from the $\text{Log}_{10}(Lm)\times Lc$ function shared a resemblance with the Adr1 motif (Figure 24). These were Zms1 and Aft2. Zms1 also matched

to the top heptamer in the up-regulated gene set (see Chapter 4). All three these TFs bind a stretch of four cytosines or guanines.

Among the enriched TFs in the optimised network, only Gcr1 and Gcr2 were differentially expressed. This suggested that apart from Gcr1 and Gcr2, and also likely Mig1 and Adr1 which could not be detected as enriched but were differentially expressed, the mode by which TFs generally were regulated was by post-translational modification and not by gene regulation. The network structure of the TF-TF network indicates Gcr1, Ste12 and Met32 as having auto-regulatory wiring. Apart from Phd1 activating Ste12 and Gcr2 activating Gcr1, not much was evident from the higher regulatory organisation. It is evident that post-translational modifications, such as the master regulator kinases, must provide the signals from the environment to this set of TFs to initiate the differential response.

Conclusions

This work presents the first genome-wide gene regulatory network for the yeast *K. marxianus*. The likelihood framework was shown to be flexible in incorporating multiple sources of evidence, bridging the gap between what can be generated rapidly for a non-model species (a genome or draft genome and RNA-seq), and the wealth of data on regulatory interactions in a model species. The analysis revealed that Gcn4, along with Arg81, Bas1 and Rtg3, controlled the down-regulation of a large number of genes encoding enzymes involved in amino acid synthesis. All of these TFs were themselves constitutively expressed, suggesting major involvement of post-translational modifications. These interactions would have been missed using the reverse-engineering approach using transcriptomic data alone which requires differential expression of the TF genes. Conversely, the down-regulation of glycolysis is explained by the down-regulation of the Gcr1 and Gcr2 genes. The true identities of regulators that controlled the up-regulated genes were less obvious. Various lines of evidence suggested that the differential activity of, firstly, Adr1 and, secondly, Mig1 was responsible for the up-regulation of the genes encoding enzymes and transporters for alternative carbon source utilisation. Yet, the likelihood framework revealed enrichment of motifs that bore a strong resemblance to those for Adr1 and Mig1 zinc finger motifs. Since the enrichment score for Adr1 was improved when the motif score was emphasised over that of the SGD dataset, it suggested that the targets did not correspond with those of *S. cerevisiae*, or that the true targets of Adr1 were not captured in SGD. A noteworthy development is that the method of optimising the enrichment statistic paves the way for automated network construction and, particularly, that it suggests which improvements could be made to improve the discovery of the true targets, depending on the regulator in question.

References

Broach JR. Nutritional control of growth and development in yeast. Genetics. 2012;192: 73-105.

Bao MZ, Schwartz MA, Cantin GT, Yates JR, Madhani HD. Pheromone-dependent destruction of the Tec1 transcription factor is required for MAP kinase signaling specificity in yeast. Cell. 2004;119(7):991-1000.

Bardwell L, Cook JG, Voora D, Baggott DM, Martinez AR, Thorner J. Repression of yeast Ste12 transcription factor by direct binding of unphosphorylated Kss1 MAPK and its regulation by the Ste7 MEK. Genes Dev. 1998;12(18): 2887-2898.

Collins SR, Kemmeren P, Zhao X, Greenblatt JF, Spencer F, Holstege FCP, Weissman JS, Krogan NJ. Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. Mol Cell Proteomics. 2007;6: 439-450.

Darling AE, Mau B, Perna NT. ProgressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. 2010. PloS ONE 5(6): e11147. doi:10.1371/journal.pone.0011147.

Duda RO, Hart PE, Stork DG. Pattern classification. 1st ed. John Wiley and Sons. 2001.

Elble R, Tye B. Both activation and repression of a-mating-type-specific genes in yeast require transcription factor Mcm1. Proc Nati Acad Sci USA. 1991;88: 10966-10970.

Elion EA, Satterberg B, Kranz JE. FUS3 phosphorylates multiple components of the mating signal transduction cascade: evidence for STE12 and FAR1. Mol Biol Cell. 1993;4(5): 495-510.

Gimeno CJ, Fink GR. Induction of pseudohyphal growth by overexpression of PHD1, a *Saccharomyces cerevisiae* gene related to transcriptional regulators of fungal development. Mol Cell Biol. 1994;14(3): 2100-12.

Holland MJ, Yokoi T, Holland JP, Myambo K, Innis MA. The GCR1 gene encodes a positive transcriptional regulator of the enolase and glyceraldehyde-3-phosphate dehydrogenase gene families in *Saccharomyces cerevisiae*. Mol Cell Biol. 1987;7(2): 813-20.

Jansen R, Yu H, Greenbaum D, Kluger Y, Krigan N, Chung S. A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. Science. 2003;302(5644): 449-453.

Lertwattanasakul N, Kosaka T, Hosoyama A, Suzuki Y, Rodrussamee N, Matsutani M, Murata M, Fujimoto M, Suprayogi S, Tsuchikane K, Limtong S, Fujita N, Yamada M. Genetic basis of the highly efficient yeast *Kluyveromyces marxianus*: complete genome sequence and transcriptome analyses. Biotechnol Biofuels. 2015;8(47). doi: 10.1186/s13068-015-0227-x.

Madhani HD, Fink GR. Combinatorial Control Required for the Specificity of Yeast MAPK Signaling. Science. 1997;275(5304): 1314-1317.

Marbach D, Costello JC, Küffner R, Vega N, Prill RJ, Camacho DM, Allison KR, The DREAM5 Consortium, Kellis M, Collins JJ, Stolovitzky G. Wisdom of crowds for robust gene network inference. Nat Methods. 2013;9(8):796-804. doi:10.1038/nmeth.2016.

Patil KR, Nielsen J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc Natl Acad Sci USA. 2005; 102(8): 2685–2689. PMID: 15710883.

Radman-Livaja M, Rando OJ. Nucleosome positioning: how is it established, and why does it matter? Dev Biol. 2010;339(2): 258-266. doi:10.1016/j.ydbio.2009.06.012.

Roberts RL, Fink GR. Elements of a single MAP kinase cascade in *Saccharomyces cerevisiae* mediate two developmental programs in the same cell type: mating and invasive growth. Genes Dev. 1994;8(24): 2974-85.

Schabort DWP, Letebele PK, Steyn L, Kilian SG, du Preez JC. Differential RNA-seq, multi-network analysis and metabolic regulation analysis of *Kluyveromyces marxianus* reveals a compartmentalised response to xylose. PLoS ONE. 2016;11(6): e0156242. doi:10.1371/journal.pone.0156242.

Springer C, Künzler M, Balmelli T, Braus GH. Amino acid and adenine cross-pathway regulation act through the same 5'-TGACTC-3' motif in the yeast HIS7 promoter. J. Biol. Chem. 1996;271(47): 29637–29643.

Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9): 1105-1111.

Uemura H, Jigami Y. Role of GCR2 in transcriptional activation of yeast glycolytic genes. Mol Cell Biol. 1992;12(9): 3834-42.

Young ET, Dombek KM, Tachibana C, Ideker T. Multiple pathways are co-regulated by the protein kinase Snf1 and the transcription factors Adr1 and Cat8. J Biol Chem. 2003;278(28): 26146–26158.

Chapter 7

Regulation of transcription factors by kinases

Abstract

Kluyveromyces marxianus has been shown to exhibit a strong differential response to glucose or xylose as the carbon source. This response seems to be regulated by a few transcription factors, which were revealed in previous chapters using independent approaches. Notably, the genes encoding these significantly enriched transcription factors, including Gcn4, Ste12 and Phd1, were themselves not differentially expressed. Post-translational modifications thus had to be responsible for altering their activities. These may include phosphorylation as a common post-translational modification, but also a variety of other chemical derivatisations and proteolytic degradation. This chapter aims to elucidate the possible kinase-transcription factor interactions that might be responsible for regulating the activity of enriched transcription factors in *K. marxianus*. The data analyses suggested Ssn3 and Pho85 to be regulating the transcription factors Gcn4 and Ste12 by phosphorylation, which both had enriched regulatory target gene sets. The method proposed here uses exclusively RNA-seq data of the targets of transcription factors, along with a homology-drafted post-translational modification network and suitable network statistics.

Introduction

Understanding gene regulation is an active field in research and many fields of biological research can benefit from knowledge of gene regulation, ranging from microbial biotechnology to the search for cures for hereditary diseases. A major disruptive technology that has emerged in recent years is next-generation sequencing (NGS). A multitude of applications has arisen recently [reviewed in van Dijk et al. 2014, Buermans and den Dunnen 2014]. Major strides have been made in terms of the experimental protocols and raw data processing to reveal new forms of data, including RNA abundance estimation and isoform elucidation using RNA-seq [Trapnell 2010]. Such experiments have drastically altered our view of the biochemical complexity in eukaryotes. However, post-translational modifications (PTMs) play an important role in regulation of all cellular processes [Voet et al. 2016]. Indeed, many of the transcription factors (TFs) that seem to be regulating the differential glucose and xylose response in *Kluyveromyces marxianus*, including Gcn4, Phd1 and Ste12, were not differentially expressed themselves, suggesting a major role for PTMs in regulating the activity of these TFs. Phosphorylation is the most common PTM and many of these have been characterised for *Saccharomyces cerevisiae*. The dimension of PTMs is, however, not accessed by NGS experiments such as RNA-seq. It may still, however, be possible to reveal differential activity of post-translational modifiers, such as kinases, by means of NGS data together with a network enrichment based methodology using a homology-drafted PTM network based on a related model species. It is shown in this chapter how a kinase network was constructed for *K. marxianus* based on experimental data for *S. cerevisiae*. Differential enrichment statistics of TFs using RNA-seq data were calculated as before (Chapter 6) and used as a replacement for differential phospho-proteomics data. Kinase enrichment statistics, based on the TF targets of the kinases, were subsequently calculated. A potentially important innovation demonstrated here was to consider all the TFs reachable by a kinase via kinase cascades as the target set, at any given depth (network distance), termed long-range enrichment. In this manner, even upstream kinases might be revealed. Additionally, a Bayesian likelihood method is demonstrated that prunes the network to reveal the most likely interactions or subnetworks. Since obtaining genome-wide NGS data certainly is less complicated compared to genome-wide phospho-proteomics and other techniques based on mass spectrometry, the method proposed here may have interesting potential.

Materials and Methods

Cultivations of *K. marxianus*, RNA-seq data generation and analyses were described in the Materials and Methods section, Chapter 3. The genome-wide gene regulatory networks were from Chapters 5 and 6.

Constructing a kinase interaction network for *K. marxianus*

The *Saccharomyces* Genome Database (SGD) provided a rich source of post-translational modifications in *S. cerevisiae*, and it was found to be an improvement over that from the annotations from UniProt. YeastMINE provides a convenient framework for retrieving interactions stored in SGD. Computer code was developed to parse the data for YeastMINE and InterMINE to collapse redundant rows and create interactions in a convenient native format, amenable to visualisation and enrichment statistics in *Reactomica*. The list of phosphorylation interactions was very large and contained multiple sources of evidence per interaction, thus constituting a redundant set. Each *S. cerevisiae* gene represented in the interaction dataset was homology mapped to the *K. marxianus* set of proteins by command-line BLASTP, which was run under Windows 7. The E-value cutoff was set to 1E-5 and only the best scoring hit was kept. Subsequently, the interactions in which the target was a regulator of transcription were selected, or more specifically, a transcription factor associated with a DNA binding site represented in the JASPAR motif database.

A Bayesian classifier approach to kinase enrichment networks

A systematic approach was developed to select only interactions with a high likelihood of being regulated only by the kinase of interest and not by any other. A Bayesian classifier calculates the likelihood that one hypotheses (H_1) is more likely, given the data, compared to a competing hypothesis (H_0) [Duda et al. 2001, Collins et al. 2007]. In this case, for each interaction the average enrichment of all the targets (transcription factors) of an effector (kinase) was used as H_1. For each Kinase-TF interaction, the H_0 was also calculated, which is the average enrichment of kinases that also phosphorylate the target (transcription factor) but is not the kinase in the actual interaction (j!=k). The likelihood ratio based on average enrichment *Le* is then calculated as the likelihood ratio H_1/H_0.

$$Le(Ak \rightarrow Bt) = \frac{H1}{H0}$$

$$Le(Ak \rightarrow Bt) = \frac{\frac{1}{T}\sum_t^T S(Bt|Ak)}{\frac{1}{K-1}\sum_{j(j!=k)}^K \frac{1}{T}\sum_t^T S(Bt|Ak)}$$

A_k represents the *k*'th kinase in the complete network of *K* kinases, B_t the *t*'th TF among the *T* target TFs of A_k, and $S(B_t|A_k)$ the enrichment score of the *t*'th transcription factor which is a target of the kinase A_k. The enrichment scores $S(B_t|A_k)$ were calculated using the Z-score method [Ideker et al. 2002, Oliveira et al. 2008, Schabort et al. 2016, Chapter 3] as below.

$$\mathrm{S} = \frac{\mathrm{Z(total, Test)} - \mathrm{Mean(Z, Background)}}{\mathrm{Standard\ deviation(Z, Background)}}$$

Since the number of kinases that affect a given TF was generally low, using the enrichment scores may provide a more sensitive likelihood method as opposed to using the numbers of enriched and non-enriched TFs, as is often used in Bayesian networks. Finally, one of two methods were used to select a final network. Firstly, top scoring interactions were chosen for inclusion in the network. Secondly, kinase subnets, based on the average likelihood values *Le* obtained after the Bayesian classifier, were chosen.

The methods were first applied to the gene regulatory network based and the draft genome of *K. marxianus* UFS-Y2791 (Chapter 5), and repeated using the gene regulatory network based on the complete genome of strain DMKU3-1042 (Chapter 6).

Long-range enrichment and network distances

To calculate a long-range enrichment at arbitrary networks distance, a breadth-first scan was performed. All TFs along the path were collected and their target gene set enrichment scores calculated. These were used as the scores S(Bt).

Results

Figure 1 shows the phosphorylation network homology mapped to *K. marxianus*. Subsequently, the interactions in which the target was a transcription factor were picked. Figure 2 shows this network of 145 interactions, involving 52 kinases and 62 transcriptional regulators. These were transcriptional regulators derived from the effectors of gene regulatory interactions from YeastMINE. Of these, 18 did not map to JASPAR DNA binding motifs and were thus co-activators and co-repressors, including kinases and other post-translational modifiers, but which were not present in the gene regulatory networks constructed in Chapters 5 and 6. A further step was performed to pick only those interactions with the targets associated having a DNA binding site in the JASPAR database, leaving 110 interactions involving 44 transcription factors and 43 kinases (Figure 3).

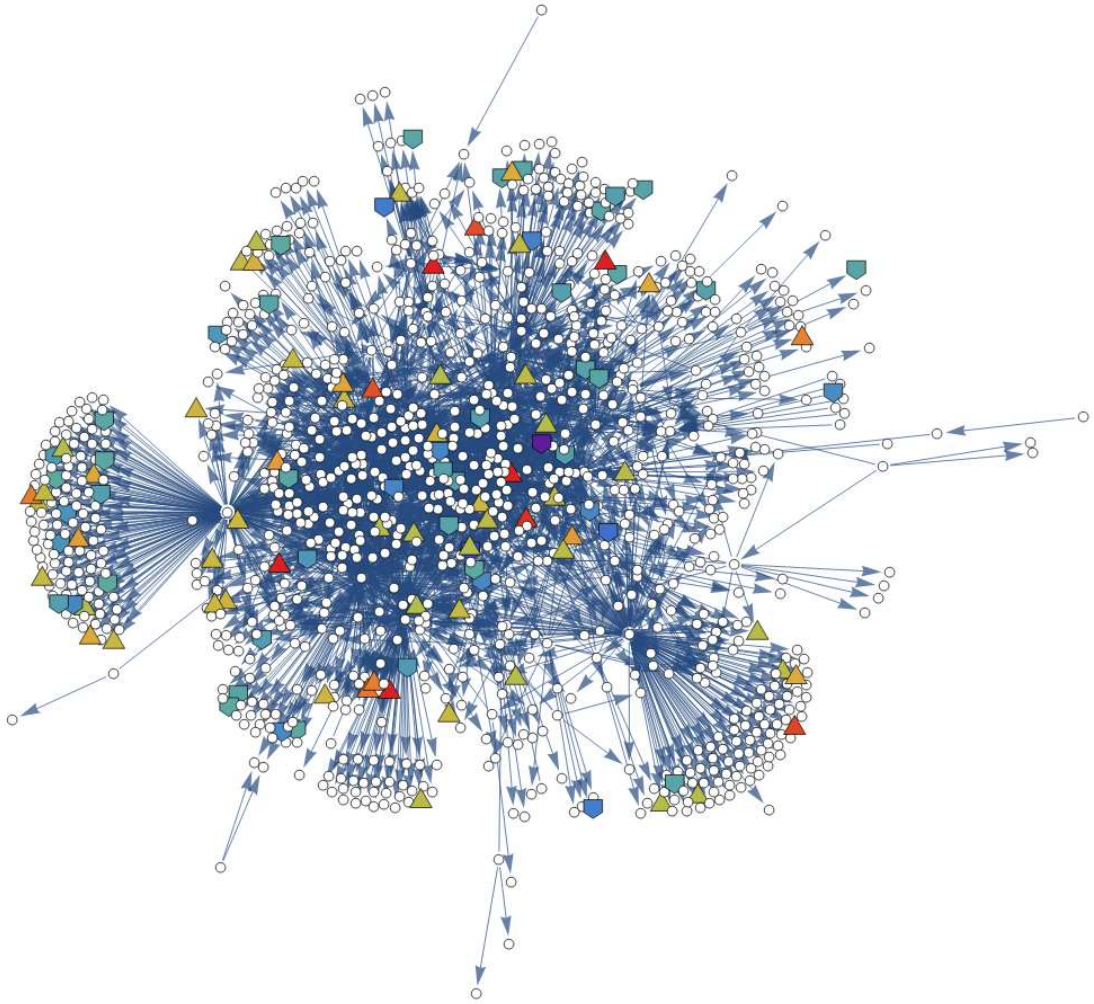


Figure 1. Phosphorylation network in *K. marxianus* drafted on the interactions in *S. cerevisiae*. TFs are indicated as triangles with colouring representing the direction of regulation – warm colours (red) indicates up-regulation while cold colours (blue and purple) indicates down-regulation. Kinases and non-TF proteins are indicated as white circles if constitutively expressed. Non-TF phosphorylation targets are indicated with pentagons with the same colouring scheme if differentially expressed at transcriptional level. Large clusters of targets indicate a central hub kinase. The scheme reveals that generally, kinases as effectors are constitutively expressed at transcriptional level.

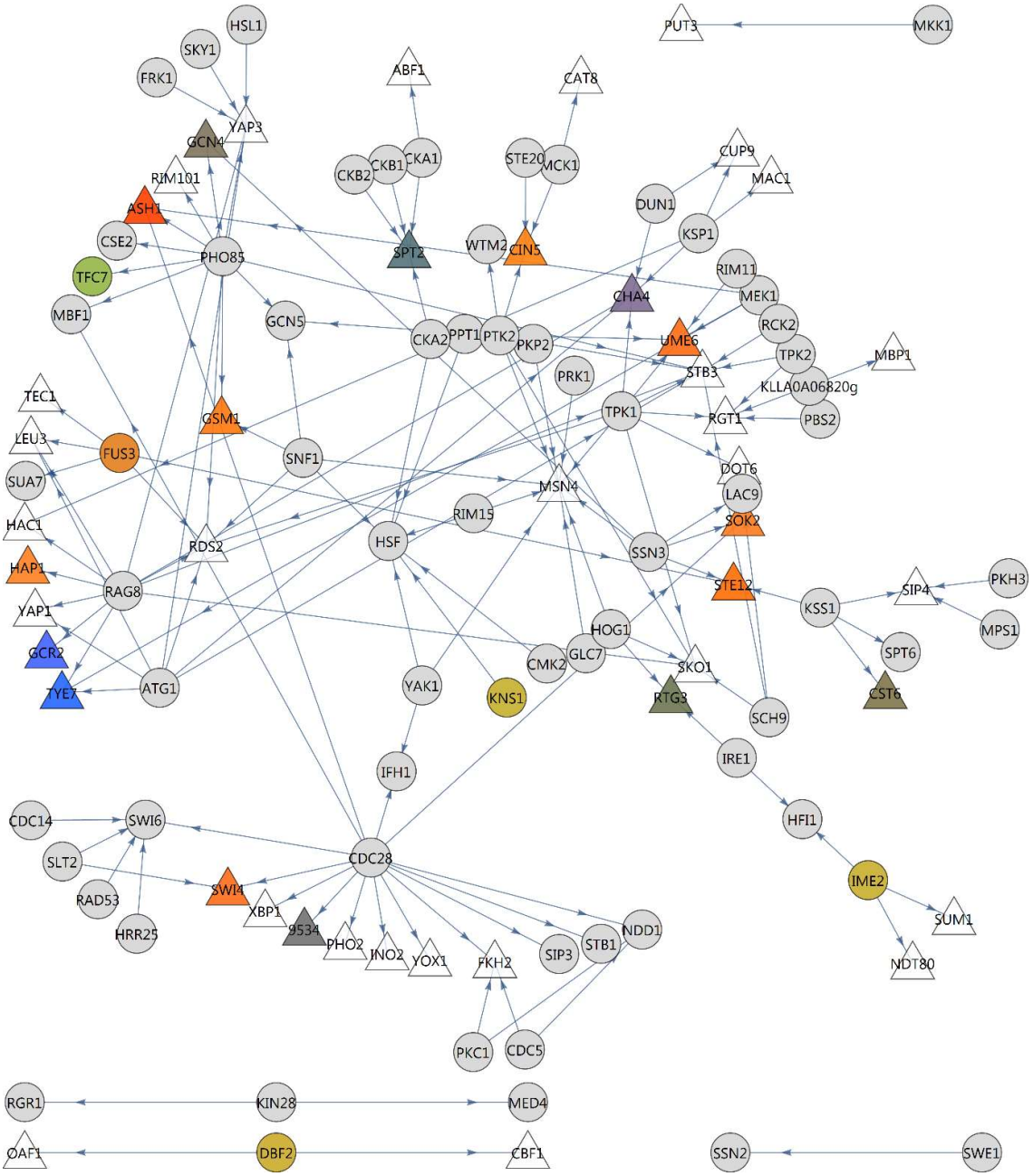


Figure 2. Phosphorylation network in *K. marxianus* drafted on interactions in *S. cerevisiae*, in which the targets are regulators of gene expression, including TFs, co-activators and co-repressors. TFs are indicated as triangles and the colouring scheme indicates the enrichment statistics based on differential expression of all its target genes, including non-regulatory genes. Warm colours (red) indicate a target gene set that was generally up-regulated; cold colours (blue) indicate a target gene set that was generally down-regulated; murkiness indicates a target gene set that was approximately equally up and down-regulated. Non-TF proteins are indicated as circles, with the colouring scheme based on the transcriptional level differential expression of the relevant gene. Warm colours (red) indicate up-regulation and cold colours (blue) indicate down-regulation at transcriptional level.

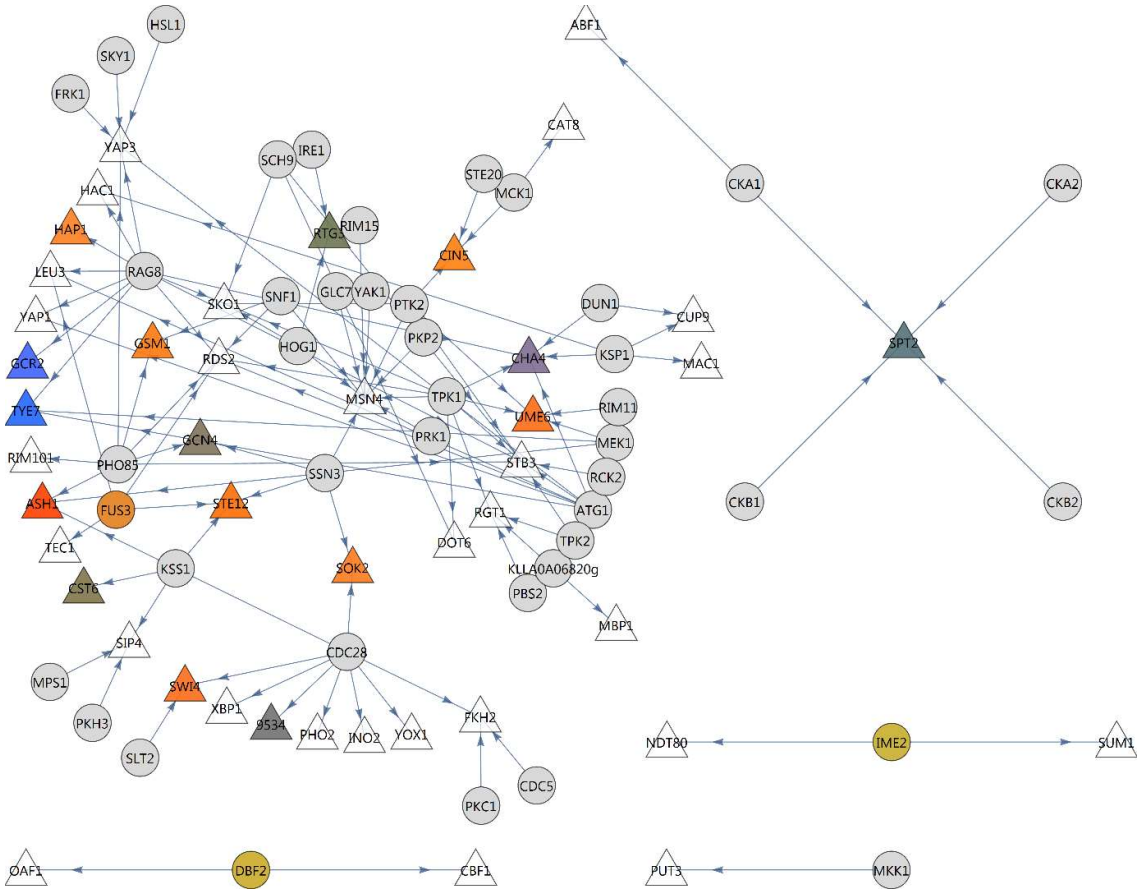


Figure 3. Phosphorylation network in *K. marxianus* drafted on interactions in *S. cerevisiae*, in which the targets were limited to transcription factors that have a DNA binding site in JASPAR. Kinases are indicated by circles while TFs are indicated by triangles. The colouring scheme is the same as in Figure 2.

Of the most significantly enriched TFs from the gene regulatory network based on the draft genome (Chapter 5), Gcn4, Gcr2, Swi4, Sok2, Ash1 and Tye7 were all included, while Gal4, YGR067C and Stp1 were absent. By visual inspection, it appeared that Pho85 was an important regulator as three of its targets were enriched (Gcn4, Ash1, Gsm1), leaving three (Rim101, Yap3, Stb3) non-enriched. Also, Ssn3 was associated with Gcn4, Ste12 and Sok2, leaving only one (Msn4). The CDC28 cyclin dependent kinase had three TF targets enriched (Ash1, Swi4, Sok2), of which their targets were mostly up-regulated. It had a larger fraction of non-enriched transcription factors (Fkh2, Yox1, Ino2, Pho2, Xbp1 and the Dal81 motif for which no gene homolog was found in *K. marxianus*). A high fraction of enrichment was found for Kss1 with two TFs enriched (Ste12 and Cst6) but Sip4 was not enriched. Rag8 had four targets enriched (Hap1, Gcr2, Tye7 and Cha4) and seven non-enriched (Yap1, Leu3, Hac1, Yap3, Sko1, Stb3, Rds2).

Figure 4 shows the resulting network when only subnetworks with average interactions likelihoods above 1 were included. Table 1 shows the underlying data. Ssn3, Mek1 and Pho85 were the top-scoring subsystems with likelihoods above 1.5. Ssn3 and Pho85 both putatively interacted with Gcn4.

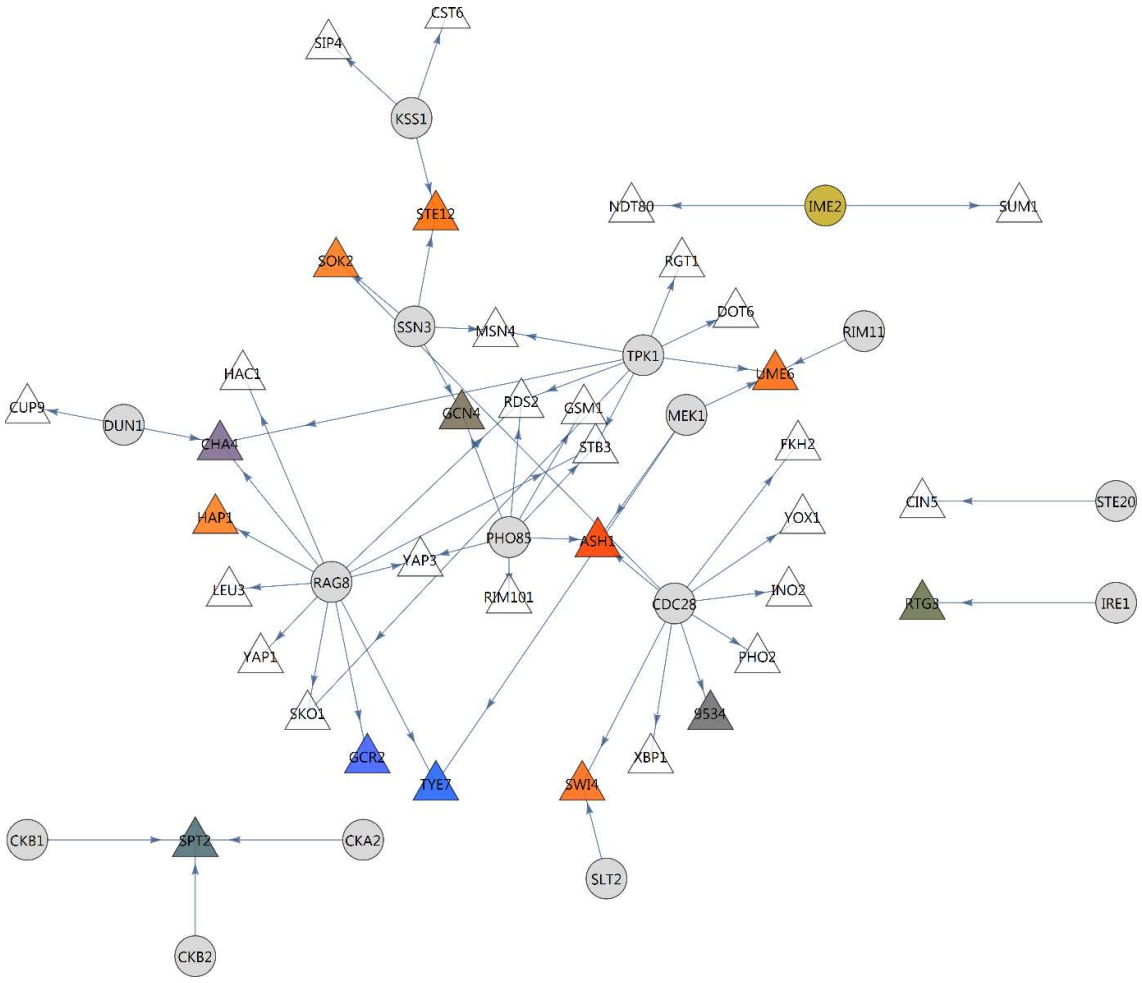


Figure 4. The kinase-TF network when only interactions with Bayesian classifier likelihood based on average enrichments, *Le*, were selected to be above 1. The likelihood *Le* of an interaction A->B was calculated as the ratio of average enrichments of the targets of kinase A, normalised by the average enrichment of all other kinases that may phosphorylate B. Kinases are indicated by circles while TFs are indicated by triangles. The colouring scheme is the same as in Figure 2.

Table 1. Calculation of the likelihoods of interactions based on average enrichments using a Bayesian classifier selected to be above 1. The likelihood *Le* of an interaction A->B was calculated as the ratio of average enrichments of the targets of kinase A, normalised by the average enrichment of all other kinases that may phosphorylate B.

Kinase	Kinase	TF	TF (JAS)	TF	AvEnr(H1)	AvEnr(H0)	Le	AvLe
SSN3	g673.t1	MSN4	9586	g3836.t1	2.65	1.01	2.62	1.77
SSN3	g673.t1	SOK2	9629	g2024.t1	2.65	1.94	1.36	1.77
SSN3	g673.t1	STE12	9637	g3393.t1	2.65	1.47	1.80	1.77
SSN3	g673.t1	GCN4	9547	g1996.t1	2.65	2.04	1.30	1.77
MEK1	g984.t1	TYE7	9653	g3069.t1	2.19	1.30	1.68	1.53
MEK1	g984.t1	UME6	9656	g2782.t1	2.19	1.40	1.56	1.53
MEK1	g984.t1	ASH1	9520	g1041.t1	2.19	1.62	1.35	1.53
PHO85	g2440.t1	RDS2	9606	g1299.t1	1.44	0.84	1.70	1.51
PHO85	g2440.t1	GSM1	9552	g852.t1	1.44	1.10	1.31	1.51
PHO85	g2440.t1	STB3	9634	g1338.t1	1.44	0.64	2.25	1.51
PHO85	g2440.t1	ASH1	9520	g1041.t1	1.44	1.62	0.89	1.51
PHO85	g2440.t1	RIM101	9612	g2725.t1	1.44	1.44	1.00	1.51
PHO85	g2440.t1	YAP3	9660	g3426.t1	1.44	0.53	2.73	1.51
PHO85	g2440.t1	GCN4	9547	g1996.t1	1.44	2.04	0.70	1.51
SLT2	g556.t1	SWI4	9645	g4004.t1	3.04	2.14	1.42	1.42
STE20	g2415.t1	CIN5	9528	g3405.t1	1.31	0.94	1.39	1.39
RIM11	g425.t1	UME6	9656	g2782.t1	1.91	1.40	1.36	1.36
IRE1	g2659.t1	RTG3	9620	g545.t1	1.89	1.41	1.34	1.34
RAG8	g4252.t1	RDS2	9606	g1299.t1	1.00	0.84	1.18	1.26
RAG8	g4252.t1	TYE7	9653	g3069.t1	1.00	1.30	0.76	1.26
RAG8	g4252.t1	GCR2	9549	g2462.t1	1.00	1.00	1.00	1.26
RAG8	g4252.t1	SKO1	9626	g4710.t1	1.00	0.71	1.41	1.26
RAG8	g4252.t1	YAP1	9659	g2401.t1	1.00	0.86	1.16	1.26
RAG8	g4252.t1	LEU3	9568	g2029.t1	1.00	0.75	1.34	1.26
RAG8	g4252.t1	HAP1	9556	g781.t1	1.00	1.00	1.00	1.26
RAG8	g4252.t1	CHA4	9527	g2709.t1	1.00	0.77	1.29	1.26
RAG8	g4252.t1	STB3	9634	g1338.t1	1.00	0.64	1.56	1.26
RAG8	g4252.t1	HAC1	9554	g341.t1	1.00	0.82	1.22	1.26
RAG8	g4252.t1	YAP3	9660	g3426.t1	1.00	0.53	1.90	1.26
TPK1	g1893.t1	RDS2	9606	g1299.t1	0.64	0.84	0.76	1.21
TPK1	g1893.t1	SKO1	9626	g4710.t1	0.64	0.71	0.90	1.21
TPK1	g1893.t1	CHA4	9527	g2709.t1	0.64	0.77	0.82	1.21
TPK1	g1893.t1	UME6	9656	g2782.t1	0.64	1.40	0.46	1.21
TPK1	g1893.t1	STB3	9634	g1338.t1	0.64	0.64	1.00	1.21
TPK1	g1893.t1	MSN4	9586	g3836.t1	0.64	1.01	0.63	1.21
TPK1	g1893.t1	RGT1	9611	g3741.t1	0.64	0.19	3.40	1.21
TPK1	g1893.t1	DOT6	9595	g261.t1	0.64	0.37	1.74	1.21
DUN1	g4717.t1	CUP9	9532	g125.t1	0.88	0.76	1.16	1.15
DUN1	g4717.t1	CHA4	9527	g2709.t1	0.88	0.77	1.13	1.15
CKA2	g2293.t1	SPT2	9631	g3039.t1	1.97	1.72	1.14	1.14

Kinase	Kinase	TF	TF (JAS)	TF	AvEnr(H1)	AvEnr(H0)	Le	AvLe
CKB2	g3674.t1	SPT2	9631	g3039.t1	1.97	1.72	1.14	1.14
CKB1	g1731.t1	SPT2	9631	g3039.t1	1.97	1.72	1.14	1.14
CDC28	g1576.t1	FKH2	9541	g2016.t1	1.24	0.41	3.00	1.11
CDC28	g1576.t1	YOX1	9677	g48.t1	1.24	1.24	1.00	1.11
CDC28	g1576.t1	INO2	9565	g2711.t1	1.24	1.24	1.00	1.11
CDC28	g1576.t1	PHO2	9600	g940.t1	1.24	1.24	1.00	1.11
CDC28	g1576.t1	ASH1	9520	g1041.t1	1.24	1.62	0.76	1.11
CDC28	g1576.t1	SOK2	9629	g2024.t1	1.24	1.94	0.64	1.11
CDC28	g1576.t1	9534	9534	g3314.t1	1.24	1.24	1.00	1.11
CDC28	g1576.t1	XBP1	9658	g2807.t1	1.24	1.24	1.00	1.11
CDC28	g1576.t1	SWI4	9645	g4004.t1	1.24	2.14	0.58	1.11
KSS1	g4162.t1	SIP4	9624	g3032.t1	1.25	0.92	1.36	1.07
KSS1	g4162.t1	CST6	9530	g235.t1	1.25	1.25	1.00	1.07
KSS1	g4162.t1	STE12	9637	g3393.t1	1.25	1.47	0.85	1.07

The Ssn3 (better known as Srb10 or Cdc8) subsystem showed the strongest enrichment of all kinase subsystems with Gcn4, Sok2 and Ste12 enriched. Its other target, the multi stress response regulator Msn4, was not enriched, consistent with the lack of observing an enriched stress response in Gene Ontology enrichment (see Chapter 3, Schabort et al. 2016).

Explaining away kinase-TF interactions by differential gene expression

Additionally, the possible effect of kinases on transcription factor activity could be explained away by considering differential gene expression of the transcription factor genes, which is a simpler explanation for the observed differential activity. Figure 5 shows the same network, but rendered with the differential expression of mRNA values from RNA-seq. The TF Ash1 which was enriched in the Pho85 target subnetwork was up-regulated 2.4-fold in the RNA-seq data. By contrast, none of the targets of Ssn3 were differentially regulated at the gene level, in support of differential kinase activity of Ssn3.

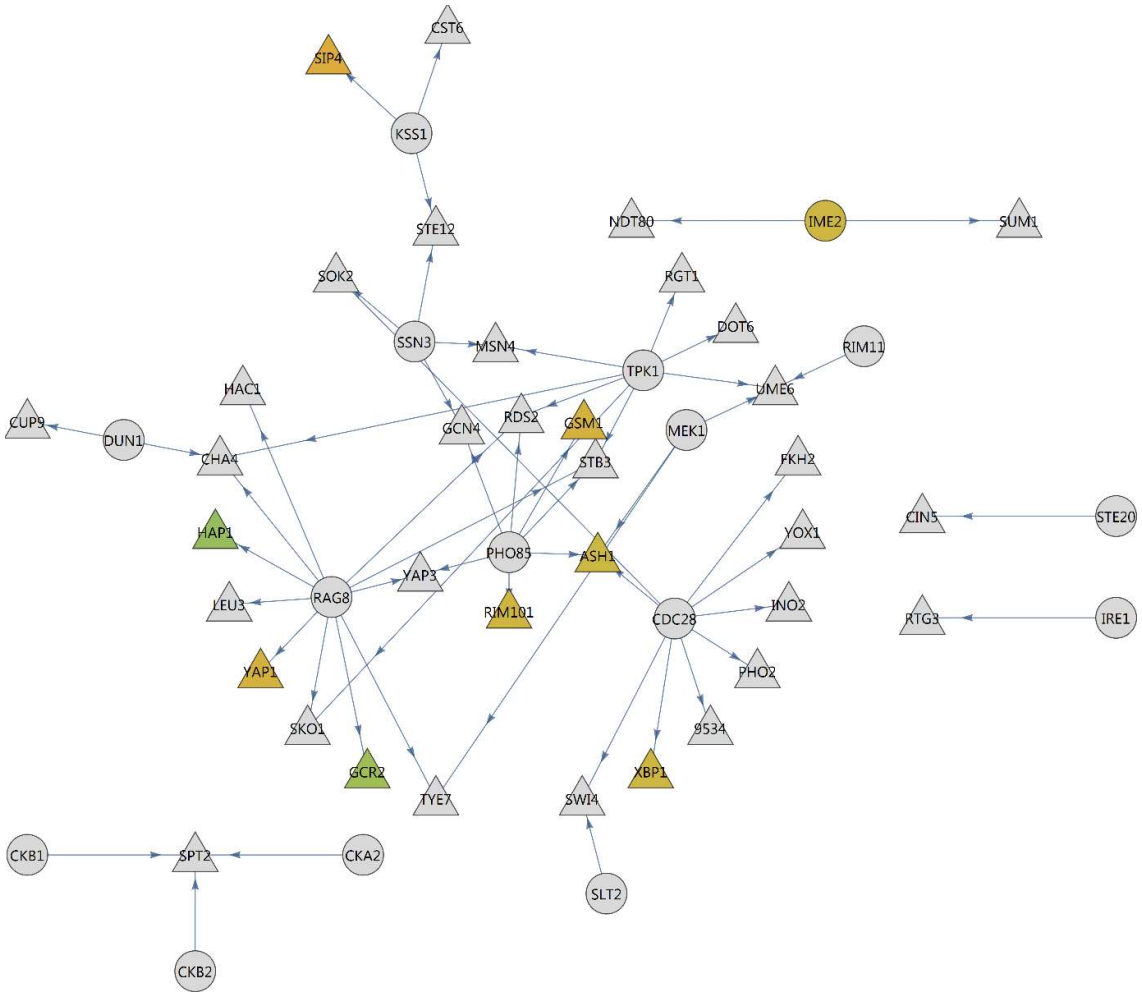


Figure 5. The kinase-TF network after Bayesian classifier likelihood selection, rendered with differential gene expression values of TFs instead of enrichment values. Enriched target TFs of Pho85 Ash1 and Gsm1 were moderately up-regulated, eliminating the need for kinase activity.

Phosphorylation sites in Gcn4

In principle, the presence of a conserved phosphorylation site in the Gcn4 protein of *K. marxianus* would further support the regulation of Gcn4 activity by the ubiquitin-dependent degradation, involving site-specific phosphorylation [Chi et al. 2001, Shemer 2002]. O'Shea et al. [1991] determined the first crystal structure of Gcn4. It is a two-stranded coil with very few amino acids buried in the core of the protein. Therefore, most of the potential phosphorylation sites that can be identified by sequence analysis may be exposed to the surface. It is, however, not trivial to map the amino acid residue numbers between different species, as the lengths of proteins orthologues may vary between species. For *S. cerevisiae*, the Gcn4 protein has 281 residues, whereas in *K. marxianus* UFS-Y2791 it has 357 amino acids. Generally, the lengths of amino acid sequences corresponded to those of the

reference sequences. Sequence alignment had to be performed to properly map sites between species. The sequences of Gcn4 in *K. marxianus* UFS-Y2791 and *S. cerevisiae* were aligned using BLASTP at NCBI for two sequences. Figure 6 shows that the threonine 165 -proline 166 pair that is the critical site for phosphorylation by Pho85-Pcl5 and likely also by Srb10, was indeed conserved between the two species. The site is at residues 237 and 238 in *K. marxianus*. This phosphorylated threonine site is recognised by the ubiquitination system in *S. cerevisiae* [Chi et al. 2001, Shemer 2002].

Figure 6. Alignment of Gcn4 amino acid sequence in *K. marxianus* UFS-Y2791 (query) and *S. cerevisiae* (subject). The conserved phosphorylation site is indicated in the red block.

Use of the gene regulatory network based on the complete genome of DMKU3-1042

The same analyses were carried out using the optimised gene regulatory network based on the complete genome for strain DMKU3-1042 (Chapter 6). Figure 7 shows a similar network as was constructed for the draft genome. At a network distance of 1, which corresponded to direct kinase-TF interactions, the well-known master kinases Pho85, Ssn3 and Pkc1 were on top of the list of mean enrichments (Table 2).

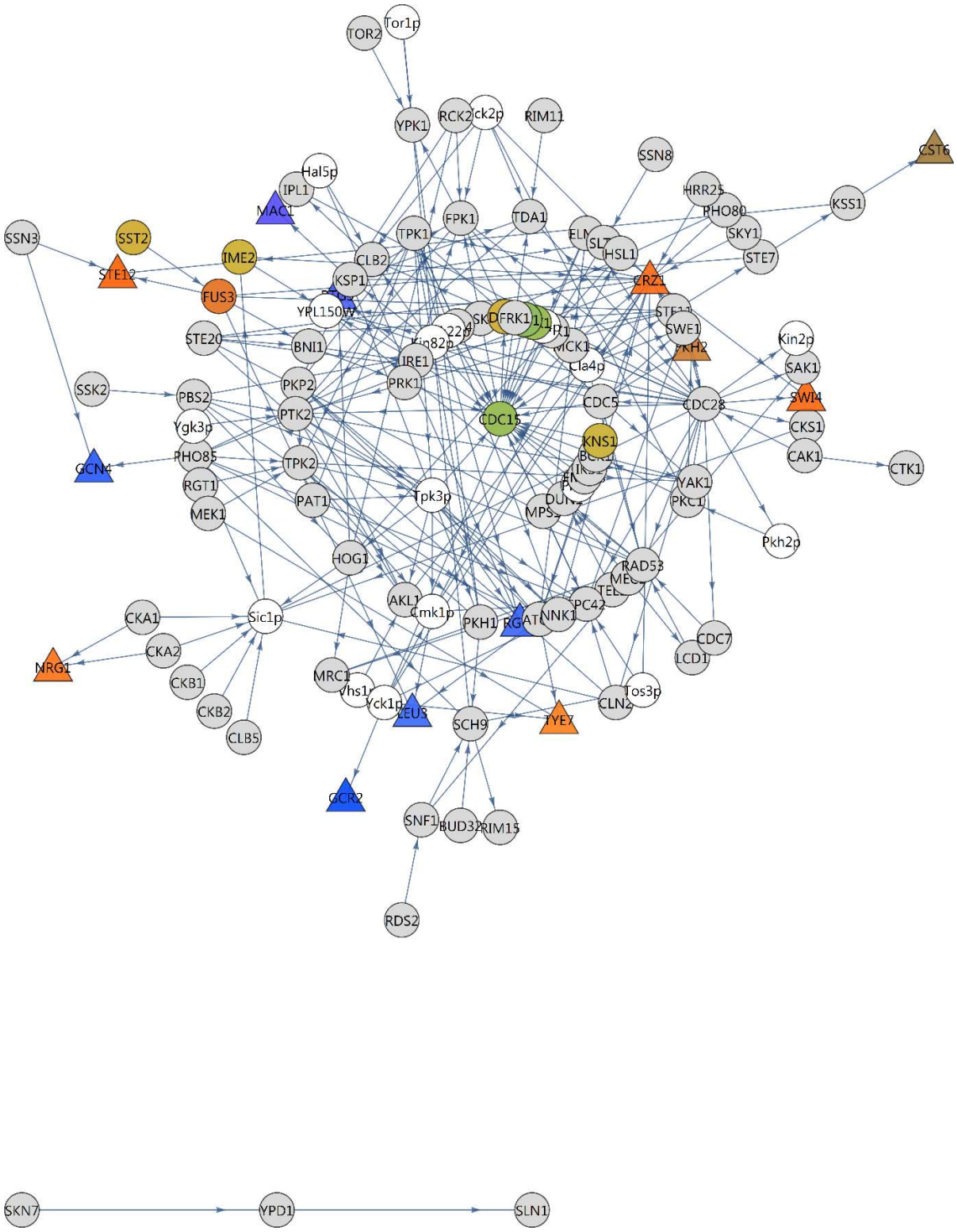


Figure 7. Phosphorylation network in *K. marxianus* drafted on interactions in *S. cerevisiae*, in which the targets were the transcription factors included in the enrichment-optimised gene regulatory network based on the complete genome (Chapter 6). Kinases are indicated by circles while TFs are indicated by triangles. The colouring scheme is the same as in Figure 2.

Table 2. Enrichment statistics of long-range enrichment of kinases via interaction with transcription factors at a network depth of 1. The analysis was based on the enrichment-optimised regulatory network based on the DMKU3-1042 genome, containing only the 38 TFs included in the optimised TF regulatory network from Chapter 6. Kinases not shown obtained a score of zero.

ID	name	Gene signt.	log2FC	FC	n Ints	n Kin	n TFs	n Enr	Av Enr
gene4081	PHO85	no	0	1	8	6	2	2	7.23
gene4089	SSN3	no	0	1	2	0	2	2	6.83
gene4374	PKC1	no	0	1	2	1	1	1	6.81
gene260	CDC5	no	0	1	5	4	1	1	6.81
gene2023	CDC28	no	0	1	23	21	2	2	5.71
S000005098	Yck2p	-	-	-	5	4	1	1	5.55
S000003147	Tos3p	-	-	-	3	2	1	1	5.55
gene639	HRR25	no	0	1	1	0	1	1	5.55
gene4740	PHO80	no	0	1	1	0	1	1	5.55
gene375	NNK1	no	0	1	1	0	1	1	5.55
gene3314	SWE1	no	0	1	3	2	1	1	5.55
gene307	SKY1	no	0	1	1	0	1	1	5.55
gene2807	MCK1	no	0	1	5	4	1	1	5.55
gene2516	AKL1	no	0	1	1	0	1	1	5.55
gene2239	IME2	yes	1.38	2.61	2	1	1	1	5.55
gene1729	STE20	no	0	1	4	3	1	1	5.55
gene3808	KSP1	no	0	1	6	4	2	2	5.15
gene2308	KSS1	no	0	1	2	0	2	2	4.94
gene1929	CKA2	no	0	1	2	1	1	1	4.89
gene1356	CKA1	no	0	1	2	1	1	1	4.89
S000001177	Yck1p	-	-		7	3	4	4	4.69
gene4568	SLT2	no	0	1	2	1	1	1	4.61
gene3919	MEK1	no	0	1	2	1	1	1	4.51
gene1985	ATG1	no	0	1	5	2	3	3	4.46
S000001910	Cmk1p	-	-		3	1	2	2	4.43
gene3298	TPK1	no	0	1	7	5	2	2	4.40
gene3412	IRE1	no	0	1	4	3	1	1	4.03
gene2579	HOG1	no	0	1	5	4	1	1	4.03
gene873	FUS3	yes	4.64	25.01	3	1	2	2	4.02
S000003701	Hal5p	-	-		2	1	1	1	3.25
S000001649	Tpk3p	-	-		9	8	1	1	3.25
gene3731	PBS2	no	0	1	3	2	1	1	3.25

To investigate whether differentially active upstream kinases could be detected by long-range enrichment, the analyses were performed at various maximal depths (network distances from each kinase via other kinases). The overall distribution of long-range enrichment scores indicated increasing fractions of kinases with mediocre scores with an increase in network depth (Figure 8). Figure 9 reveals four classes of kinases: (*a*) kinases with scores that remained at their original values of zero, which were those kinases for which the targets had led only to other kinases, (*b*) kinases with scores that remained at their original non-zero values, indicating that all TFs along the path were discovered at a depth of one, (*c*) kinases with scores that dropped from an initial high value, indicating that at least one kinase was a target, leading to TFs with low enrichment, and (*d*) kinases with scores that increased at some point, indicating that at least one kinase was a target, which led to TFs with high enrichment. In general, short-range interactions could be assigned with higher confidence as opposed to long-

range interactions, since fewer assumptions were made. Enrichment of kinases belonging to class (*b*) were thus assigned with more confidence as opposed to classes (*c*) and (*d*).

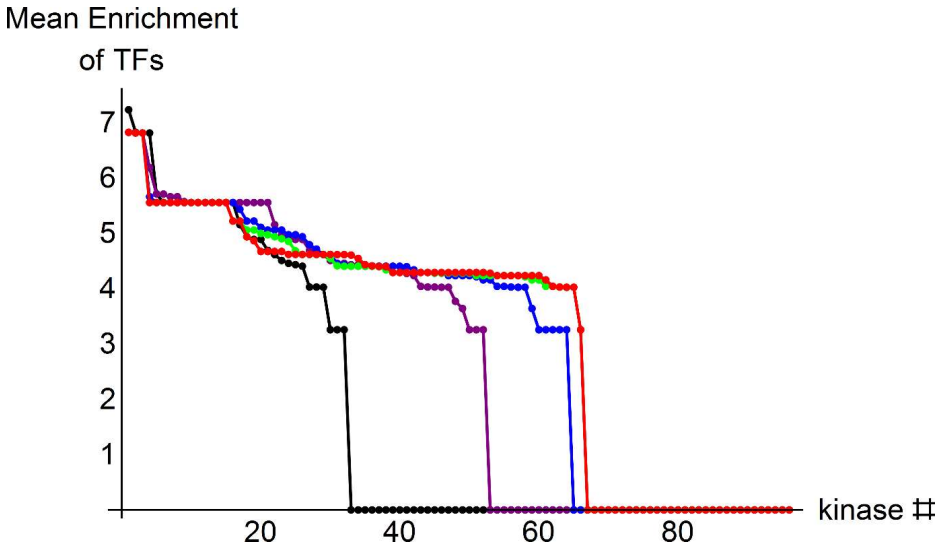


Fig 8. Distribution of enrichment values of kinases at various network depths. Scores were calculated as the mean enrichment values of the TFs that a kinase can phosphorylate itself or via a kinase cascade. Black, depth = 1; Purple, depth = 2; Blue, depth = 3; Green, depth = 4; Red, depth = 5.

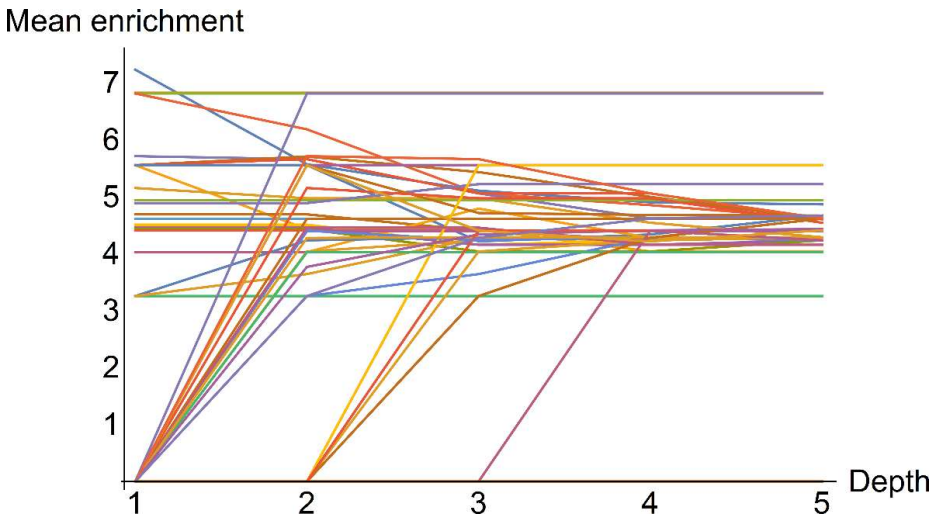


Figure 9. Long-range enrichment of kinases via interaction with transcription factors as a function of network depth, using only TFs present in the enrichment-optimised regulatory network based on the *K. marxianus* DMKU3-1042 strain. Four classes of kinases are distinguishable: classes of kinases: (*a*) kinases with scores that remained at their original values of zero (*b*) kinases with scores that remained at their original non-zero values (*c*) kinases with scores that dropped from an initial high value, and (*d*) kinases with scores that reaches a maximum enrichment and subsequent decrease in enrichment.

Since this TF-kinase network was constructed only from enriched TFs in the enrichment-optimised gene regulatory network (Chapter 6), this analysis resulted in either high-scoring kinase-TF interactions or zero-scoring interactions. To develop an improved scoring method, another kinase-TF network was constructed, which contained all kinase-TF interactions in the original phosphorylation network (Figure 10). In the cases when a TF was not present among the 38 TFs of the optimised TF network, a value of zero was assigned as enrichment score for a TF. In Figure 11 it is evident that fewer kinases obtained high scores using the improved method. Figure 12 additionally shows that the top kinases kept the same high scores as a function of network depth. On closer inspection it is evident that seven (Akl1, Pho80, Hrr25, Ssn3, Sky1, Nnk1, Kss1) of the top eight kinases were in the category of kinases that kept high enrichments (Table 3). Other kinases (Pkh2p, Ssn8, Vhs1p, Ste7) reached an optimum only by discovering other kinases first.

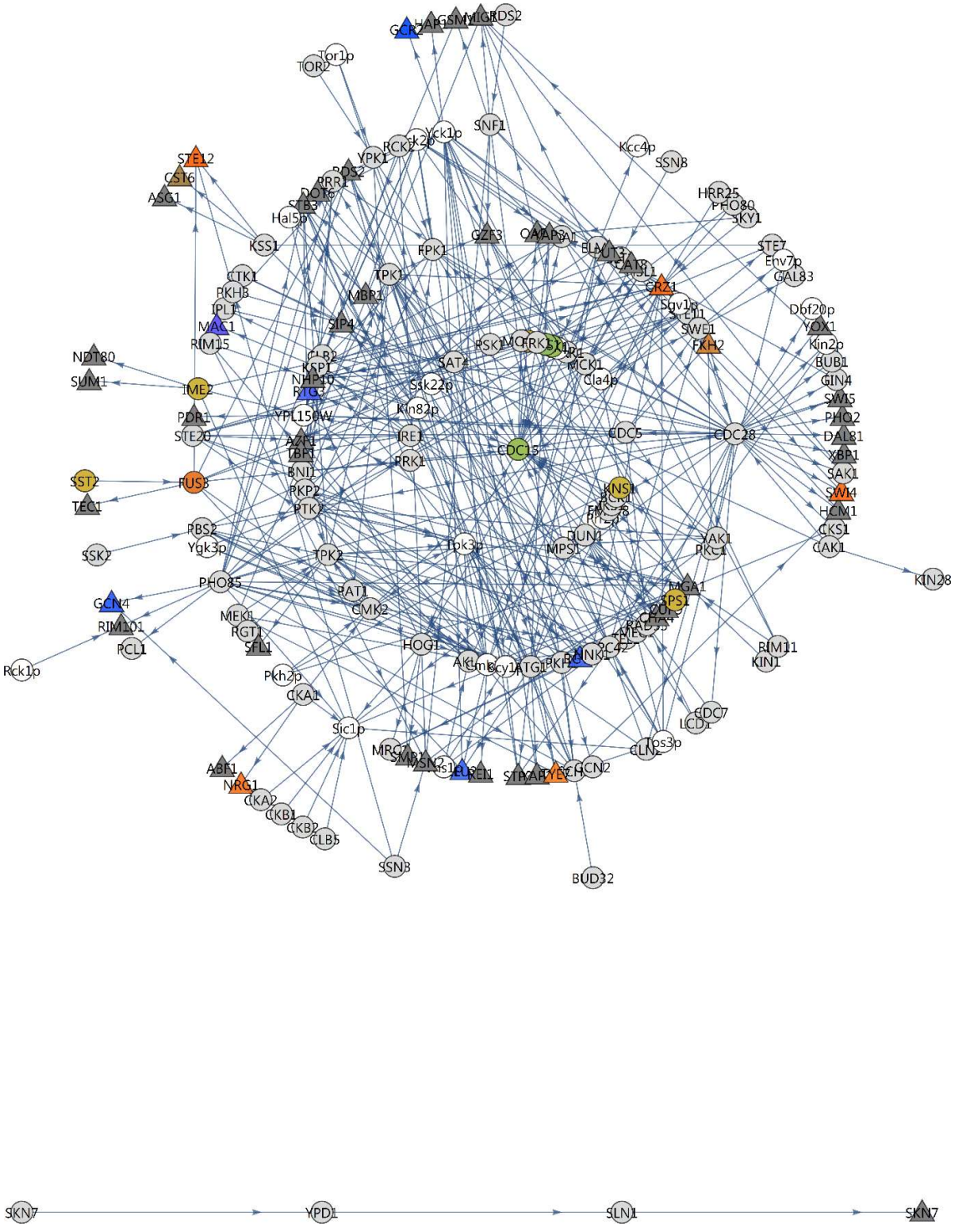


Figure 10. Phosphorylation network in *K. marxianus* drafted on interactions in *S. cerevisiae*, in which the targets were any of the gene regulators, regardless of their inclusion in the enrichment-optimised gene regulatory network from Chapter 6. Kinases are indicated by circles while TFs are indicated by triangles. The colouring scheme is the same as in Figure 2.

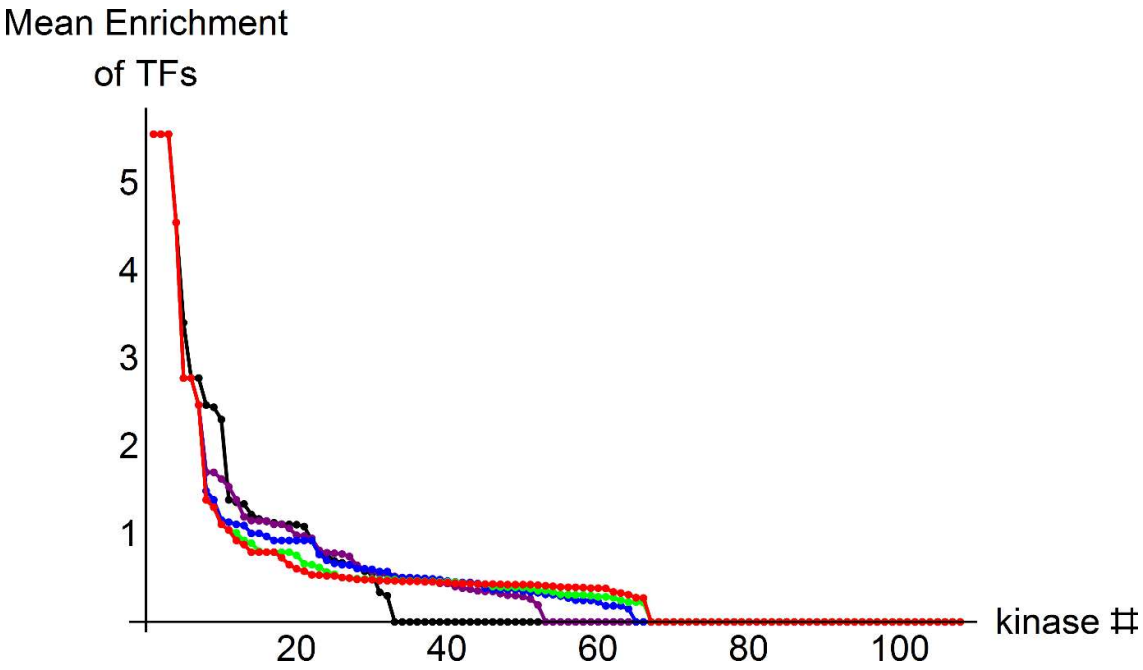


Fig 11. Distribution of enrichment values of kinases at various network depths. Scores were calculated as the mean enrichment values of the TFs that a kinase can phosphorylate itself or via a kinase cascade. Black, depth = 1; Purple, depth = 2; Blue, depth = 3; Green, depth = 4; Red, depth = 5.

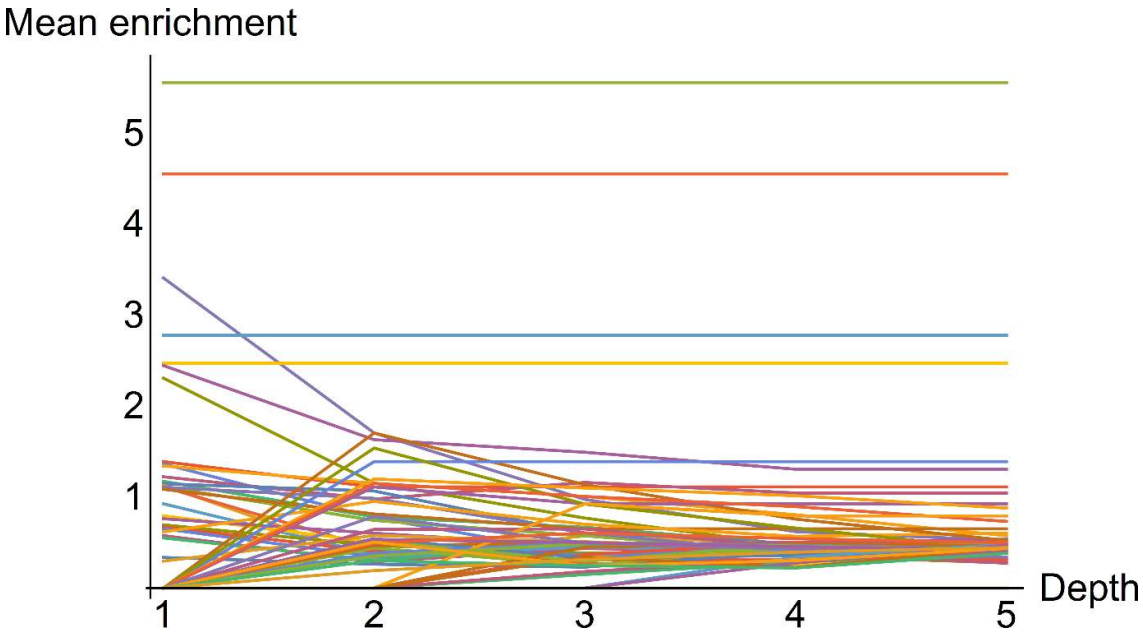


Figure 12. Long-range enrichment of kinases via interaction with transcription factors as a function of network depth, using all TFs present in the phosphorylation network. An enrichment score of zero was assigned to a TF if it did not exist in the enrichment-optimised gene regulatory network.

Table 3. Statistics of the top 20 candidate kinases suggested by long-range enrichment via interaction with transcription factors as a function of network depth. Depth is indicated by numbers 1 to 5. The values x;(y);(z) represent: mean enrichment of TFs; (number of TFs enriched / number of TFs reachable along cascade); (Log2(Fold change) of kinase gene). If the kinase was not differentially expressed in RNA-seq data, a value of 0 was given for Log2(Fold change). Dashes indicate that the kinase gene was not annotated in the DMKU3-1042 genome. Entries in bold indicate maximal mean enrichment scores.

Gene	1	2	3	4	5
AKL1	**5.55(1/1)(0)**	**5.55(1/1)(0)**	**5.55(1/1)(0)**	**5.55(1/1)(0)**	**5.55(1/1)(0)**
PHO80	**5.55(1/1)(0)**	**5.55(1/1)(0)**	**5.55(1/1)(0)**	**5.55(1/1)(0)**	**5.55(1/1)(0)**
HRR25	**5.55(1/1)(0)**	**5.55(1/1)(0)**	**5.55(1/1)(0)**	**5.55(1/1)(0)**	**5.55(1/1)(0)**
SSN3	**4.55(2/3)(0)**	**4.55(2/3)(0)**	**4.55(2/3)(0)**	**4.55(2/3)(0)**	**4.55(2/3)(0)**
PKC1	**3.41(1/2)(0)**	1.7(1/4)(1)	0.973(1/7)(0)	0.619(1/11)(0)	0.524(1/13)(0)
SKY1	**2.78(1/2)(0)**	**2.78(1/2)(0)**	**2.78(1/2)(0)**	**2.78(1/2)(0)**	**2.78(1/2)(0)**
NNK1	**2.78(1/2)(0)**	**2.78(1/2)(0)**	**2.78(1/2)(0)**	**2.78(1/2)(0)**	**2.78(1/2)(0)**
KSS1	**2.47(2/4)(0)**	**2.47(2/4)(0)**	**2.47(2/4)(0)**	**2.47(2/4)(0)**	**2.47(2/4)(0)**
CKA2	**2.44(1/2)(0)**	1.63(1/3)(0)	1.49(2/7)(0)	1.31(2/8)(0)	1.31(2/8)(0)
SLT2	**2.31(1/2)(0)**	1.15(1/4)(0)	0.769(1/6)(0)	0.461(1/10)(0)	0.384(1/12)(0)
Pkh2p	0(0/2)(-)	**1.7(1/4)(-)**	1.14(1/6)(-)	0.757(1/9)(-)	0.524(1/13)(-)
SSN8	0(0/1)(**0**)	**1.54(1/3)(0)**	0.922(1/5)(0)	0.659(1/7)(0)	0.419(1/11)(0)
Vhs1p	0(0/3)(-)	**1.39(1/4)(-)**	**1.39(1/4)(-)**	**1.39(1/4)(-)**	**1.39(1/4)(-)**
IME2	**1.39(1/4)(1.4)**	1.11(1/5)(1.4)	1.11(1/5)(1.4)	1.11(1/5)(1.4)	1.11(1/5)(1.4)
CDC5	**1.36(1/5)(0)**	0.773(2/16)(0)	0.368(4/55)(0)	0.285(5/85)(0)	0.326(7/99)(0)
FUS3	**1.34(2/6)(4.6)**	1.15(2/7)(4.6)	1.01(2/8)(4.6)	0.805(2/10)(4.6)	0.575(2/14)(4.6)
CKA1	**1.22(1/4)(0)**	0.978(1/5)(0)	1.16(2/9)(0)	1.04(2/10)(0)	1.04(2/10)(0)
STE7	0(0/2)(**0**)	**1.2(3/11)(0)**	1.1(3/12)(0)	1.01(3/13)(0)	0.879(3/15)(0)
Yck1p	**1.17(4/16)(-)**	0.782(4/24)(-)	0.667(5/33)(-)	0.479(5/46)(-)	0.505(7/61)(-)

The TF-kinase network was also pruned by assigning a likelihood to each interaction, based on the ratio of mean enrichment of a kinase (as calculated from all its direct targets), versus the average enrichment of all other kinases that may phosphorylate the TF in question, and taking the interactions with highest likelihood. Taking only the interactions with likelihoods greater than one led to a reduced explained-away network containing the most likely kinase-TF interactions (Figure 13). An alternative approach (likelihood subnetworks) was to keep those kinases which had high average likelihoods for their direct interactions, based on the explained-away method. In the latter chase, Ssn3, Cdc5, Pho80, Akl1 and Hrr25 were identified (Figure 14).

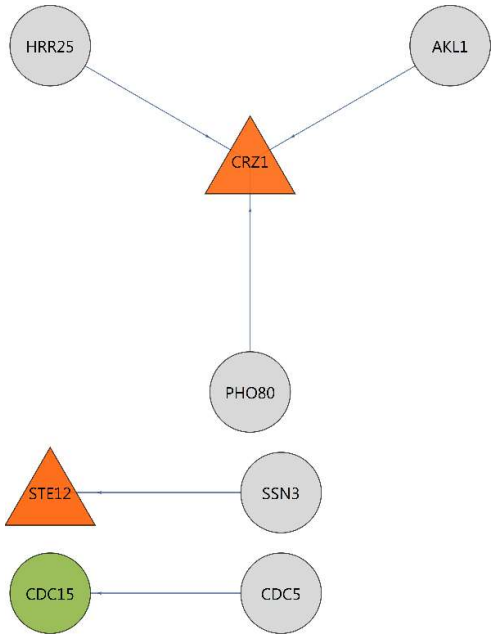


Figure 13. Kinase-TF network showing only interactions with a high likelihood, after explaining away differential activity that could instead have arisen by the average effect of all other effectors of a target. Kinases are indicated by circles while TFs are indicated by triangles. The colouring scheme is the same as in Figure 2.

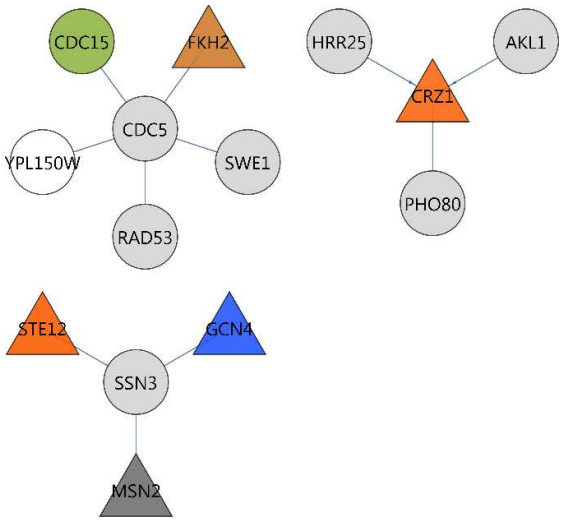


Figure 14. Kinase-TF network showing only kinases with a high average likelihood of target interactions, after explaining away differential activity that could instead have arisen by the average effect of all other effectors of a target. Kinases are indicated by circles while TFs are indicated by triangles. The colouring scheme is the same as in Figure 2.

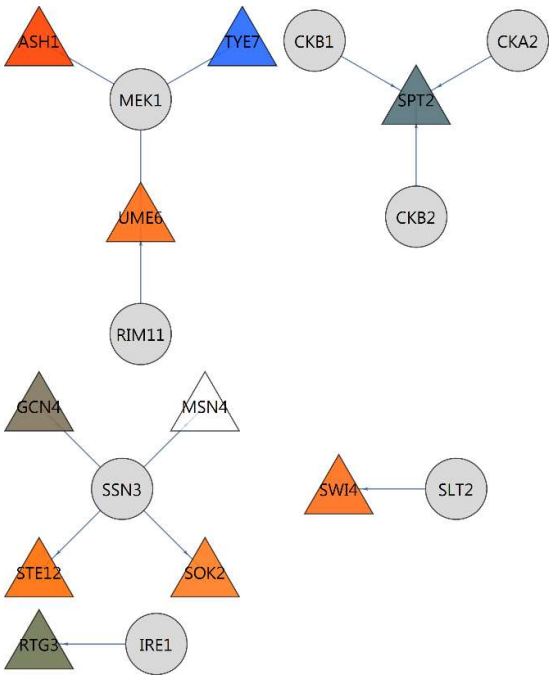


Figure 15. Kinase-TF network showing only kinases subnetworks with a high average likelihood of target interactions, after explaining away differential activity that could instead have arisen by the average effect of all other effectors of a target. Kinases are indicated by circles while TFs are indicated by triangles. The colouring scheme is the same as in Figure 2.

The same likelihood subnetworks approach was used on the TF-kinase network, which again extracted Ssn3, but also Mek1, Rim11, Ckb1, Ckb2, Cka2, Ire1 and Slt2 (Figure 15).

Discussion

In this chapter it was shown that the knowledge on the phosphorylation network of the model species could in principle be used to create a putative phosphorylation network of *K. marxianus*. An innovation in this chapter was to calculate enrichments for kinases based on the enrichments of the transcription factors that they phosphorylate. This was extended to long-range enrichment, which could do the same through a phosphorylation network at any depth. In such a way, differentially active upstream kinases such as Snf1 may be discoverable. In general, short-range interactions can be assigned with a higher confidence as opposed to long-range interactions, since fewer assumptions are made. The upstream kinases such as Snf1 were notably absent from this analysis. Since Snf1 was expected to be differentially active between the conditions with glucose and xylose as carbon source, which very likely

involved the Snf1 targets Adr1 and Mig1 (Chapter 5), this lack of Snf1 enrichment probably indicates that Snf1 is too far upstream to be resolved as differentially active by the long-range method.

Additionally, a Bayesian likelihood approach was taken to extract a network of interactions in which the differential activity of the targets could not be explained away by the average enrichment of all its other effector kinases, but were more likely the result of a single effector kinase. The explained-away likelihood network included the most important TF from the down-regulation category, Gcn4, as well as Ste12 from the up-regulation category. The genes for both of these TFs were constitutively expressed, suggesting their significant enrichments might be explained by the differential activity of kinases Ssn3 or Pho85. While Pho85 was considered potentially important when using the analysis based on the draft genome, it was considered less supported when using the complete genome of DMKU3-1042. Both of these kinases could lead to the phosphorylation and proteolytic degradation of Gcn4 by the ubiquitin-dependent system [Chi et at. 2001, Nelson et al. 2003, Raithatha et al. 2012, Shemer et al. 2002].

Ssn3 is better known as Srb10 or Cdc8 and is a negative regulator of stress response genes and those for pseudohyphal and invasive growth [Raithatha et al. 2012]. In *S. cerevisiae*, the expression level of Ssn3 drops dramatically in exponentially growing cells as they approach the oxidative shift [Holstege, et al. 1998], allowing stress response genes to work. The effect Ssn3 on the transcription factor Msn2 is performed by somehow restricting localisation to the cytoplasm [Chi et at. 2001]. Gcn4, Ste12 and Phd1 perform a specific phosphorylation that is recognised by the uniquitin dependent degradation system [Chi et at. 2001, Shemer et al. 2002, Nelson et al. 2003, Raithatha et al. 2012]. Ssn3 marks chromatin-bound Gcn4p for degradation [Chi et al. 2001]. Ssn3, which also binds and phosphorylates the C-terminal domain of RNA polymerase II [Hengartner et al. 1998], may function as a molecular clock, limiting the time that a transcription factor can remain bound to chromatin. It is an interesting example of the increased specificity introduced by forming complexes. The kinase Ssn3 is brought into close contact with the transcription factor Gcn4p by complexing with RNA polymerase, leading to the removal of the transcription factor. Its activation is dependent on the cAMP dependent protein kinase A pathway and independent from the pheromone response that works via Kss1 (reviewed in Raithatha et al. 2012). This would be consistent with the lower activity of Gcn4, as indicated by the many down-regulated targets of the activator Gcn4 (Chapter 7). However, the degradation of Ste12 by this same mechanism, as well as of Phd1 (not in the kinase network) was not consistent with the proposed higher activities of Ste12 and Phd1 from enrichment statistics (Chapter 7).

The Pho85 gene also was constitutively expressed in the glucose and xylose medium cultivations, as evident from the RNA-seq data on *K. marxianus*, as were the vast majority of kinases. Similar to Ssn3, Pho85 phosphorylates Gcn4, leading to its ubiquitin-dependent proteolytic degradation [Shemer et al. 2002]. Whereas Srb10 marks chromatin-bound Gcn4p, the apparent role of Pcl5-Pho85 is to mark excess, free Gcn4p for degradation [Shemer et al. 2002]. Pho85 is a cyclin-dependent kinase, requiring the cyclins Pho80, Pcl1, Pcl5, Pcl6, Pcl9 or Pcl10. Pcl5 is specific for Gcn4 degradation [Shemer et al 2002]. The Pho85-specific cyclins Pho80, Pcl1, Pcl6, Pcl9 and Pcl10 were constitutively expressed. The Pcl5 gene, however, was five-fold down-regulated. This is in accord with the fact that Pho85 is under the transcriptional regulation of Gcn4p [Shemer et al. 2002]. The interaction of Gcn4 with the Pcl5 gene was also found in the likelihood network (Chapter 7). However, since Pcl5 was down-regulated five-fold, down-regulation of the Gcn4 targets was inconsistent with Pcl5-Pho85 leading to degradation of Gcn4.

The effects of Ssn3 and Pho85 on Gcn4 may be additive or synergistic, as they target free and chromatin-associated Gcn4, respectively [Chi et al. 2001]. Other mechanisms, apart from phosphorylation, might be at play. For instance, Pho85 is also known to be regulated at the level of translation by the kinase Gcn2 [reviewed in Hinnebush 1997].

Conclusions

The data analyses suggested Ssn3 and Pho85 are likely kinases regulating the transcription factors Gcn4 and Ste12, which both had enriched regulatory target gene sets. The direction of regulation of the targets of Ste12 and Gcn4, however, was not the same. Using RNA-seq data of the target genes of TFs, the small numbers of kinase-TF interactions unfortunately did not allow a rigid statistical enrichment approach to the study of kinases via their interactions with TFs. Direct physical evidence of kinase activity, such as phosphoproteomics, seems to be necessary to make further deductions about the activity of regulating kinases. Nevertheless, the method demonstrated in this chapter would come into its own when such data becomes available, and could be used in a context of multiple network types.

References

Buermans HPJ and den Dunnen JT. Next generation sequencing technology: Advances and applications. Biochim Biophys Acta. 2014;1842 :1932–1941.

Chi Y, Huddleston MJ, Zhang X, Young RA, Annan RS. Negative regulation of Gcn4 and Msn2 transcription factors by Srb10 cyclin-dependent kinase. Genes Dev. 2001;15: 1078-1092.

Collins SR, Kemmeren P, Zhao X, Greenblatt JF, Spencer F, Holstege FCP, Weissman JS, Krogan NJ. Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. Mol Cell Proteomics. 2007;6: 439-450.

Duda RO, Hart PE, Stork DG. Pattern classification. New York: J. Wiley & Sons; 2001.

Hengartner CJ, Myer VE, Liao S, Wilson CJ, Koh SS, Young RA. Temporal regulation of RNA polymerase II by Srb10 and Kin28 cyclin-dependent Kinases. Mol Cell. 1998;2: 43-53.

Hinnebush AG. Translational regulation of yeast GCN4. J Biol Chem 1997;272(35): 21661–21664.

Holstege FCP, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. Dissecting the regulatory circuitry of a eukaryotic genome. Cell. 1998;95: 717-728.

Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signaling circuits in molecular interaction networks. Bioinformatics. 2002;18(1): 233-240.

Nelson C, Goto S, Lund K, Hung W, Sadowki I. Srb10/Cdk8 regulates yeast filamentous growth by phosphorylating the transcription factor Ste12. Nature. 2003;421: 187-190.

Oliveira AP, Patil KR, Nielsen J. Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. BMC Syst Biol 2008;2(17). doi:10.1186/1752-0509-2-17.

O'Shea EK, Klemm JD, Kim PS, Alberr T. X-ray structure of the Gcn4 leucine zipper, a two-stranded, parallel coiled coil. Science. 1991;254(5031): 539-544.

Raithatha S, Su T, Lourenco P, Goto S, Sadowski I. Cdk8 regulates stability of the transcription factor Phd1 to control pseudohyphal differentiation of *Saccharomyces cerevisiae*. Mol Cell Biol. 2012;32(3): 664-674. doi:10.1128/MCB.05420-11.

Schabort DWP, Letebele PK, Steyn L, Kilian SG, du Preez JC. Differential RNA-seq, multi-network analysis and metabolic regulation analysis of *Kluyveromyces marxianus* reveals a compartmentalised response to xylose. PLoS ONE. 2016;11(6): e0156242. doi:10.1371/journal.pone.0156242.

Shemer R, Meimoun A, Hotzman T, Kornitzer D. Regulation of the transcription factor Gcn4 by Pho85 cyclin Pcl5. Mol Cell Biol. 2002;22(15): 5395-5404.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan GJ, van Baren M, , Salzberg SL, Wold BJ, Pachter L. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nat. Biotechnol. 2010;28(5): 511–515. doi:10.1038/nbt.1621.

van Dijk EL, Auger H, Yan Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30(9): 418-426. doi:10.1016/j.tig.2014.07.001.

Voet D, Voet JG. Biochemistry, fourth edition. New York: J. Wiley & Sons; 2011.

Chapter 8

Gene regulation in the context of chromosomes

Abstract

The finding that the RNA-seq data from *Kluyveromyces marxianus* strain UFS-Y2791 could be sufficiently mapped to a complete genome of strain DMKU3-1042 opened the way for analysis in the context of complete chromosomes. This chapter reports on an investigation of whether the species displays clusters of differentially expressed genes similar to the X and Y elements found near the telomeres, as is found in *Saccharomyces cerevisiae*. To the contrary, a small number of gene clusters which contained some of the most significantly differentially expressed genes encoding enzymes involved with alternative carbon source utilisation, were found further from telomeres. These clusters contain putative binding sites for the transcription factors Mig1 and Adr1. The inulinase gene INU1 was located in one of these clusters, among genes that are involved with the response to metal ions. Putative binding sites for the metal responsive transcription factor Aft1 were also found in the gene cluster containing INU1, suggesting that expression of this industrially important enzyme might be manipulated through control of metal ion concentrations.

Introduction

Previous chapters focused on network statistical approaches to analysing RNA-seq data. Several transcription factors were revealed to be important in regulating the differential response to glucose or xylose as the carbon source (Chapters 4, 5, 6). It was also found that it was possible to map the RNA-seq data from *Kluyveromyces marxianus* strain UFS-Y2791 to a complete genome of strain DMKU3-1042 (Addendum 1). This opened the way for improved construction of gene regulatory networks (Chapter 6). Also, it made it possible to consider differential expression in the context of chromosomes. It is known that in *Saccharomyces cerevisiae*, chromatin silencing and desilencing occurs in regions on the genome, especially close to the telomeres, known as X and Y elements which are enriched in transcription factor binding sites (Chapter 1, Smith et al. 2011). Chromatin silencing spreads from these regions and may silence several genes together. These regions are approximately six to seven kilobases from the telomeres [Smith et al. 2011]. In this chapter, RNA-seq data were mapped to chromosomes to investigate whether this phenomenon could be detected in *K. marxianus* or whether other, unusually long clusters of genes were differentially expressed. Further, the organisation of the genome in terms of the loci for genes belonging to some of the common metabolic functions identified as differentially expressed was investigated. These were the up-regulated genes of peroxisomal metabolism, the 2-methyl citrate cycle and sugar transporters.

Mig1 and Adr1 are two important transcription factors involved with glucose repression and derepression [Broach 2012]. Mig1 is a transcriptional repressor in *S. cerevisiae* [Nehlin et al. 1991]. It is a zinc finger protein that recognises a conserved GC box described as [GC][CT]GGGG. It was later shown that a flanking AT box was also important, thus the motif may be better described as "[ATG][AT][AT][AT][ATG].[GC][CT]GGGG" [Lundin et al. 1994]. It was also shown by Lertwattanasakul et al. [2011] that two putative binding sites for Mig1 existed in the regulatory region of the INU1 gene in *K. marxianus* DMKU3-1042 that are perfectly conserved in four other strains of *K. marxianus*. These sequences were TTAAATCCGGGG at bp 155 from the translation start site, and TTTTTCCTGGGG at 500 bp from the translation start site. Both of these match the combined AT box, GC-box consensus described by Lundin et al. [1994]. Adr1 is an activator of many genes involved in the utilisation of alternative carbon sources, especially those encoding enzymes involved in peroxisomal β-oxidation [Young et al. 2003]. The consensus pattern for Adr1 has been described as [TGA][TC]GG[AG]G [Cheng et al. 1994]. It usually binds as a dimer in opposite directions between two and 36 bp apart, and the more precise motif can be thus be described as C[CT]CC[GA][TCA]N{2-36}[TGA][TC]GG[AG]G, the reverse being identical. Based on the enumerative method of heptamer frequency comparison, it was

shown in a previous chapter that, firstly, Adr1 and, secondly, Mig1 were strong candidates as major regulators of up-regulated genes in a xylose medium which contained no glucose, thus representing glucose derepressed conditions (Chapter 4). RNA-seq data of the Mig1 and Adr1 transcripts also supported this role (Chapter 4). Motif searches were done to determine whether such binding sites may co-localise with any up-regulated gene clusters.

Materials and Methods

Strains and cultivation

K. marxianus UFS-2791 was cultivated in a defined mineral medium containing glucose or xylose in aerobic shake flasks. RNA was extracted in mid-exponential phase. Protocols were described in Chapter 3 and in Schabort et al. [2016].

RNA-seq and differential expression

RNA-seq reads from *K. marxianus* strain UFS-Y2791 from previous work [Schabort et al. 2016] were mapped to the complete genome of *K. marxianus* strain DMKU [Lertwattanassakul et al. 2015] using TopHat2 [Trapnell et al. 2009, Kim et al. 2013] in Galaxy [Afgan et al. 2015]. For simplicity of analysis in the protocol used here, the genome was annotated using Augustus and a gene model of *K. lactis* [Stanke et al. 2008] to obtain annotation tracks. Only annotation tracks 'gene', 'transcript' and 'CDS' were selected to serve as quantitation windows for differential expression testing in CuffDiff [Trapnell 2010]. Pileups were converted to intervals using the Pileup-to-Interval tool in SAM Tools [Li et al. 2009] as implemented in Galaxy.

Enrichment for stretches of likely chromatin regulation

An algorithm was developed for Reactomica [Schabort et al. 2016], implemented in the Wolfram Mathematica language for mapping differential expression to chromosomes and to map intervals from pile-ups to DNA for visualisation. The hypergeometric distribution is a discrete statistical distribution that calculates a probability of finding a certain number of consecutive successes (aces) in a certain sample size (cards drawn), given that in the total population (the size of the deck of cards) a certain known number of successes (four aces in a deck of cards) exists. The hypergeometric probability mass function is given by the formula below, consisting of binomial operators and obtainable in The Wolfram language as the function *HypergeometricDistribution*, where *N* is the

population size, *K* the number of successes in the population, *n* the number of draws and *k* the number of successes observed in *n* draws.

$$P(X = k) = \binom{K}{k}\binom{N-K}{n-k}/\binom{N}{n}$$

The cumulative distribution function of this formula can be used to calculate the background probability of obtaining at least *k* successes by means of the function

(1 - CDF[HypergeometricDistribution[n, K, N], k])

where the variable *k* represents the number of genes that were either up or down-regulated in a sample (two separate tests were performed). The variable *N* represents the total number of genes, *K* the total number of genes either up or down-regulated, and *n* the number of consecutive genes that were analysed and adjacent to one another on the genome. Correcting for multiple comparisons by multiplying the p-values with the total gene count (5 162) resulted in a very conservative estimate for assigning up-regulated gene clusters in the genome.

Motif searches

The dimeric binding model of Adr1 binding used was similar to that of Cheng et al. [1994] as a regular expression C[CT]CC[GA][TCA]N{2-36}[TGA][TC]GG[AG]G, the reverse being identical. The core recognition site (GC-box) for Mig1 was modelled as described as [GC][TC]GG[GA]G (and C[CT]CC[AG][GC] in the opposite direction). The more restrictive GC box, AT box pattern was modelled as [ATG][AT][AT][AT][ATG]N[GC][TC]GGGG and the reverse of this pattern as CCCC[GA][GC]N[TAC][AT][AT][AT][CAT], based on the binding sequences found by Lundin et al. [1994].

Results

Differential expression values from RNA-seq data of strain *K. marxianus* UFS-Y2791 [Chapter 3, Schabort et al. 2016] were mapped to the complete chromosomes of strain DMKU as annotated by Lertwattanassakul et al. [2015] and represented in Figure 1. In most chromosomes there seems to be a pattern in that several genes close to the telomeres have low expression levels under both conditions of glucose or xylose as carbon source. Further, regions on the chromosomes were visible that displayed sequences of either up-regulated or down-regulated genes, but were located far from the telomeres.

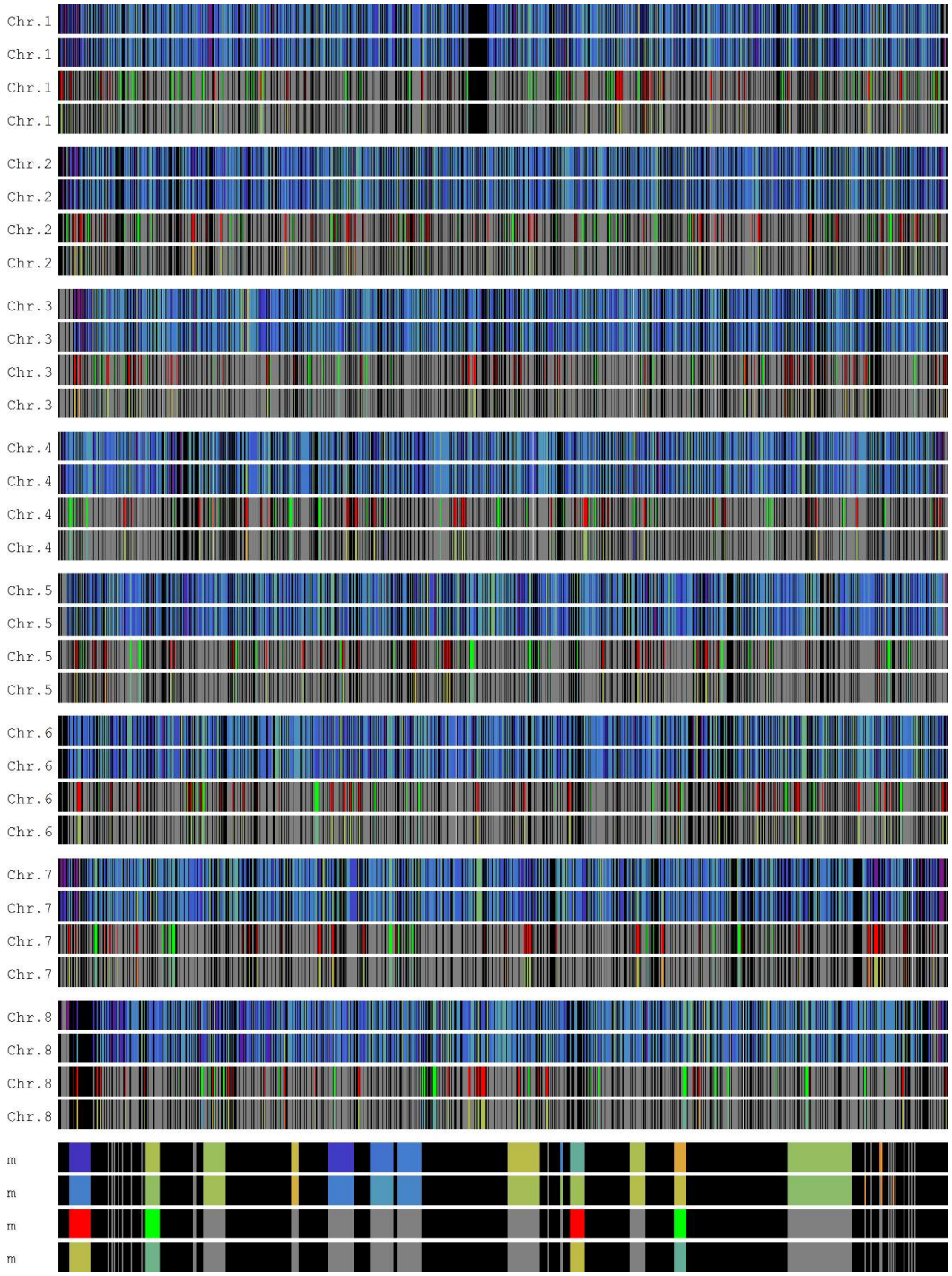


Figure 1. Visual representation of genomic chromosomes 1 to 8 and the mitochondrial chromosome mapped with RNA-seq data from *K. marxianus* UFS-2791 using the genome annotation by Lertwattanassakul et al. [2015]. Black represents intergenic regions. Track 1, normalised transcript levels with glucose as carbon source; Track 2, normalised transcript levels with xylose as carbon source. Warmer colours represent highly expressed genes with red indicating the highest; colder colours represent lowly expressed genes with violet indicating the lowest. Track 3, classifier up/down classifier scheme: red, up-regulated; green, down-regulated; grey, constitutively expressed. Track 4, log2(FC) scheme: warmer colours represent the highest positive fold changes, colder colours the highest negative fold changes and grey the constitutively expressed genes.

On initial inspection, two such gene clusters that were immediately visible on chromosome 1 were those between 1 Mbp and 1.2 Mbp, which were both up-regulated in the xylose medium compared to the glucose medium, in a surrounding setting of constitutively expressed genes (Figures 2 and 3). It was subsequently calculated whether these stretches contained more up-regulated genes than could be expected by a selection process that selected six genes at random positions. The calculated p-value indicated that the assignment was highly significant, even after severe correction for multiple comparisons in which a q-value was obtained (see Methods). This process also revealed more such regions (see below).

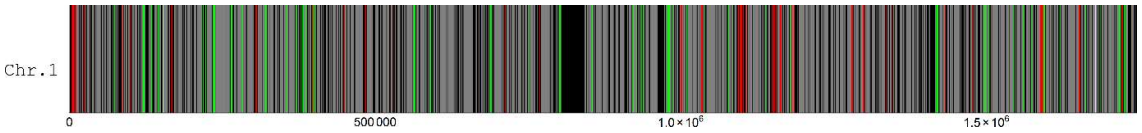


Figure 2. Visual representation of chromosome 1 mapped with RNA-seq data from *K. marxianus* UFS-2791. Notice the two clusters of up-regulated genes between 1.1 Mb and 1.2 Mb. Red, up-regulated; Green, down-regulated; Grey, constitutively expressed; Black, intergenic region.

The first gene cluster contained the inulinase gene INU1 on the 5' end, followed by ARN1 (siderophore iron transporter ARN1), KLMA_10520 (uncharacterised protein AN0679), FMP23 (Protein FMP23), ZTA1 (probable quinone oxidoreductase) and KLMA_10523 (uncharacterised protein). The second gene cluster contained three repeats of the putative high-affinity glucose transporter HGT1, as well as GLC3. These three paralogs likely originated via gene duplication.

In both of these clusters, it seems that the gene on the one periphery of the cluster was the most significantly up-regulated, with a decrease in the up-regulation further away from that side. This bears a resemblance to the mechanism of chromatin silencing which spreads from one locus to other loci [Chapter 1, Smith et al 2011].

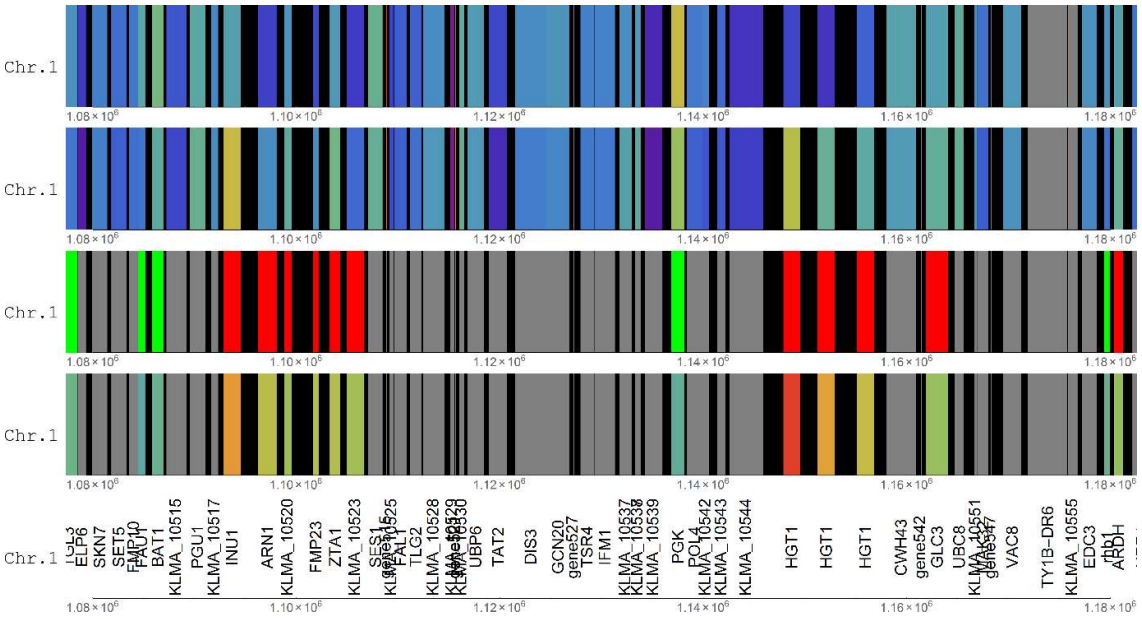


Figure 3. Visual representation of two clusters of up-regulated genes on chromosome 1. The probability of finding six out of six up-regulated genes (INU1, ARN1, KLMA_10520, FMP23, ZTA1, KLMA_10523 between 1 089 524 bp and 1 108 926 bp) by a random process was effectively zero. The probability for the up-regulation of four or more genes (HGT1, HGT1, HGT1 and GLC3) in a cluster of six genes (between 1 147 866 bp and 1 164 063 bp) by a random process was calculated at q = 0.027 after correction for 5 162 comparisons ($p = 5.3 \times 10^{-6}$).

Functional organisation and regulatory potential of the gene cluster containing inulinase

To gain insight into the functional organisation of the up-regulated gene cluster containing the biotechnologically important inulinase gene INU1, the functional annotation of the two uncharacterised proteins KLMA_10520 and KLMA_10523 was considered.

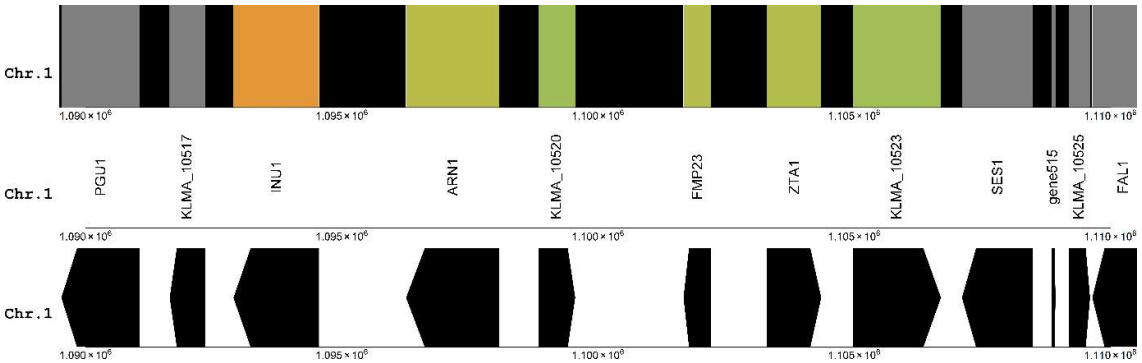


Figure 4. The gene cluster containing INU1, between 1.09 Mbp and 1.11 Mbp on chromosome 1, showing the direction of transcription.

KLMA_10520 has the GO terms "FMN binding [GO:0010181] and oxidoreductase activity [GO:0016491]" associated on the UniProt database. For KLMA_10523, a search on ProtoNet [Sasson et al. 2002] suggested that the protein may bind to the enzyme protein phosphatase type 1, based on structural similarity with other proteins in its protein structural cluster, from a variety of species, including GIP1 in *S. cerevisiae.* Several highly similar sequences were found by a BLASTP search, but all of these were to uncharacterised proteins. The best characterised was the GIP1 gene product, which is a meiosis-specific regulatory subunit of the Glc7 protein phosphatase [Tu et al. 1996]. Using BLASTP, GIP1 was found but at a low statistical significance with only a short region matching GIP1. DELTA-BLAST improved the query coverage to 35%, with improved statistical significance (E-value = 2e-05) to GIP1 (Figure 5). The matching region was 197 bp long, with a 40% similarity and 26% identity (Figure 6). A conserved domain or protein family was not reported by DELTA-BLAST.

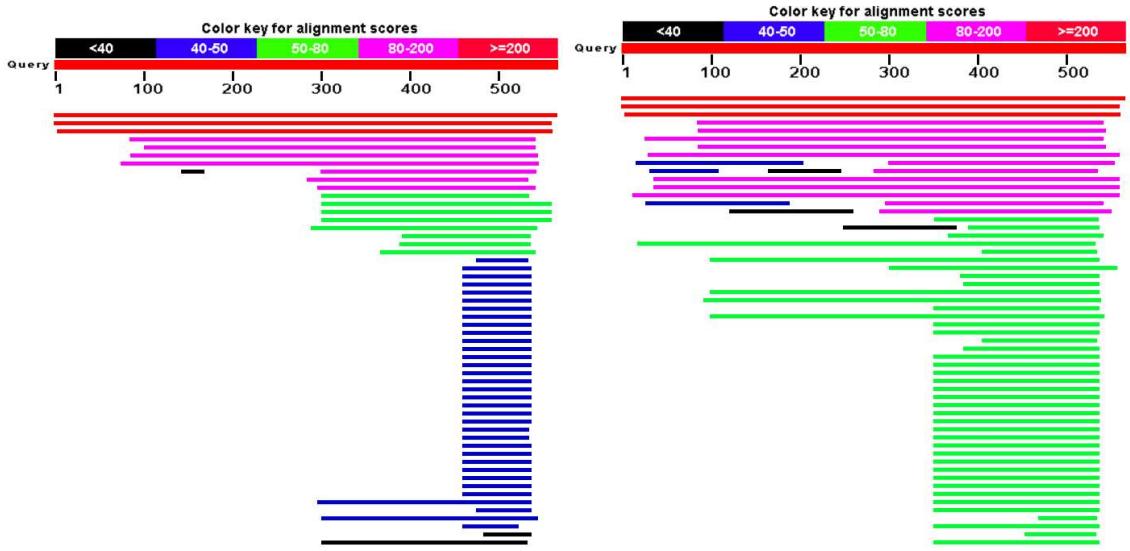


Figure 5. BLASTP and DELTA-BLAST matches to the KLMA_10523 amino acid sequence. A: GIP1 matched to 13% of the query (KLMA_10523) with an E-value of 0.006. B: GIP1 matched to 35% of the query (KLMA_10523), with an E-value of 2e-05.

Gip1p [Saccharomyces cerevisiae YJM1386]

Sequence ID: gb|AJQ14203.1| **Length:** 639 **Number of Matches:** 1

Range 1: 430 to 586 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
54.7 bits(130)	2e-05	Composition-based stats.	52/197(26%)	79/197(40%)	49/197(24%)

```
Query  354  DNEERSISSLEEYHMAENFSIRSKGFVQTGSIEHDQERE------RNLANMEVE---KGD  404
            DN    I S ++     NF  R+K F    +   D E E       N++ M+ +   K
Sbjct  430  DNLTHLIKSHQKKKRCVNF--RNKRFYDAFNPYVDNEEEAELSDSENISEMDTDLCIKDR  487

Query  405  KGDTVKFNKFSYLVIYDASKKCHYSESTDKLTLKIPHDGGSQTRAKIAVSALVNTETSLN  464
              +V+F++ S L+IY  SKK                                      LN
Sbjct  488  STSSVRFDENSRLLIYKKSKK--------------------------------------LN  510

Query  465  IEAESTPYSDSMRLKSILKRRTNEQESIEAQRARKCDEIDASDFLEFVENHENKRRSGED  524
             +   + YS ++  +SILK + N Q   E+QRA KCD +  + FL + +  E KR+  E
Sbjct  511  KDETQSGYS-TIETRSILKTKMNSQHDEESQRASKCDTVGVAQFLHYFQYTEYKRQRNEA  569

Query  525  ILVLARERQLKNYYDDQ  541
                R  QL  YY ++
Sbjct  570  ENYRLRGEQLSKYYSEE  586
```

Figure 6. Alignment of the KLMA_10523 amino acid sequence to that of GIP1 using DELTA-BLAST.

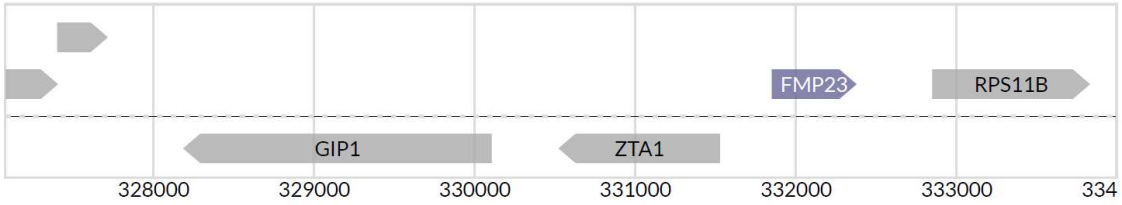


Figure 7. The region on chromosome II in *S. cerevisiae* that contains the FMP23, ZTA1 and GIP1 genes (from the UniProt database).

However, and perhaps most significantly, was the genomic context of the KLMA_10523 gene. The ZTA1 and FMP23 genes were both found to be present in *S. cerevisiae* as well, and translated in the same orientation as in *K. marxianus*. In both species, translation occurs away from each other where their coding regions are separated by a short intergenic region (compare Figures 4 and 7). The organisation of this region of the genome was thus conserved after the genome duplication and reshuffling events in the evolutionary past of the Saccharomycetes. In *S. cerevisiae*, GIP1 is found directly downstream of ZTA1 and translated in the same orientation as ZTA1. This was found to be exactly the situation in *K. marxianus* (compare Figures 4 and 7). Most probably, GIP1 and KLMA_10523 had a common ancestor. Following the same rationale, aligning KLMA_10521 with the RPS11B protein, the ribosomal subunit gene which is immediately downstream from FMP23 in *S. cerevisiae*, did not result in any significant similarity. ARN1 was found on chromosome VIII (Figure 8). Using instead the protein sequence of EFM1, which occurs immediately upstream of ARN1 (see Figure 8), matched best to YHL039W in *K. marxianus*, namely 906 968 bp to 908 632 bp on Chromosome 8 (AP012220.1), and

not on chromosome 1. The closest match to the inulinase gene INU1 of *K. marxianus* in *S. cerevisiae* was the invertase gene SUC2. In *S. cerevisiae,* SUC2 is found on chromosome IX (Figure 8).

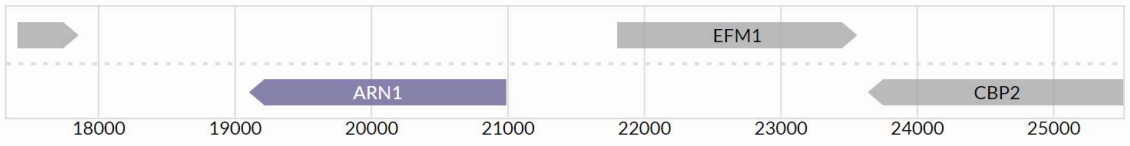


Figure 8. The region on chromosome VIII in *S. cerevisiae* that contains the ARN1 gene (from the UniProt database).

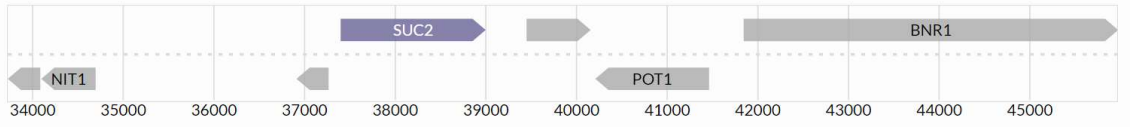


Figure 9. The region on chromosome IX in *S. cerevisiae* that contains the SUC2 gene, a homolog of the INU1 gene (from the UniProt database).

In summary, it is thus likely that FM23, ZTA1 and KLMA_10523 (GIP1) in *S. cerevisiae* originated from a single genomic region in the last common ancestor, and after the genome duplication event and subsequent extensive rearrangements was kept intact in *S. cerevisiae*, while ARN1 and KLMA_10520 were moved to other regions or lost, as is the case for INU1.

Notably, FM23 and ARN1 have both been implicated in the response to the concentrations of metal ions. It has been suggested that FMP23 (YBR047W) was involved with copper or iron balance, as in *S. cerevisiae* it is up-regulated in response to copper depletion and its regulatory region is enriched with binding sites for the Aft1 or Aft2 transcription factors [van Bakel et al. 2005]. Aft1 and Aft2 control a number of iron responsive genes [de Freitas at al. 2004]. ARN1 too is regulated by Aft1 [Yun et al. 2000]. This posed the question of whether the Aft1 and Aft2 binding sites possibly originated before the genome duplication event, which would then be present in *K. marxianus*, and concentrated over the INU1 containing gene cluster.

The exact consensus pattern for the Aft1 binding site in the regulatory region of FMP23, "GTGCACCC", did not result in any matches in the gene cluster investigated here. However, using the consensus motif “[TC]GCACC[TC]” from de Freitas at al. [2004] revealed two potential binding sites for FMP23 and ZTA1, and one that was in the region of the core promoter of ARN1, but may equally serve as a regulator of KLMA_10520 (Figure 10). In addition, the motif was also found at the region upstream of the inulinase gene INU1. In the *Saccharomyces* Genome Database (SGD), the motif is described as

"PyPuCACCCPu" ([TC][AG]CACCC[AG]), which is more specific. Using this pattern revealed that the binding site in the core promoter region of ARN1 matched the more specific pattern, as did the pattern closest to the ZTA1 start site (Figure 10, red lines). If these were true functional binding sites for Aft1, this cluster may be responsive to concentrations of metal ions. It may also have a direct application, as it could then suggest that inulinase production in this yeast might be further inducible by alterations in the concentration of metals such as iron or copper.

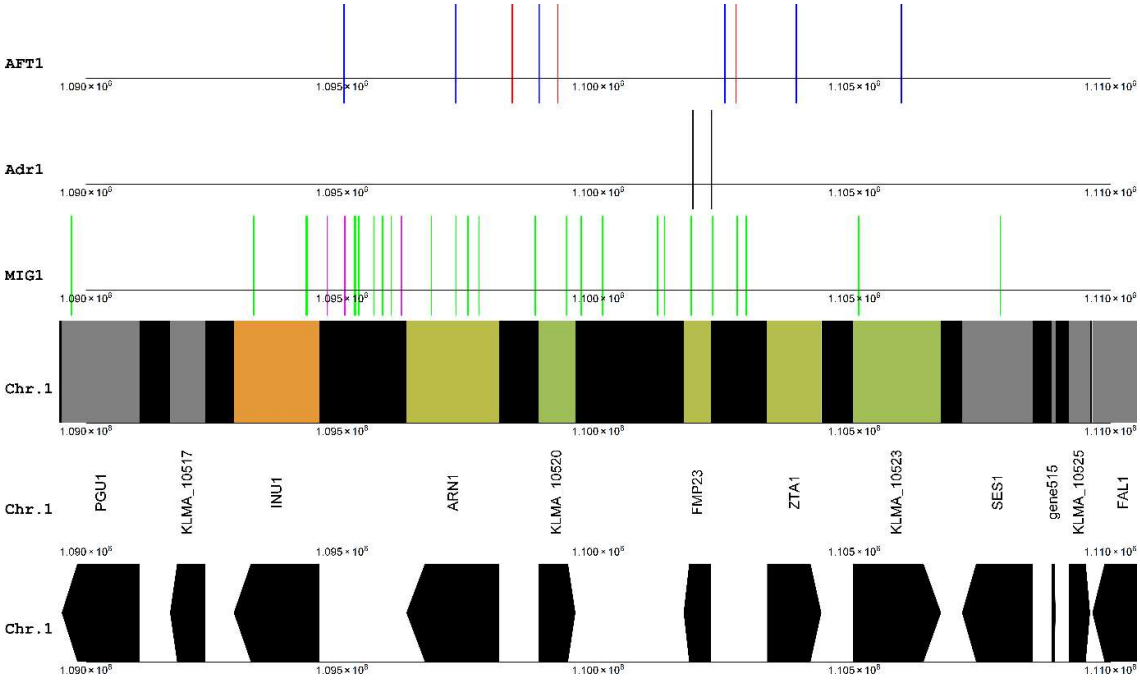


Figure 10. Candidate DNA binding sites for Adr1, Mig1 and Aft1/Aft2 in the vicinity of a gene cluster containing INU1, between 1.09 Mbp and 1.11 Mbp on chromosome 1. Black lines, candidate Adr1 binding sites using the dimeric binding model; Magenta lines, candidate Mig1 binding sites using the GC box, AT box model; Green lines, candidate Mig1 binding sites using the core GC box model; Blue lines, Aft1/Aft2 consensus motif [TC]GCACC[TC]; Red lines, Aft1/Aft2 consensus motif PyPuCACCCPu ([TC][AG]CACCC[AG]).

Since the alleviation of glucose repression is a likely explanation for the up-regulation of many genes in *S. cerevisiae*, it was expected that Adr1 and Mig1 may regulate this response (Chapter 4). The short and degenerate core binding site of Adr1 resulted in too many putative binding sites to be considered. The much more restrictive dimeric binding model revealed only two binding sites in the whole region, one of which was directly upstream of the FMP23 gene (Figure 10). Mig1 is also an important transcriptional repressor in *S. cerevisiae*. Both the GC box motif and the combined AT box, GC box motifs were scanned through the gene cluster. Three of the combined AT box, GC box motifs were

found in this region (Figure 10, magenta lines), of which two were the sites reported by Lertwattanasakul et al. [2011]. The other motif is in the same upstream intergenic region of INU1, directly downstream of the ARN1 gene. A larger number of the more degenerate GC box motifs was also detected (Figure 10, green lines), which appeared to be concentrated over the up-regulated gene cluster as opposed to the neighbouring regions. The regulation of this gene cluster may thus both be responsive to the concentration of the carbon source and to metal ions.

For the HGT1 gene cluster (Figure 11), the dimeric Adr1 binding sites were notably absent, except for one in between two of the HGT1 gene copies. The same HGT1 gene also had a candidate site for Aft1 binding. However, each of the three HGT1 copies had candidate binding sites for the GC box, AT box Mig1 model in their upstream regulatory regions, which were the only such sites in the whole region. Like with the INU1 cluster, candidate Mig1 sites were concentrated around the genes with the most significant up-regulation.

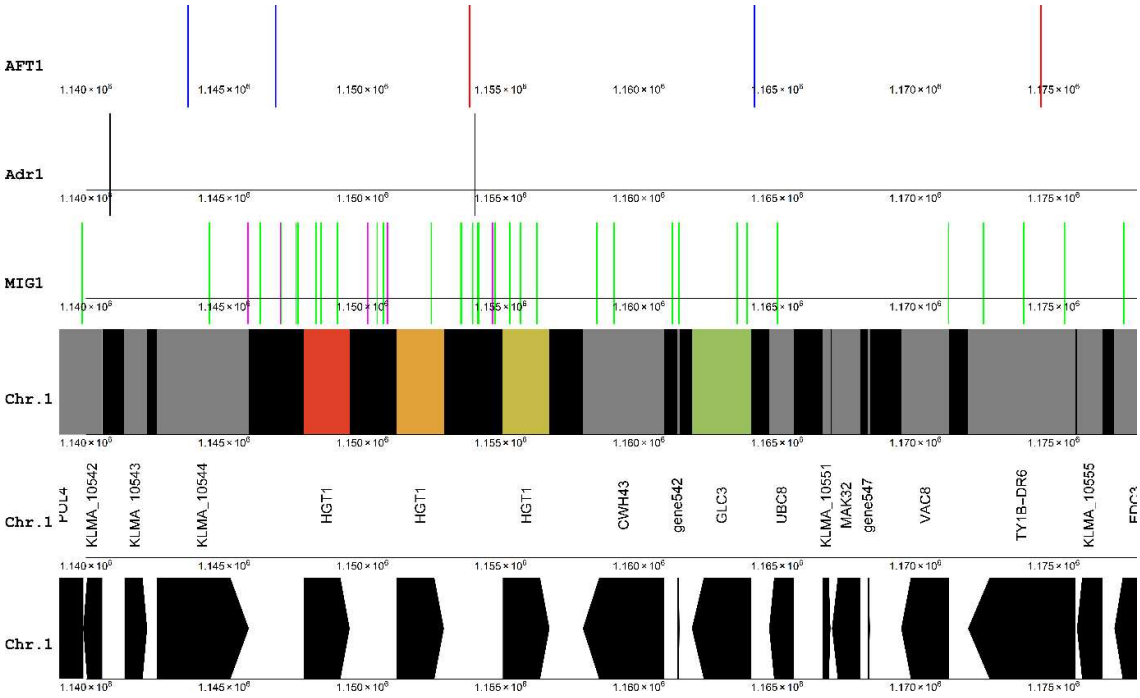


Figure 11. Candidate DNA binding sites for Adr1, Mig1 and Aft1/Aft2 in the vicinity of a gene cluster containing three copies of HGT1, between 1.14 Mbp and 1.17 Mbp on chromosome 1. Black lines, candidate Adr1 binding sites using the dimeric binding model; Magenta lines, candidate Mig1 binding sites using the GC box, AT box model; Green lines, candidate Mig1 binding sites using the core GC box model; Blue lines, Aft1/Aft2 consensus motif [TC]GCACC[TC]; Red lines, Aft1/Aft2 consensus motif PyPuCACCCPu ([TC][AG]CACCC[AG]).

Genes encoding for the up-regulated 2-methylcitrate cycle are up-regulated and co-localised

A gene cluster with four up-regulated genes (FOX2, KLMA_70428, ICL2 and KLMA_70430) was found on chromosome 7 (Figure 12). The FOX2 gene encodes one of the enzymes of β-oxidation, a pathway which was strongly up-regulated in the xylose medium, along with many other peroxisomal genes (Chapter 3, Schabort et al. 2016, Lertwattanassakul et al. 2015). No other peroxisomal genes were, however, found close to the FOX2 gene. Further investigation showed that peroxisomal genes were distributed among chromosomes. The up-regulated isocitrate lyase 2 gene, ICL2, was also found within the cluster. This ICL2 gene does not encode the glyoxylate cycle enzyme (Icl1), but instead encodes an isocitrate lyase isozyme belonging to the 2-methylcitrate cycle. All three genes encoding enzymes of the 2-methylcitrate cycle were strongly up-regulated in the xylose medium [Chapter 3, Schabort et al. 2016]. Strikingly, the other two genes of the 2-methylcitrate cycle, 2-methylcitrate dehydratase (PDH1) and citrate synthase 3 (CIT3) were found to be located close to the FOX2-containing gene cluster and adjacent to each other (Figure 12).

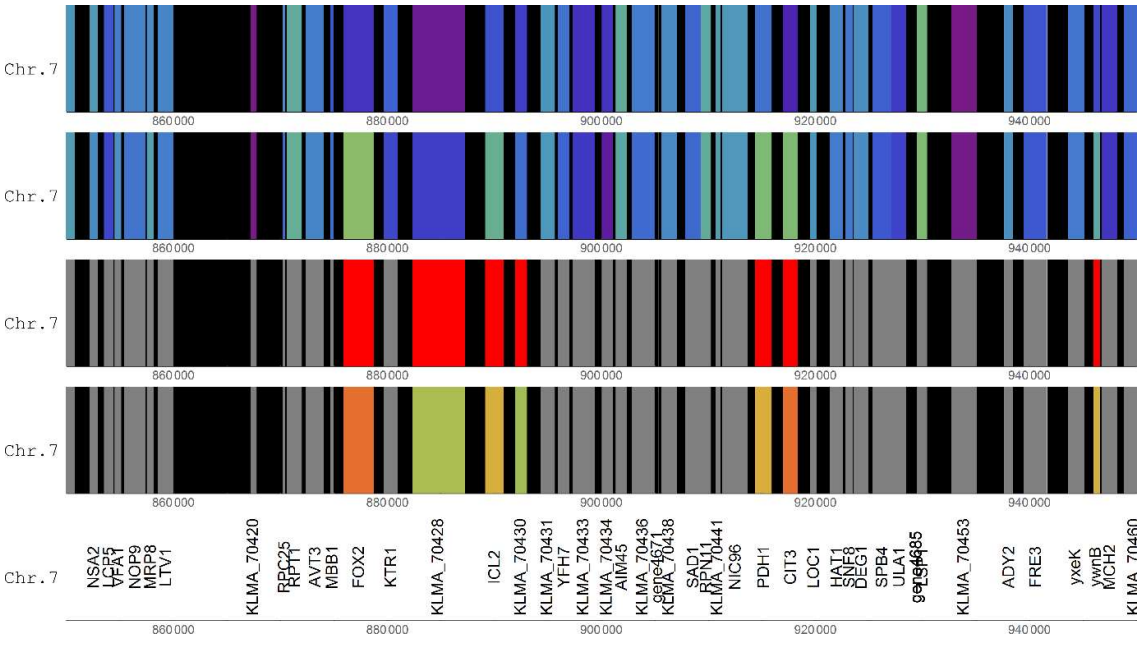


Figure 12. A cluster of up-regulated genes on chromosome 7. The probability for the up-regulation of four or more genes (FOX2, KLMA_70428, ICL2 and KLMA_70430) in a cluster of six genes (between 874 649 bp and 895 644 bp) by a random process was calculated at q = 0.027 after correction for 5162 comparisons ($p = 5.3\times10^{-6}$).

Using the dimeric binding model for Adr1 revealed one candidate binding site upstream of FOX2 (Figure 13), which is in accordance with data for *S. cerevisiae*, where FOX2 was shown to be under the regulation of Adr1 [Young et al. 2003]. Adr1 was also found to be the most likely candidate regulator of up-regulated genes based on the enumerative method of heptamer frequency comparisons [Chapter 4]. Four other candidate Adr1 sites were found in proximity of the gene cluster, but were either inside the coding regions or downstream of the genes. The GC box, AT box motif for Mig1 resulted in one candidate binding site upstream of the PDH1 gene, immediately downstream of CIT3.

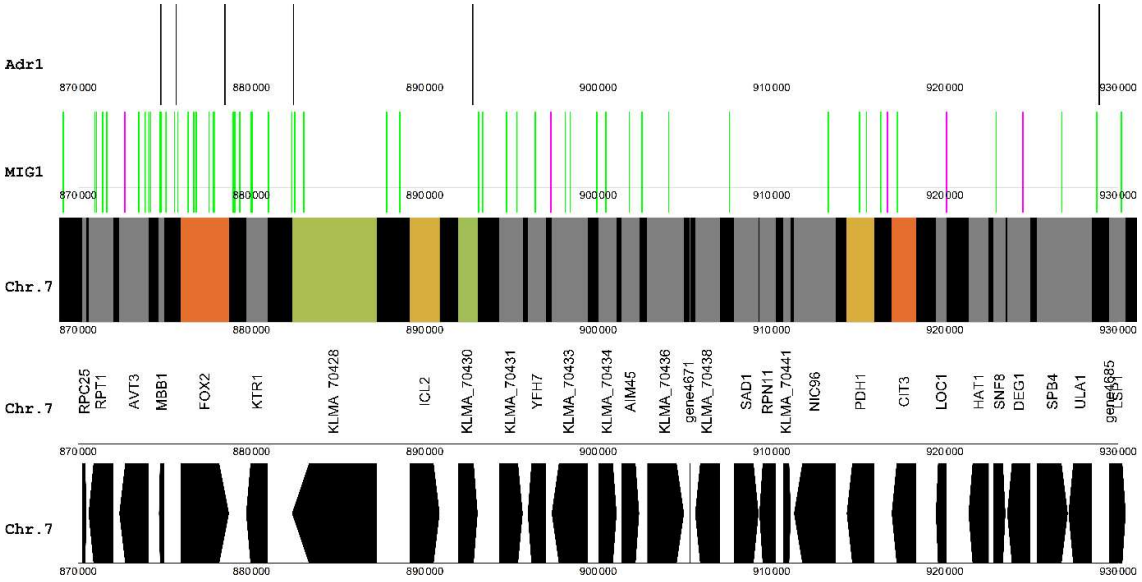


Figure 13. Candidate DNA binding sites for Adr1 and Mig1 in the vicinity of a gene cluster on chromosome 7 containing FOX2. Black lines, candidate Adr1 binding sites using the dimeric binding model; Magenta lines, candidate Mig1 binding sites using the GC box, AT box model; Green lines, candidate Mig1 binding sites using the core GC box model.

Discussion

The silencing of genes due to chromatin compaction and the reverse process occur in a cascaded fashion which spreads across parts of a chromosome [Chapter 1]. In this work, up-regulated gene clusters were detected that may be the result of desilencing due to glucose derepression in the xylose medium. This resembled the expression patterns of subtelomeric regions found in *S. cerevisiae* [Smith et al. 2011], but were not located close to telomeres. It seems that the most significantly up-regulated genes in these clusters were along the periphery of the clusters, with a gradual decrease in differential expression towards the other side. This may support a model of spreading of desilencing, especially considering a population of cells in which some cells show a more extensive form of de-silencing,

covering more of the genes in the cluster. Such cell-to-cell variation is known as stochastic noise, and may have various benefits such as faster adaptation to changing conditions by some cells (discussed in Chapter 1). Glucose repression involves transcription factors, including Adr1 and Mig1. These were also highlighted to be likely important in the differential response to glucose and xylose in *K. marxianus* [Chapter 4]. The mechanism of Mig1 repression is to recruit chromatin silencing proteins such as Tup1 [Treitel et al. 1995] leading to the silencing of chromatin. Thus the exact location of Mig1 sites, such as the distance from the transcription start site and the orientation of its binding, may be less important as opposed to those transcription factors that interact with RNAPII via the Mediator complex. Candidate GC box, At box Mig1 sites were found in the regulatory regions of these clusters, and their abundance also appeared to correlate with the most significantly up-regulated genes, notably for the INU1 containing cluster, the three HGT1 copy cluster and for the PDH1 gene. It was interesting no note that the highly up-regulated FOX2 gene of peroxisomal β-oxidation was found in a cluster together with the ICL2 gene of the up-regulated 2-methylcitrate cycle, and that the other two genes of this pathway were found to be up-regulated, located close by and adjacent to each other. Adr1, which is known to regulate peroxisomal genes in particular, was found to have a candidate site in the regulatory region of the FOX2 gene. It was also interesting to find that the INU1 gene encoding inulinase was found in a cluster of other up-regulated genes known to be associated with the response to metal ion concentrations in *S. cerevisiae*. Since this segment of the genome has been conserved in *S. cerevisiae* even after severe reshuffling of the genome in the evolutionary past, it is speculated that the same regulatory mechanisms originated in the earlier common ancestor. In accord with this notion, candidate binding sites for the metal responsive Aft1/Aft2 binding sites were found for these genes, as they also occur in *S. cerevisiae*.

Conclusions

The genomic context of complete chromosomes provided another route of exploration of RNA-seq data in the differential expression response to glucose and xylose as carbon sources in *K. marxianus* UFS-Y2791. The up-regulated gene clusters identified in this work presented and interesting perspective on gene regulation. Gene regulation of pathways such as the 2-methylcitrate cycle may be coordinated both by transcription factors and their localisation along the chromosome, in which a spread of chromatin silencing or desilencing could coordinate the regulation of a gene cluster. For others like the genes encoding peroxisomal β-oxidation, the coordination may be more dependent on common transcription factor binding sites. The finding of co-localisation of the INU1 gene with metal-responsive genes, as well as candidate binding sites for a metal responsive transcription factor, poses

an interesting perspective on the use of co-regulated gene clusters. Perhaps such knowledge could be used in a biotechnological scenario in which the alteration of metal ion concentrations, in particular copper and iron, may be used to further improve the production of inulinase, along with a non-fermentable carbon source to impose glucose derepression.

References

Afgan E, Sloggett C, Goonasekera N, Makunin I, Benson D, Crowe M, Gladman S, Kowsar Y, Pheasant M, Horst R, Lonie A. Genomics Virtual Laboratory: a practical bioinformatics workbench for the cloud. PLoS One. 2015;10(10): e0140829. doi: 10.1371/journal.pone.0140829.

Broach JR. Nutritional control of growth and development in yeast. Genetics. 2012;192: 73-105.

Cheng C, Kacherovsky N, Dombek KM, Camier S, Thukral SK, Rhim E, Young ET. Identification of potential target genes for Adr1p through characterization of essential nucleotides in UAS1. Mol Cell Biol. 1994;14(6): 3842-3852.

de Freitas JM, Kim JH, Poynton H, Su T, Wintz, Fox T, Holman P, Loguinov A, Keles S, van der Laan M, Vulpe C. Exploratory and confirmatory gene expression profiling of mac1Δ. J Biol Chem. 2004;279(6): 4450-4458.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg S. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4): R36. doi:10.1186/gb-2013-14-4-r36.

Lertwattanasakul N, Kosaka T, Hosoyama A, Suzuki Y, Rodrussamee N, Matsutani M, Murata M, Fujimoto M, Suprayogi S, Tsuchikane K, Limtong S, Fujita N, Yamada M. Genetic basis of the highly efficient yeast *Kluyveromyces marxianus*: complete genome sequence and transcriptome analyses. Biotechnol Biofuels. 2015;8(47). doi: 10.1186/s13068-015-0227-x.

Lertwattanasakul N, Rodrussamee N, Suprayogi LS, Thanonkeo P, Kosaka T, Yamada M. Utilization capability of sucrose, raffinose and inulin and its less-sensitiveness to glucose repression in thermotolerant yeast *Kluyveromyces marxianus* DMKU 3–1042. AMB Express. 2011;1(20): 1–11. doi: 10.1186/2191-0855-1-20.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25 (16): 2078–2079.

Lundin M, Nehlin JO, Ronne H. Importance of a flanking AT-rich region in target site recognition by the GC box-binding zinc finger protein MIG1. Mol Cell Biol. 1994;14(3): 1979-1985.

Nehlin O, Carlberg M, Ronne H. Control of yeast GAL genes by MIG 1 repressor: a transcriptional cascade in the glucose response. EMBO. 1991;10(11): 3373-3377.

Sasson O, Vaaknin A, Fleisher H, Portugaly E, Bilu Y, Linial N, Linial M. ProtoNet: hierarchical classification of the protein space. Nucleic Acids Res. 2003;31(1): 348-352.

Schabort DWP, Letebele PK, Steyn L, Kilian SG, du Preez JC. Differential RNA-seq, multi-network analysis and metabolic regulation analysis of *Kluyveromyces marxianus* reveals a compartmentalised response to xylose. PLoS ONE. 2016;11(6): e0156242. doi:10.1371/journal.pone.0156242.

Smith JJ, Miller LR, Kreisberg R, Vazquez L, Wan Y, Aitchison JD. Environment-responsive transcription factors bind subtelomeric elements and regulate gene silencing. Mol Syst Biol. 2011;7(455): 1-15. doi:10.1038/msb.2010.110.

Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics. 2008;24(5):637-644. doi: 10.1093/bioinformatics/btn013.

Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9): 1105-1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan GJ, van Baren M, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nat. Biotechnol. 2010;28(5): 511–515. doi:10.1038/nbt.1621.Wx`

Treitel, MA, Carlson M. Repression by SSN6-TUP1 is directed by MIG1, a repressor/activator protein. Proc Natl Acad Sci USA. 1995; 92: 3132-3136.

Tu J, Song W, Carlson M. Protein phosphatase type 1 interacts with proteins required for meiosis and other cellular processes in *Saccharomyces cerevisiae*. Mol Cell Biol. 1996;16(8): 4199-4206.

van Bakel H, Strengman E, Wijmenga C, Holstege FCP. Gene expression profiling and phenotype analyses of *S. cerevisiae* in response to changing copper reveals six genes with new roles in copper and iron metabolism. Physiol Genomics. 2005;22: 356-367.

Young ET, Dombek KM, Tachibana C, Ideker T. Multiple pathways are co-regulated by the protein kinase Snf1 and the transcription factors Adr1 and Cat8. J Biol Chem. 2003;278(28): 26146–26158.

Yun C, Ferea T, Rashford J, Ardon O, Brown PO, Botstein D, Kaplan J, Philpott CC. Desferrioxamine-mediated iron uptake in *Saccharomyces cerevisiae*. J Biol Chem. 2000;257(14): 10709-10715.

Chapter 9

Elucidation of new condition-dependent roles for fructose-1,6-bisphosphatase linked to cofactor balances

This paper was published online in PLoS ONE on 27 May 2017. See supporting information online at https://doi.org/10.1371/journal.pone.0177319**.**

Schabort DWP, Kilian SG, du Preez JC. Elucidation of new condition-dependent roles for fructose-1,6-bisphosphatase linked to cofactor balances.

Author contributions

Schabort DWP: Conceptualisation of investigation, writing of programmes, bioinformatics and analysis of RNA-seq data, writing of the manuscript.

Kilian SG: Review of manuscript.

du Preez JC: Editing and review of manuscript.

Abstract

The cofactor balances in metabolism is of paramount importance in the design of a metabolic engineering strategy and understanding the regulation of metabolism in general. ATP, NAD^+ and $NADP^+$ balances are central players linking the various fluxes in central metabolism as well as biomass formation. $NADP^+$ is especially important in the metabolic engineering of yeasts for xylose fermentation, since NADPH is required by most yeasts in the initial step of xylose utilisation, including the fast-growing *Kluyveromyces marxianus*. In this simulation study of yeast metabolism, the complex interplay between these cofactors was investigated; in particular, how they may affect the possible roles of fructose-1,6-bisphosphatase, the pentose phosphate pathway, glycerol production and the pyruvate dehydrogenase bypass. Using flux balance analysis, it was found that the potential role of fructose-1,6-bisphosphatase was highly dependent on the cofactor specificity of the oxidative pentose phosphate pathway and on the carbon source. Additionally, the excessive production of ATP under certain conditions might be involved in some of the phenomena observed, which may have been overlooked to date. Based on these findings, a strategy is proposed for the metabolic engineering of a future xylose-fermenting yeast for biofuel production.

Introduction

Recently, differential RNA-seq transcriptomics of *Kluyveromyces marxianus* was performed with glucose or xylose as the carbon source under aerobic conditions [1, 2]. It is to be expected that much of the differential response results from glucose derepression, as is the case with *Saccharomyces cerevisiae* in the absence of glucose, where the response is due to carbon source responsive transcription factors such as Adr1 and Mig1. Although the overall pattern of regulation in *K. marxianus* grown with D-xylose as carbon source instead of glucose resembled that of glucose derepression in *S. cerevisiae*, including the strong up-regulation of peroxisomal metabolism [1, 2], it did not represent a complete gluconeogenic response. For instance, the glyoxylate cycle was not up-regulated. Other signals are seemingly required for up-regulation of the glyoxylate cycle, which would be required if the cells were growing on acetyl-CoA arising from the β-oxidation of lipids. Some of the genes under glucose repression in *S. cerevisiae* have been dubbed “gluconeogenic” genes. Of particular interest is the fructose-1,6-bisphosphatase (FBP) reaction, catalysed by the Fbp1 protein and encoded by the FBP1 gene, which is under the control of the transcription factor Mig1 and glucose repression [3]. What distinguishes this gene from other genes subject to glucose repression is its centrality to energy

metabolism. If the FBP1 gene is expressed, it is likely to have an effect on energy metabolism, whereas unused transporters and β-oxidation would not have an effect if the carbon source were glucose or xylose. It is thus possible that the FBP1 gene might take on a different role, not related to gluconeogenesis, under some derepressed conditions. The RNA-seq data on *K. marxianus* [2] revealed that the FBP1 gene was up-regulated 27-fold when xylose served as carbon source. Moreover, it was recently discovered that in each of more than six hundred tumours of clear cell renal cell carcinoma analysed in humans, the expression level of the FBP1 gene was decreased [4]. Hence, a deeper understanding of the regulation and dis-regulation of the FBP reaction may as well be important for the treatment of cancer.

The up-regulation of the FBP1 gene in a recombinant *S. cerevisiae* strain fermenting xylose has been postulated as a mechanism to increase NADPH production by a type of cyclic pentose phosphate pathway (PPP) [5]. NADPH is also required for xylose utilisation by a yeast such as *K. marxianus*, in which xylose reductase utilises NADPH for reducing power [6, 1]. However, the FBP reaction could also have a direct influence on ATP levels. In mammalian muscle cells, both the FBP and phosphofructokinase (PFK) reactions act simultaneously in a substrate cycle, effectively dissipating free energy by synthesising and hydrolysing ATP. This mechanism allows a greater dynamic range in regulation of the flux, making use of adenylate kinase and AMP as a signal amplifier mechanism for dynamic responses to changes in ATP concentration [7, 8]. In other cells the substrate cycle could also serve as a heat generation method, as in the case of the flight muscles of bumblebees [9]. Heat generation is, however, unlikely to be a physiological response in yeasts, as is the signal amplifier mechanism that facilitates the dynamic response in glycolytic flux in mammals.

Another type of non-shivering thermogenesis is present in mammals, in which brown fat tissue catabolises acetyl CoA from lipids to generate heat. To avoid excessive ATP production, or equivalently, to avoid a limitation in ADP, uncoupling mechanisms exist. Hence, ATP can be excessive and metabolism is not always geared towards producing the maximal yield of ATP from a substrate. In fact, the uncoupling proteins found in mammals are an elaborate mechanism to avoid excessive ATP production. These allow the pumping of protons across the mitochondrial membrane, generating heat in the process and avoiding the need for protons to pass through the F_1F_0 ATPase. Thermogenin (UCP1) and similar proteins in mammals are the primary uncoupling proteins that allow the passage of protons [10]. However, to date, dedicated uncoupling proteins have not been found in yeasts, and UPC1-like activity in the yeast *Yarrowia lipolytica* is due to the promiscuous activity of an anion carrier protein [11, 12].

A number of questions arise from the above observations: (*a*) Could an experimental condition or a genetic manipulation exist that would cause an ATP imbalance by excessive production of ATP (and thus an ADP limitation) in a yeast such as *K. marxianus*, or is there always a sufficient demand for ATP by processes such as biomass formation? (b) If such a situation indeed existed and considering that dedicated uncoupling proteins apparently are absent in yeasts, would substrate cycles like those induced by the FBP and PFK reactions be good candidate replacement mechanisms for uncoupling proteins under such conditions? (*c*) Further, since the FBP reaction is such a central reaction, with links to both ATP as well as to NADPH (as it is associated with gluconeogenesis), what is the true purpose of up-regulation of FBP in *K. marxianus* in a xylose medium when there is no apparent need for the FBP reaction? The transaldolase and transketolase reactions of the non-oxidative PPP, which catalyse the route of carbon entry from xylose, indeed produce fructose-6-phosphate which is upstream of FBP in the glycolytic pathway. (d) Finally, might there be as yet unidentified genetic manipulations which could theoretically enable anaerobic xylose fermentation for bioethanol production, given the complexities imposed by cofactor balances and flux constraints? These complex questions call for a rigid mathematical modelling framework such as Flux Balance Analysis (FBA). In this report, the complex interplay between cofactor balances, major metabolic pathways and redox cofactor specificity was investigated using FBA as a predictive simulation framework.

Methods

The RNA-seq transcriptomic data were obtained as described elsewhere [2]. Briefly, *Kluyveromyces marxianus* strain UFS-2791 was cultivated at 35°C under aerobic conditions in a chemically defined medium containing glucose or xylose as the carbon source. The FBA model, capturing 56 reactions throughout central metabolism with reaction blocks for biomass formation, electron transport and oxidative phosphorylation, was described previously (see Schabort et al. [2], supplementary materials). The biomass formation formula was obtained from Fischer et al. [13] (See supplementary Table S5 in [2]). The phosphate/oxygen ratio (P/O ratio), which refers to the number of ADP to ATP conversions by ATP synthase per oxygen atom, was assumed to be 2.5 to simulate metabolism in Crabtree negative yeasts. The FBA simulation framework was described by Schilling et al. [14]. FBA was implemented in *Reactomica* using the Wolfram language. Flux constraints were defined by the stoichiometric matrix *S* and the exchange flux matrix *E* as below, which describe the mass balances of each metabolite as rows in the matrices.

$$Sv - Ee = 0 \quad (-\infty \leq v \leq \infty)$$

The intracellular flux vector v and the exchange flux vector e were calculated as a single vector, using optimisation with linear programming. In the above equation, *E* is an identity matrix that maps the

vector of exchange fluxes to metabolite balances, where non-zero entries were only present for metabolites that can cross the cell boundary, or those which were allowed to be produced in excess, such as ATP. The metabolites and cofactors that were allowed to accumulate are indicated in the Results section. The upper bounds for all intracellular fluxes were left unconstrained, while irreversible reactions were constrained to a lower flux value of zero (See supplementary Table S5 in [2]). The uptake flux of the carbon source was constrained to a value of 10 mmol h^{-1} g $biomass^{-1}$ specific flux. No substrate cycles, which could result in very high fluxes, were allowed. These were identified by Flux Variability Analysis [15], implemented also in *Reactomica*. The objective function was optimisation of the growth rate. Fluxes are interpretable in terms of their relative ratios compared to the molar uptake rate of the carbon source.

Results

The results of the flux balance analyses of *K. marxianus* grown on glucose and xylose as respective carbon sources are presented below. For clarity, a visual guide to the reversibility of reactions and the names of reactions is provided as supplementary material S1 Figures.

Is an increase in NADPH production the likely role for up-regulation of FBP1 in the xylose medium?

Simulations of aerobic metabolism in glucose and xylose media were initially performed. Figs 1 and 2 show the fluxes in the reference model in simulated glucose and xylose media, respectively. Note that for both simulations, the flux was limited by the same value for the sugar transporter at 10 mmol h^{-1} g^{-1}, specific to cell dry weight. Although the experimental xylose uptake and growth rates were approximately 50% of those for glucose, the power of FBA lies in calculating fluxes relative to a reference flux; in this case, the sugar uptake flux. Also, the measured production rates of ethanol and acetate were not included here as hard constraints, to facilitate exploring the theoretical limits of the model. The biomass formation rate in these simulations have arbitrary units and should be treated in a comparative manner among conditions. Throughout, fluxes and exchange rates were interpreted in a comparative sense and the units of mmol h^{-1} g^{-1} were omitted.

Note that the reversible glucose-6-phosphate isomerase reaction flux switches direction between the conditions of glucose and xylose utilisation under simulated aerobic conditions (Figs 1 and 2). These simulations revealed that, although glucose-6-phosphate isomerase had to operate in the

gluconeogenic direction when using xylose as an *in silico* carbon source, this was not the case for the PFK/FBP step and the FBP reaction did not become active when switching to the xylose *in silico* medium, in contrast to suggestions from literature that the FBP reaction was required to produce additional NADPH when xylose was the carbon source [5].

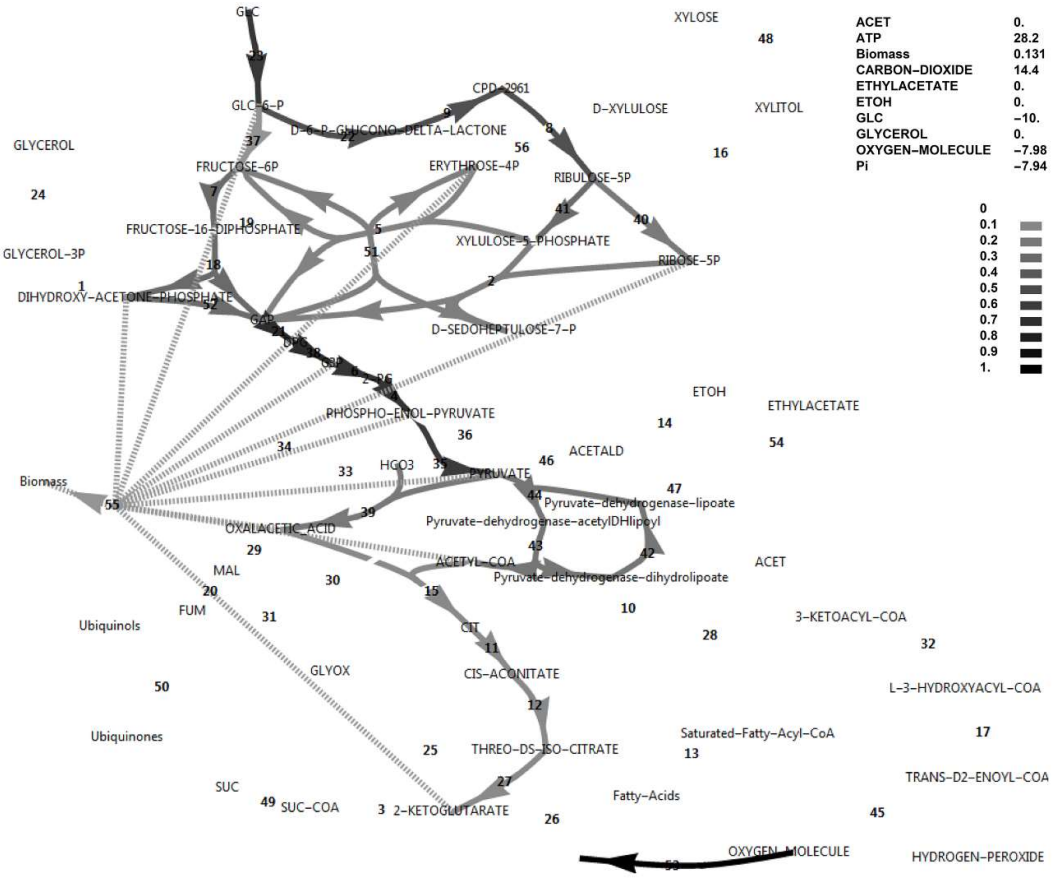


Fig 1. FBA simulation with glucose as *in silico* carbon source. Note that the glucose uptake flux was set to -10 and fluxes were interpreted relative to this flux as biomass specific fluxes, and that ATP is over-produced in this simulation at 28.2 as an exchange flux. This exchange flux represents reactions not directly accounted for in the model. Fluxes and exchange fluxes in these simulations are unitless, since they are determined by the upper bound of the carbon source uptake flux, set at a value of 10, and interpretation is done in the comparative sense.

Reaction names, followed by MetaCyc ID's are as follows (all maps of central metabolism contain the same annotations): 1, glycerol-3-phosphate dehydrogenase (NAD+) (1.1.1.8-RXN); 2, transketolase (1TRANSKETO-RXN); 3, 2-oxoglutarate dehydrogenase (2OXOGLUTARATEDEH-RXN); 4, phosphopyruvate hydratase (2PGADEHYDRAT-RXN); 5, D-fructose 6-phosphate:D-glyceraldehyde-3-phosphate glycolaldehyde transferase (2TRANSKETO-RXN); 6, phosphoglycerate mutase (3PGAREARR-RXN); 7, 6-phosphofructokinase (6PFRUCTPHOS-RXN); 8, phosphogluconate dehydrogenase (decarboxylating) (6PGLUCONDEHYDROG-RXN); 9, 6-phosphogluconolactonase (6PGLUCONOLACT-RXN); 10, acetate-CoA ligase (ACETATE--COA-LIGASE-RXN); 11, aconitate hydratase (ACONITATEDEHYDR-RXN); 12, aconitate hydratase (ACONITATEHYDR-RXN); 13, 2,3,4-saturated fatty acyl-CoA synthetase (ACYLCOASYN-RXN); 14, alcohol dehydrogenase (ALCOHOL-DEHYDROG-RXN); 15, citrate-S-synthase (CITSYN-RXN); 16, D-xylulose reductase (D-XYLULOSE-REDUCTASE-RXN); 17, enoyl-CoA hydratase (ENOYL-COA-HYDRAT-RXN); 18, fructose-bisphosphate aldolase (F16ALDOLASE-RXN); 19, fructose-bisphosphatase (F16BDEPHOS-RXN); 20, fumarate hydratase (FUMHYDR-RXN); 21, glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) (GAPOXNPHOSPHN-

RXN); 22, glucose-6-phosphate dehydrogenase (GLU6PDEHYDROG-RXN); 23, glucokinase (GLUCOKIN-RXN); 24, glycerol-1-phosphatase (GLYCEROL-1-PHOSPHATASE-RXN); 25, isocitrate lyase (ISOCIT-CLEAV-RXN); 26, isocitrate dehydrogenase ($NADP^+$) (ISOCITDEH-RXN); 27, isocitrate dehydrogenase (NAD^+) (ISOCITRATE-DEHYDROGENASE-NAD+-RXN); 28, acetyl-CoA-C-acyltransferase (KETOACYLCOATHIOL-RXN); 29, malate dehydrogenase (MALATE-DEH-RXN); 30, malate dehydrogenase (oxaloacetate-decarboxylating/malic enzyme) ($NADP^+$) (MALIC-NADP-RXN); 31, malate synthase (MALSYN-RXN); 32, 3-hydroxyacyl-CoA dehydrogenase) (OHACYL-COA-DEHYDROG-RXN); 33, phosphoenolpyruvate carboxylase (PEPCARBOX-RXN); 34, phosphoenol pyruvate carboxykinase (ATP) (PEPCARBOXYKIN-RXN); 35, pyruvate kinase (PEPDEPHOS-RXN); 36, pyruvate, water dikinase (PEPSYNTH-RXN); 37, glucose-6-phosphate isomerase (PGLUCISOM-RXN); 38, phosphoglycerate kinase (PHOSGLYPHOS-RXN); 39, pyruvate carboxylase (PYRUVATE-CARBOXYLASE-RXN); 40, ribose-5-phosphate isomerase (RIB5PISOM-RXN); 41, ribulose-phosphate 3-epimerase (RIBULP3EPIM-RXN); 42, dihydrolipoyl dehydrogenase (RXN0-1132); 43, dihydrolipoyllysine-residue acetyltransferase (RXN0-1133); 44, pyruvate dehydrogenase (acetyl-transferring (RXN0-1134); 45, Acyl-CoA oxidase (RXN-11026); 46, pyruvate decarboxylase (RXN-6161); 47, acetaldehyde dehydrogenase (NAD^+) (RXN66-3); 48, NADPH-dependent D-xylose reductase (RXN-8773); 49, succinate-CoA ligase (ADP-forming))SUCCCOASYN-RXN); 50, succinate dehydrogenase (ubiquinone) (SUCCINATE-DEHYDROGENASE-UBIQUINONE-RXN); 51, transaldolase (TRANSALDOL-RXN); 52, triose-phosphate isomerase (TRIOSEPISOMERIZATION-RXN); 53, electron transport (vETC); 54, ethyl acetate synthesis (vEthylAcetate); 55, growth reaction (biomass formation) (vGrowth). 56, xylulokinase (XYLULOKIN-RXN).

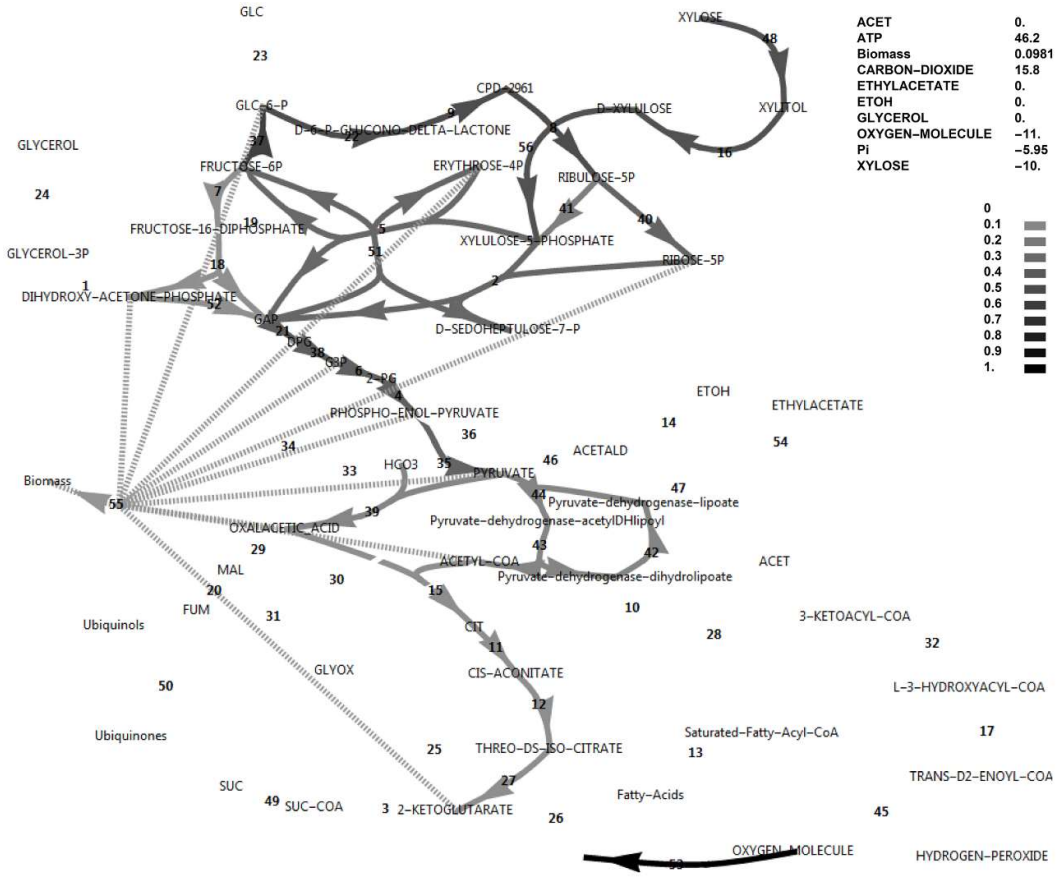


Fig 2. FBA simulation with xylose as *in silico* carbon source. Note that the xylose uptake flux was set to -10 and fluxes were interpreted relative to this flux as biomass specific fluxes, and that ATP is accumulated in this simulation at 46.2. The reaction names are as in Fig 1.

By optimising for excessive unbalanced NADPH production from xylose, instead of the growth rate, it was found that, in principle, two modes of cyclic PPP flux were possible (Figs 3 and 4). Activating the FBP reaction (Fig 3) resulted in three-fold the molar yield of NADPH on substrate compared to the model without it (Fig 4), with an NADPH balance of 90 versus 30.

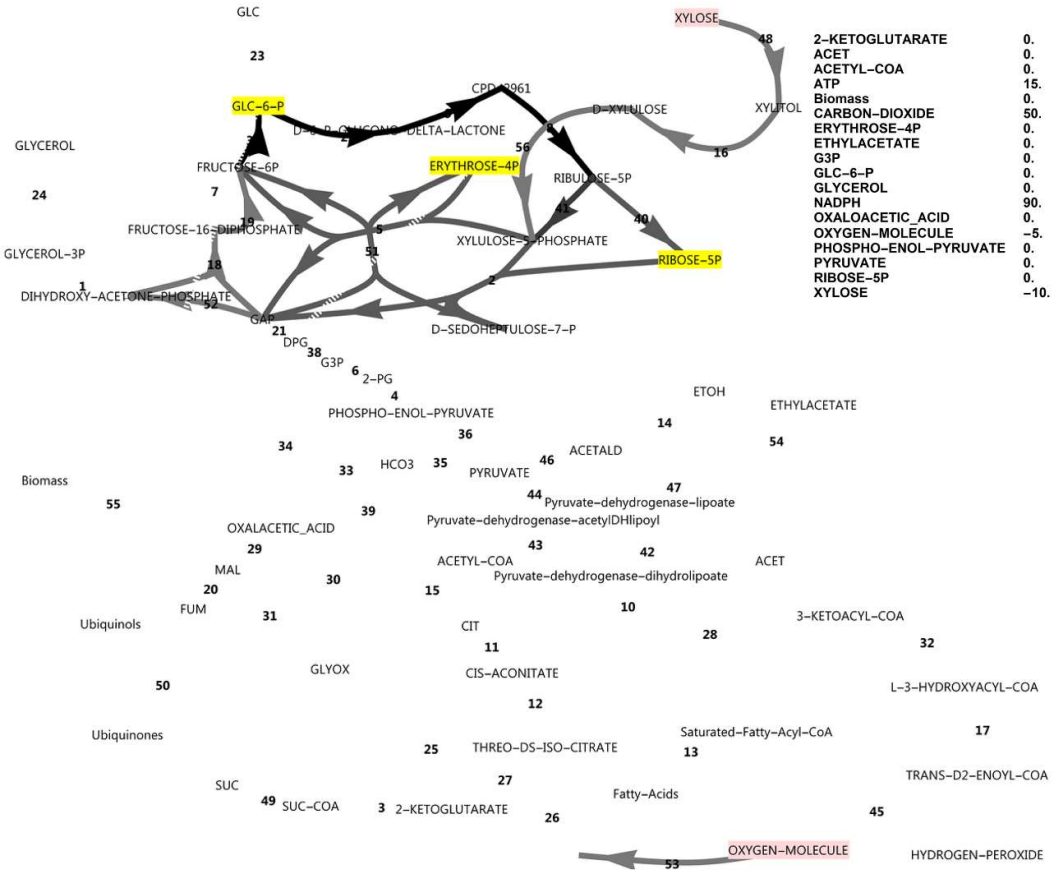


Fig 3. FBA simulation with xylose as *in silico* carbon source, inducing complete PPP cycling. Cofactors were allowed to exchange with the environment and FBP was active. Production of NADPH was optimised and complete cycling did not require glyceraldehyde-3-phosphate to accumulate. Nodes in pink indicate consumption and nodes in blue indicate production. Greyscale colours represent fluxes as a fraction of the highest flux in the simulation (black). None of the metabolites that were allowed to accumulate did accumulate (all yellow). The reaction names are as in Fig 1.

However, when optimising for biomass production with a closed NADPH balance while both the unidirectional FBP and PFK reactions were activated, the flux was always glycolytic via PFK from glucose-6-phosphate to fructose-1,6-bisphosphate, with no flux through FBP. Deactivating PFK and activating FBP (attempting to force a complete cyclic PPP flux, with a gluconeogenic direction) allowed steady state only when NADPH was allowed to accumulate in the model (Fig 5) concomitant with a

minor decrease in growth rate (0.0895 vs 0.0981), but FBP still did not carry flux. Null mutants of the PFK1 and PFK2 genes could thus be expected to have excessive reducing power in the form of NADPH when utilising xylose. This would manifest *in vivo* as a limitation in $NADP^+$ regeneration from NADPH. The NADPH imbalance would be further increased in a scenario of a 'gluconeogenic' net flux via aldolase and FBP towards glucose-6-phosphate, which produces three-fold more NADPH per xylose molecule (shown in Fig 3). Thus, in this model where two molecules of NADPH are produced via the oxidative PPP, up-regulation of the FBP1 gene cannot result in a higher growth rate by supplying more NADPH, but in fact has the opposite effect.

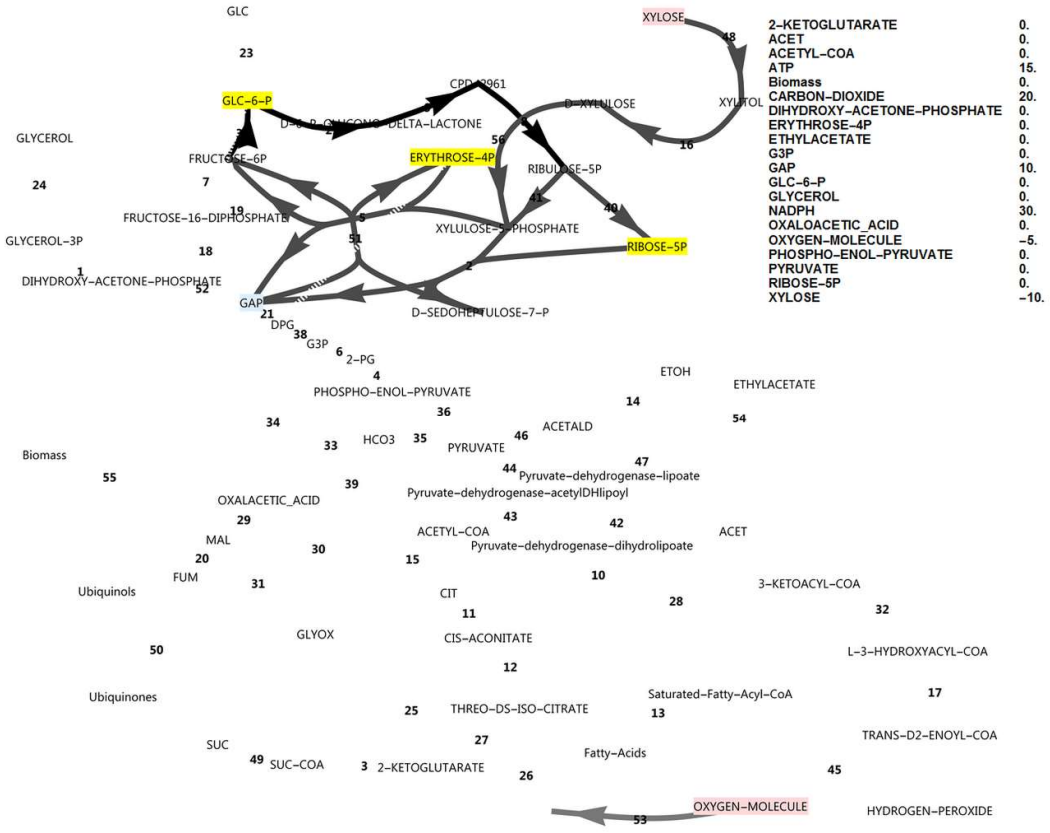


Fig 4. FBA simulation with xylose as *in silico* carbon source, inducing incomplete PPP cycling. Cofactor balances were open and FBP was inactive. Production of both NADPH and glyceraldehyde-3-phosphate were optimised. Nodes in pink indicate consumption, nodes in blue indicate production and nodes in yellow were allowed to accumulate, but did not accumulate. Greyscale colours represent fluxes as a fraction of the highest flux in the simulation (black). The reaction names are as in Fig 1.

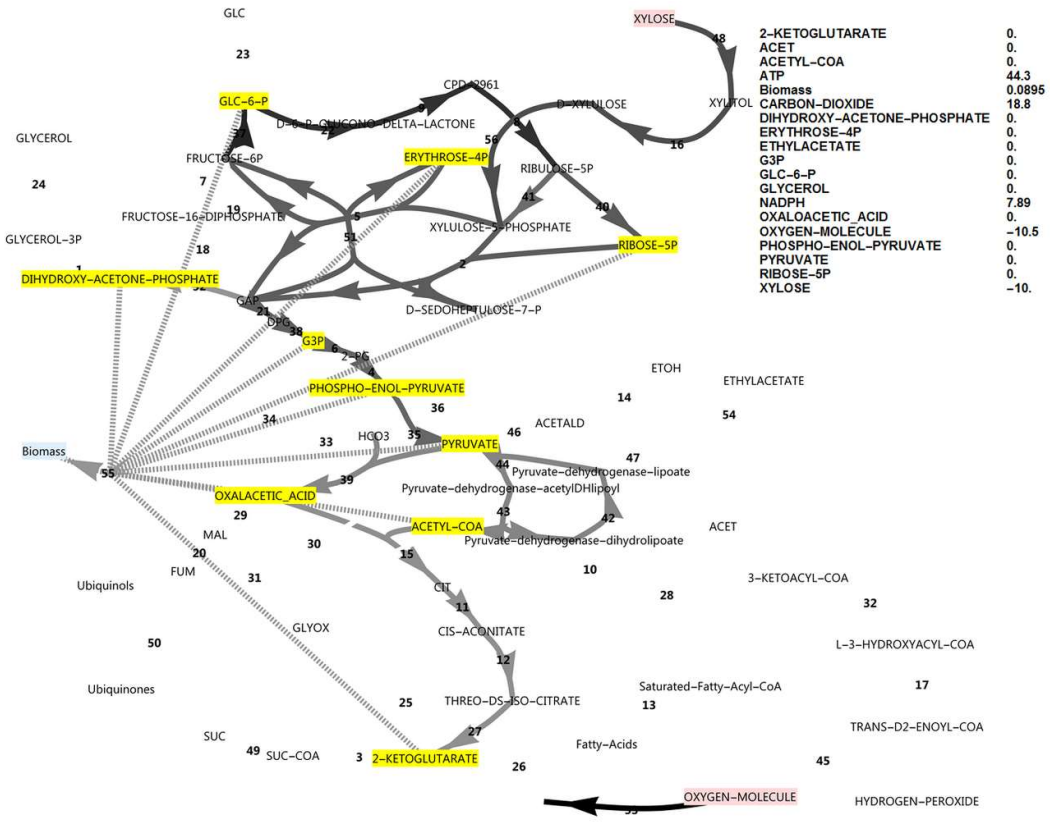


Fig 5. FBA simulation with xylose as the *in silico* carbon source, allowing overproduction of all biomass precursors as well as ATP and NADPH, with PFK inactive and FBP active. Note the absence of FBP flux and the substantial overproduction of only NADPH. Nodes in pink indicate consumption and nodes in yellow were allowed to accumulate, but did not accumulate. The reaction names are as in Fig 1.

An interesting observation was that intermediates such as glucose-6-phosphate would accumulate if the NADPH balance was closed and if both PFK and FBP were inactivated (Fig 6). The growth rate obtained in this scenario was somewhat lower than when the NADPH balance was open (0.0709 vs 0.0895). Thus, null mutants of the PFK1 and PFK2 genes may likely accumulate an intermediate such as glucose-6-phosphate during xylose utilisation, resulting in an overproduction of cell wall components, trehalose or glycogen. This finding could have an interesting biotechnological application. Nevertheless, this observation demonstrates that up-regulation of FBP in the absence of PFK cannot lead to a more balanced NADPH redox state and consequently a higher growth rate, as it would lead to a further excess of NADPH which may manifest as an accumulation of glucose-6-phosphate.

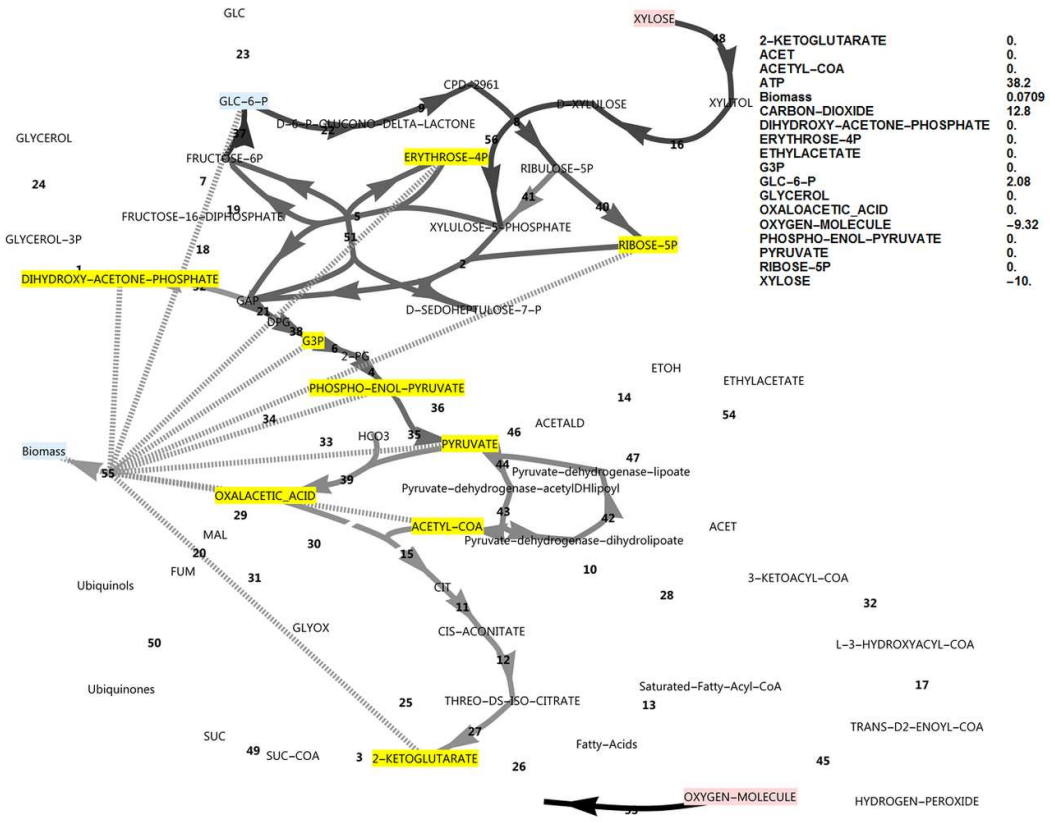


Fig 6. FBA simulation with xylose as *in silico* carbon source allowing overproduction of all biomass precursors as well as ATP, while closing the NADPH balance. PFK was inactive and FBP was active. Note the absence of FBP flux and the overproduction of glucose-6-phosphate. The reaction names are as in Fig 1.

A cell may respond the constraint of NADPH overproduction, or $NADP^+$ limitation, in one of several ways. Transdehydrogenases that oxidise NADPH while reducing NAD^+ would be one such mechanism, but transdehydrogenases are absent in yeasts in general, except for *Pichia angusta* [16]. Conversely, trans-oxidation-reduction reaction cycles may occur in which NADPH is in effect exchanged for NADH in a cyclic pathway, might occur. Also, the cofactor preference of the oxidative PPP for $NADP^+$ or NAD^+ would affect the NADPH balance and, therefore, the potential role of FBP. The cofactor preference of oxidative PPP enzymes is known to vary among yeast species. The effect of cofactor preference in the oxidative PPP and the possibility of trans-oxidation-reduction reaction cycles in other parts of metabolism are explored in a later section. However, tightly linked to the redox cofactor balances is the ATP balance**.** These are linked not only via the oxidative metabolism in mitochondria, but also via the flux constraints in central metabolism, in which a given pathway may produce or consume ATP, NADH, NADPH and their partners in various ratios. Given that the total uptake rate of the carbon flux is constant in these simulations, alternative catabolic pathways with different cofactor stoichiometry in effect could cause a negative correlation between the production rates of different cofactors.

Focussing still on upper central metabolism and the central role of FBP, the role of the ATP balance in flux routing is explored next.

The ATP excess hypothesis and new roles for the FBP reaction and glycerol production

In the simulations discussed above, the ATP balance was not closed and ATP was allowed to accumulate. This exchange flux of ATP represents reactions that hydrolyse ATP and that are not included in the growth reaction and not accounted for in the modelled system. These may include the energy requirement for the synthesis and degradation cycles of biopolymers, the action of active transporters or unidentified metabolic substrate cycles. On both substrates, a substantial positive ATP exchange flux was calculated, but notably the exchange flux of ATP was 64% higher in the xylose *in silico* medium compared to the glucose *in silico* medium (46.2 vs 28.2, Figs 1 and 2). Using xylose as carbon source, ATP supply was thus less likely to be growth rate limiting than when glucose was the carbon source, which is counter-intuitive. Also note that if the sugar transporter was to be changed to an active transporter, the effect would be negligible under these high ATP-yielding aerobic simulated conditions. Notably, it was found that closing the ATP balance caused a drastically decreased growth rate (0.0379 vs 0.0983), accompanied by both glycerol production and a flux through the pyruvate dehydrogenase bypass (PDB) (Fig 7). In this regard, glycerol production is a strategy to avoid NADH production in lower glycolysis and subsequent ATP synthesis resulting from electron transport, and is not due to a limited electron acceptor activity for regenerating NAD^+, as is observed in *S. cerevisiae* under conditions supporting anaerobic growth [17]. The PDB hydrolyses two ATP equivalents through the acetate-CoA ligase step, whereas aldehyde dehydrogenase may produce one of either NADH or NADPH. Activation of the PDB led to a small increase in the growth rate but was non-essential for *in silico* growth, whereas glycerol production was essential.

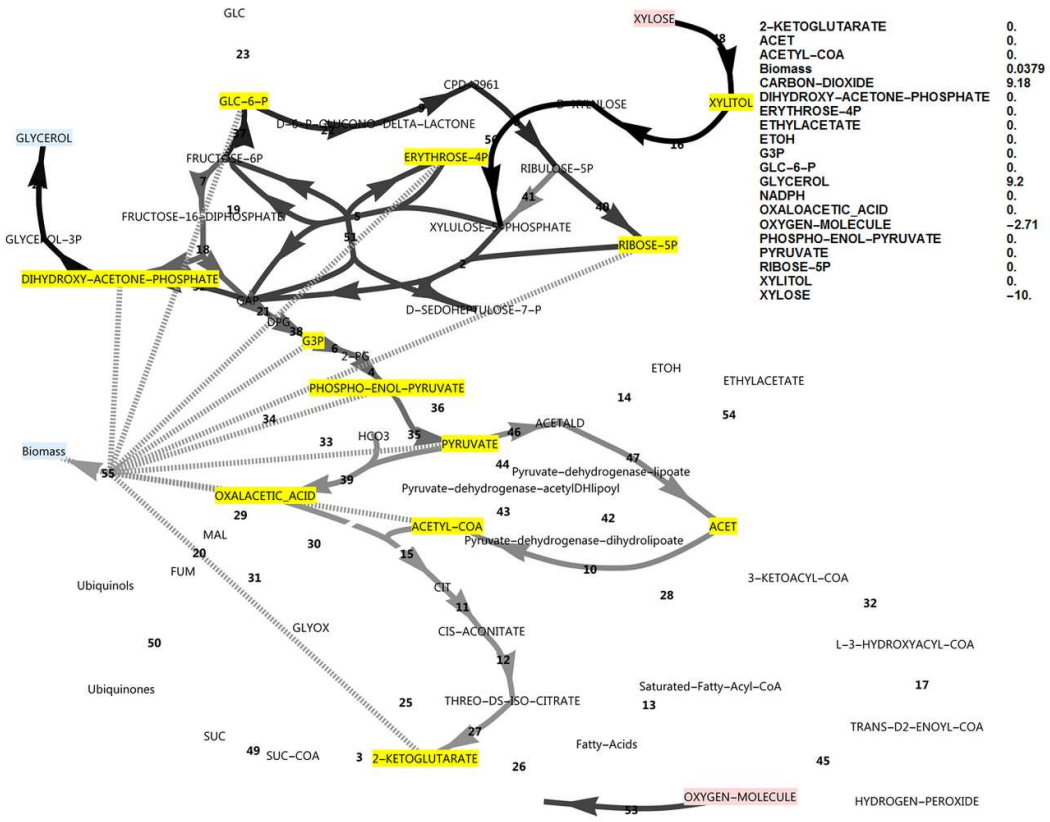


Fig 7. FBA simulation with xylose as *in silico* carbon source, allowing overproduction of all biomass precursors as well as NADPH, while closing the ATP balance. FBP activity was inactive while PFK was active. Note the production of glycerol and the appearance of flux in the pyruvate dehydrogenase bypass via acetate. The reaction names are as in Fig 1.

This state thus resembles phenotypically, an oxygen limited growing phenotype which produces glycerol, involving a low oxygen consumption rate (-2.71). This effect is however induced as an ATP avoidance strategy. *In vivo*, it would manifest as a limitation in ADP, which may have its effect via enzyme kinetics on various enzymes.

Dedicated mechanisms in the electron transport chain for uncoupling ATP synthesis in yeasts have not been reported to date. Therefore, other mechanisms may be required to deal with the higher relative ATP production from xylose. The FBP reaction may perform exactly such a function by inducing an ATP hydrolysing substrate cycle together with PFK. Like the other glycolytic enzymes, the PFK1 and PFK2 genes were down-regulated in the xylose medium (Table 1). Notably, while the two subunits of PFK1 and PFK2 were down-regulated from 1640.1 and 1856.4 FPKM (fragments per kilobase per million reads) to 341.0 and 395.4 FPKM, respectively, FBP1 was up-regulated from 13.7 to 374.6 FPKM. Thus, the mRNA levels of the PFK and FBP genes were brought from extremely different levels to very similar

levels. The presence of highly similar expression levels of genes encoding enzymes has indeed been observed in cases of confirmed substrate cycling, such as in mammalian muscle cells and bumblebees [7, 9]. Also, based on the striking similarity between simulated flux patterns and RNA-seq levels of genes in central metabolism [2], it seems that such deductions could be made in a pragmatic sense, linking the accurate transcript abundance levels from RNA-seq to protein levels and, ultimately, to an approximation of fluxes in the comparative sense.

Table 1. Relative expression levels (in FPKM) of mRNA in glucose and xylose media as determined by RNA-seq.

Gene name	Glucose	Xylose	Fold change from glucose to xylose
PFK1	1640.1	341	0.21
PFK2	1856.4	395.4	0.21
FBP1	13.7	374.6	27.34

Activating both the FBP and PFK activities and balancing ATP restored the rapid growth rate and avoided glycerol formation (Fig 8). The PFK/FBP substrate cycle induced here served as an alternative ATP sink in the absence of the ATP overproduction flux, representing ATP uncoupling, ATPases, or ATP-dependent membrane transporters. In this situation, activation and de-activation of the PDB made no difference to the growth rate or metabolite balances. The differential transcriptomic response of an engineered *S. cerevisiae* strain to xylose under anaerobic conditions was recently determined, which also showed a four-fold up-regulation of FBP1 in the presence of xylose [18]. Details of this substrate cycle can be seen in the supplementary material S1 Figures.

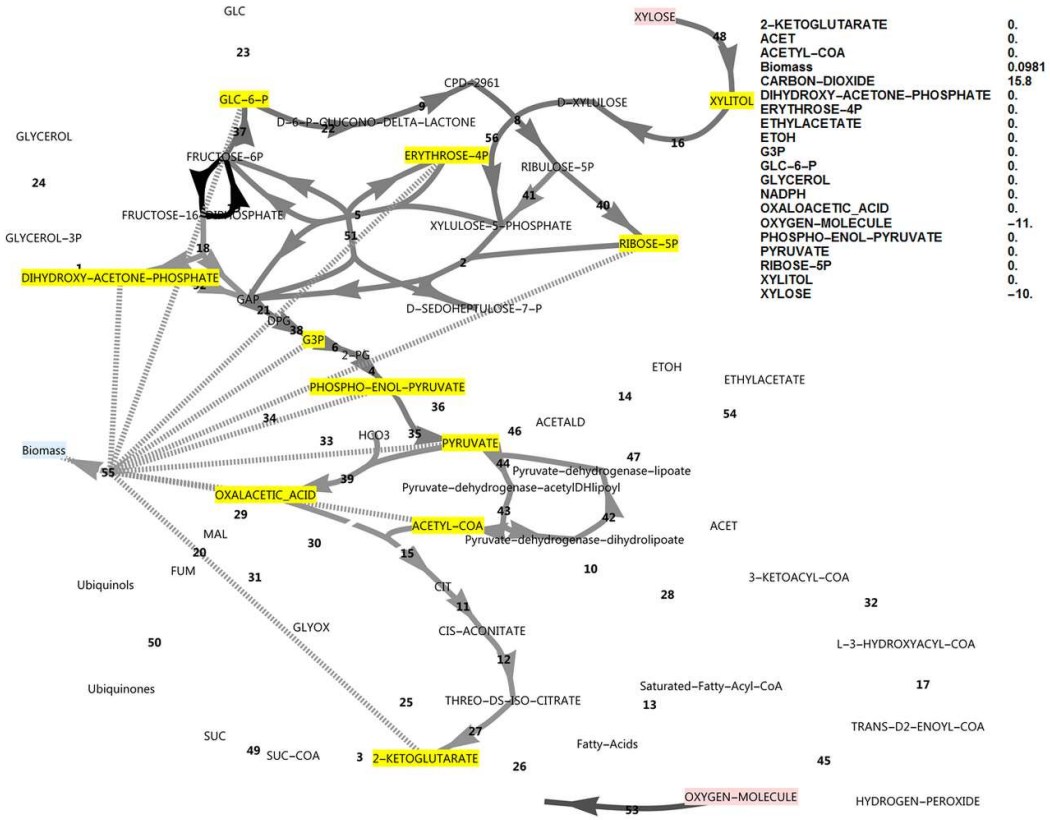


Fig 8. FBA simulation with xylose as *in silico* carbon source allowing overproduction of all biomass precursors as well as NADPH, while closing the ATP balance and with both FBP and PFK reactions active. Note the high growth rate, the absence of glycerol production and pyruvate dehydrogenase bypass fluxes, and the FBP/PFK substrate cycle that is responsible for the balance in ATP and ADP. The reaction names are as in Fig 1. The PFK/FBP substrate cycle serves as an alternative ATP sink in the absence of the ATP exchange flux, which represents various mechanisms of ATP utilisation.

Effect of cofactor specificity of oxidative PPP enzymes

Although data on the subject of redox cofactor specificity is scarce, it has been found that in a number of bacteria [19, 20], the oxidative PPP enzymes such as glucose-6-phosphate dehydrogenase are not exclusively specific to $NADP^+$, but can also use NAD^+. In an attempt to obtain an increased cyclic PPP flux in the model, the cofactor specificity of glucose-6-phosphate dehydrogenase was changed in the model from $NADP^+$ to NAD^+ while allowing all biomass precursors and ATP to accumulate. As there was a deficiency of NADPH production in this model, cyclic PPP flux through FBP was observed as well as a higher growth rate (0.0777, Fig 9) than when FBP was absent (0.0556, Fig 10). A further increase in ATP exchange flux was present in this cyclic PPP mode (85.3, Fig 9), which was nearly double that of the reference model with a glucose-6-phosphate dehydrogenase specificity for $NADP^+$ (46.2, Fig 2), and three-fold that of the initial model with glucose as carbon source (28.2, Fig 1). Thus, by allowing

oxidative PPP enzymes to utilise NAD^+ instead of $NADP^+$, the FBP reaction may adopt a dual role. It would allow increased production of NADPH by the oxidative PPP, but at the same time this cyclic flux induces an increased overproduction of ATP, requiring the presence of the FBP/PFK substrate cycle - the second role of the FBP reaction. Absence of both FBP and PFK results in glycerol production, a lower growth rate and a substantial ATP overproduction (Fig 10). Although PDB can hydrolyse some of the excessive ATP, it cannot be compared to the potential of a substrate cycle to hydrolyse ATP, as it cannot carry a flux higher than 50% of the flux in lower glycolysis.

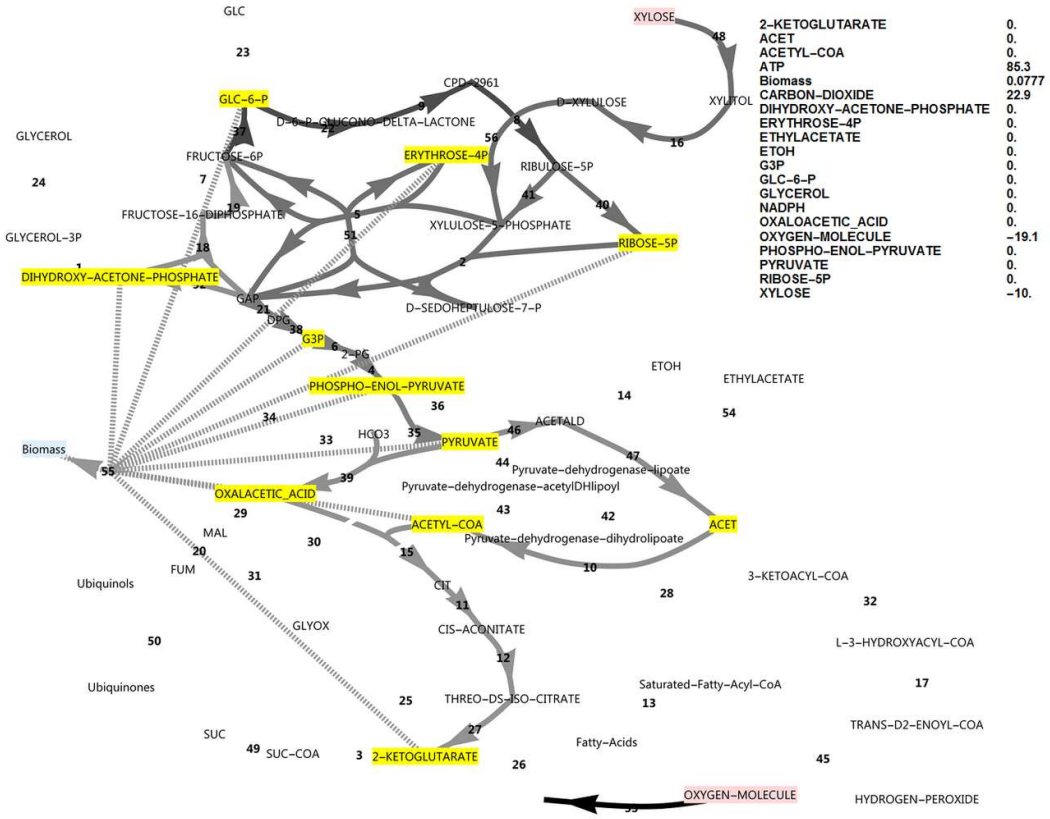


Fig 9. FBA simulation with xylose as *in silico* carbon source, assuming that the oxidative PPP produces one NADPH and one NADH. All biomass precursors as well as NADPH and ATP were allowed to accumulate and with FBP active and PFK inactive. Note the complete cyclic PPP flux, a large ATP exchange flux and the PDB flux. The reaction names are as in Fig 1.

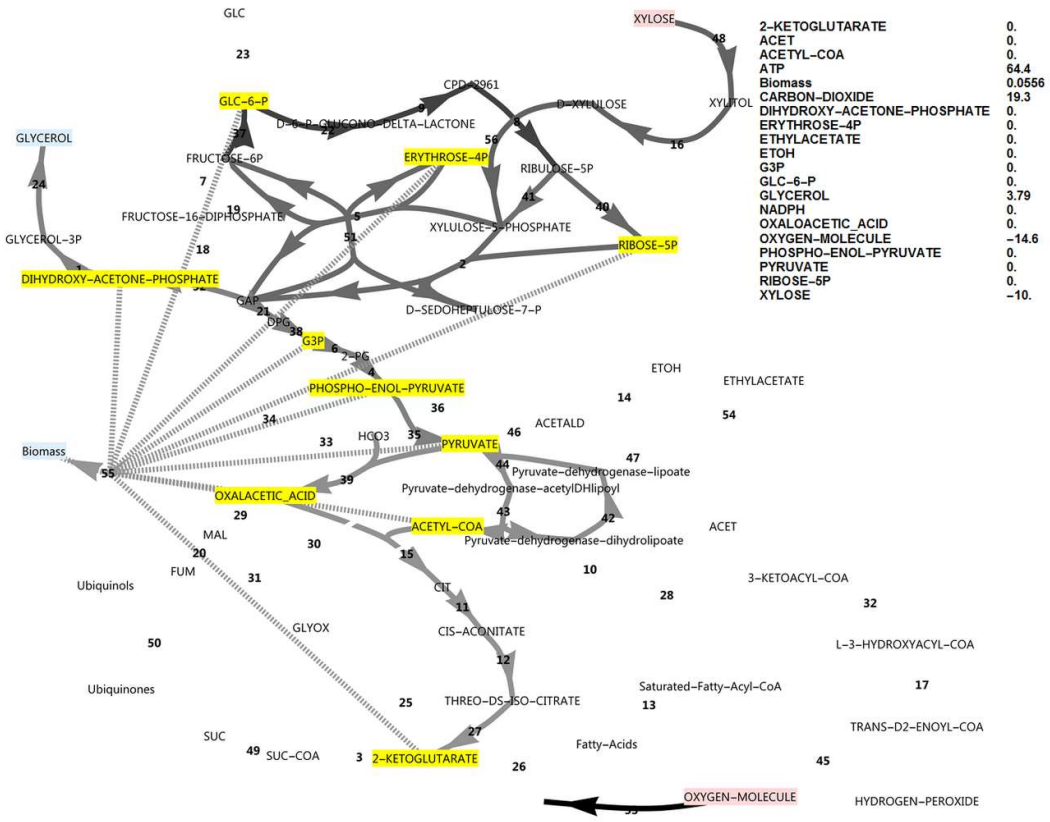


Fig 10. FBA simulation with xylose as *in silico* carbon source, assuming that the oxidative PPP produces one NADPH and one NADH. All biomass precursors as well as NADPH and ATP were allowed to accumulate and with PFK active and FBP inactive. Note the presence of glycerol production, the absence of PFK flux and a large ATP exchange flux. The growth rate was lower compared to the model in Fig 9 where FBP was active. The reaction names are as in Fig 1.

Simulated anaerobic conditions

The FBA model, which is independent from the RNA-seq data, also provides a means to explore the theoretical potential of future engineered yeast strains. From this perspective, metabolism is only limited by the flux constraints, the activity bounds assumed for the uptake rate of the carbon source, and thermodynamic feasibility of the individual reactions, and assumes that gene expression levels can be suitably set by the metabolic engineer. In contrast to the aerobic scenario, under simulated anaerobic conditions (Fig 11) ATP production was balanced with NADPH approximately balanced. As a small overproduction of NADPH occurred, additional flux through the cyclic oxidative PPP was not required and thus neither was FBP required for additional NADPH production nor for ADP regeneration. Flux through glucose-6-phosphate isomerase, functioning in the gluconeogenic direction, was in the model only for the formation of cell wall glucans. Most important, however, was that oxaloacetate-decarboxylating ($NADP^+$-requiring) malic enzyme activity was required to be active

in order to obtain a steady state in these simulations. Together with pyruvate carboxylase and malate dehydrogenase of the TCA cycle, these enzymes formed a substrate cycle in the simulation that effectively oxidised the excess NADH, originating from xylulose reductase activity, to NAD^+ and supplied additional NADPH for xylose reductase. At the same time, excess ATP was consumed by the cycle, which resulted in the balancing of ATP production and utilisation. These reactions are also shown in Fig 12. Limiting the capacity of malic enzyme *in silico* also manifested as xylitol accumulation, a phenomenon often observed in yeasts growing on xylose.

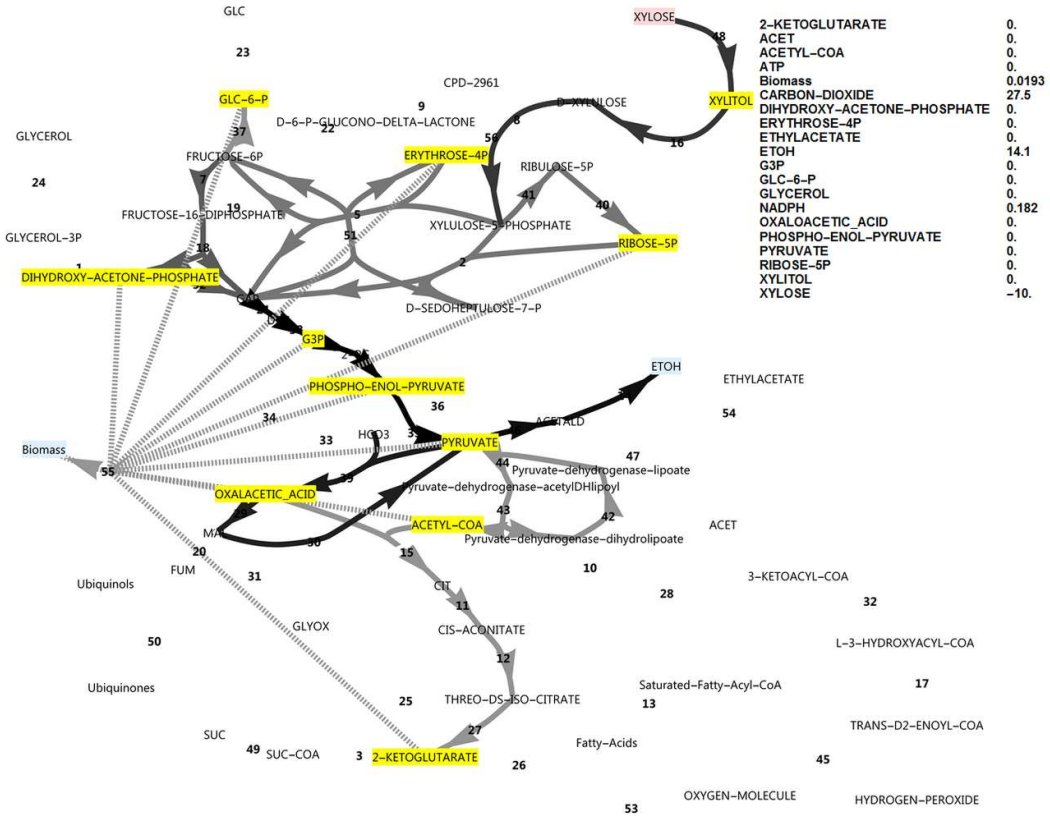


Fig 11. FBA simulation with xylose as *in silico* carbon source under anaerobic conditions. All biomass precursors as well as NADPH and ATP were allowed to accumulate and with both FBP and PFK active. Note the presence of ethanol production, absence of FBP flux or ATP accumulation, and the presence of a cyclic flux involving malic enzyme (reaction 30) which was required for *in silico* growth. Limiting malic enzyme activity caused the accumulation of xylitol. The reaction names are as in Fig 1.

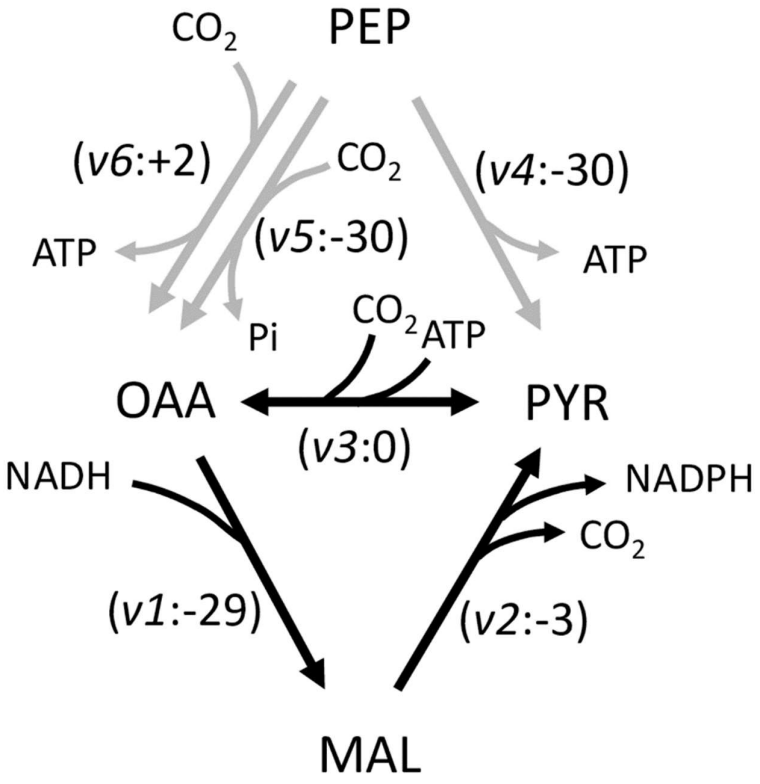


Fig 12. Reactions in anaplerosis that may enable a malic enzyme cycle. Reactions in black indicate the cycle. Reaction names are as follows: *v1*, malate dehydrogenase; *v2*, malic enzyme ($NADP^+$-dependent/decarboxylating malate dehydrogenase); *v3*, pyruvate carboxylase; *v4*, pyruvate kinase; *v5*, phosphoenolpyruvate carboxylase; *v6*, phosphoenolpyruvate carboxykinase. OAA: oxaloacetate. PYR: pyruvate. MAL: malate. PEP: phosphoenolpyruvate. Pi: orthophosphate.

The gene for malic enzyme, MAE1, was found in our annotation against UniProt, and was constitutively expressed under the aerobic conditions tested in glucose and xylose media using RNA-seq [2]. To our knowledge, xylose fermentation by *K. marxianus* under anaerobic or oxygen-limited conditions has not been described thus far, which might be the relevant condition for the proposed malic enzyme cycle. It is also important to note that no flux was predicted in the oxidative PPP flux under the anaerobic condition utilising xylose. Sufficient NADPH for growth was produced by the oxaloacetate-decarboxylating malic enzyme. Ribose-5-phosphate was derived from xylulose-5-phosphate via ribulose-5-phosphate. PDB, which reportedly is utilised under anaerobic conditions [21], only contributed a minor effect to the growth rate, as an *in silico* knock-out of the PDB enzyme only resulted in a minor decrease in growth rate. From a flux perspective, its role is thus not clear for anaerobic conditions.

Responses of an engineered *S. cerevisiae* strain to xylose under anaerobic conditions was recently determined [18], showing a 6.5-fold down-regulation of the MAE1 gene in a xylose medium. However, since *S. cerevisiae* did not evolve for fermenting xylose, the latter observation cannot be extrapolated to *K. marxianus*. There was, however, a four-fold up-regulation of FBP1 in the recombinant *S. cerevisiae* strain when fermenting xylose.

Discussion

Cofactor balance is an important consideration in developing a metabolic engineering strategy. However, even without taking enzyme kinetics into account, the multiple reactions that involve NAD^+, $NADP^+$ and ATP, combined with a variable and uncertain cofactor specificity for redox cofactors, render central metabolism complex to understand. FBA as a framework was demonstrated here to provide insight into the implications of alternative cofactor specificity of oxidative PPP enzymes. The potential role of the FBP reaction in cofactor balance was explored and, at the same time, its possible involvement in the ATP balance by forming a substrate cycle with PFK. It was shown that FBP could indeed contribute to additional NADPH production during xylose utilisation by causing a cyclic PPP flux, but only in a scenario where the oxidative PPP enzymes were not exclusively specific for $NADP^+$ and only under aerobic conditions. Additionally, it was shown that in an *in silico* xylose medium, the excess ATP production was substantially greater than on glucose. This effect was increased further when the cofactor specificity of oxidative PPP enzymes for NAD^+ increased. Thus, FBP may have a dual function under aerobic conditions – both to increase NADPH production in yeasts with one NAD^+ -specific oxidative PPP enzyme and to hydrolyse excessive ATP.

Considering metabolic engineering of a future xylose-fermenting, bioethanol-producing yeast, at least three main aspects would be critical for success. Firstly, anaerobic growth requires an additional fermentative route for oxidation of the additional NADH that is produced by xylitol reductase and which cannot stoichiometrically be oxidised by alcohol dehydrogenases, as there is a shortage of electron acceptors. Secondly, it has been shown in *S. cerevisiae* that the glycolytic flux is strongly coupled to the yield of ATP during catabolism [22]. For instance, by using recombinant active transporters for sugars instead of facilitated diffusion in *S. cerevisiae*, the ethanol yield on sugar was increased while the biomass yield was decreased [23]. Since under xylose utilisation there seems to be a further increase in ATP overproduction, negative feedback on glycolysis by a high ATP concentration might amplify this problem. Thirdly, for every xylose molecule utilised, one NADPH would need to be oxidised which would have to be regenerated from $NADP^+$ in another pathway,

which might be the oxidative PPP, putting further flux constraints on the system. Notably, it was shown in this report that a potential malic enzyme cycle involving oxaloacetate-decarboxylating, $NADP^+$-dependent malic enzyme could theoretically fulfil all three roles, enabling xylose fermentation, which currently is not theoretically possible in *K. marxianus* and other natural yeasts that do not possess a xylose isomerase gene. The cycle would oxidise the excessive NADH that cannot be oxidised by alcohol dehydrogenase due to flux constraints. In doing so, it also provides additional NADPH for the xylose reductase step for xylose utilisation. Finally, it forms an additional ATP sink, which may further pull glycolytic flux forward to ethanol production. Thermodynamic feasibility of this cycle involving malic enzyme should to be calculated based on the concentrations of ATP, ADP, Pi, NADH, NAD^+, NADPH and $NADP^+$ for this species, which is currently unknown. In addition, the PFK/FBP cycle may constitute an additional ATP sink, if it was found that ATP was inhibiting glycolytic flux. This cycle might, in fact, already be active in xylose-utilising *K. marxianus*, since our RNA-seq data showed that the FBP1 gene was derepressed in the xylose medium [2].

Other unanticipated findings were brought to light. Aerobic glycerol production as a physiological strategy to avoid excessive ATP production was also proposed here. Glycerol production is often observed in yeast fermentations and has primarily been associated with oxygen-limited growth conditions, but here the simulation demonstrated a different mechanism. The role of the PDB was also investigated. It was shown that allowance for a PDB flux increased the simulated growth rate under aerobic conditions when ATP was not allowed to be over-produced, as it hydrolysed ATP in the ATP over-producing scenario of aerobic xylose utilisation. In the case when oxidative PPP enzymes were more specific for NAD^+, PDB may also improve growth since it produces additional NADPH. However, the potential contribution of PDB to these effects was low, as this flux was constrained by the flux in lower glycolysis. In a scenario of a defined mineral medium where the organism has to synthesise all biomass components from a sugar as carbon and energy source, as was simulated here, the potential flux through the PDB is further decreased as anaplerotic reactions from lower glycolytic intermediates subtract from the potential flux to PDB to replenish oxaloacetate for amino acid synthesis.

Two additional potential ATP sinks exist in yeasts. The first is another cycle in anaplerosis in which pyruvate and phosphoenolpyruvate are interconverted and dissipate free energy by effectively hydrolysing ATP [24, 25]. PEP carboxylase was not found in the *K. marxianus* UFS-Y2791 annotation, whereas PEP carboxykinase was constitutively expressed at low levels, as was pyruvate carboxylase [2]. Therefore, there was thus no indication of such a cyclic pathway being active. Another is the H^+-

ATPase of the cell membrane [17, 26, 27] as well as the ABC drug efflux pumps of *S. cerevisiae* [28]. H^+-ATPase exports protons that originated from the proton symport of nutrients, including NH_4^+**.** The ATP-hydrolysing capacity of the efflux pumps, however, results in the pumping of protons and NH_4^+, and hence cannot act as condition-independent replacements for uncoupling proteins.

Conclusions

This work shows that cofactor balances should not be interpreted separately. Furthermore, it highlighted the importance of experimentally determining the cofactor specificity of oxidative PPP enzymes, as the prediction or calculation of fluxes will change, depending on these parameters. The ATP balance is highly relevant to metabolic engineering, since ATP not only has a negative feedback on glycolysis, but also a high ATP concentration might lead to excessive biomass formation, reducing the yield of a primary fermentation product such as ethanol. Thus, from a practical perspective, the proposed substrate cycle induced by FBP may be useful in decreasing the effective ATP yield on glucose, which may lead to an increase in ethanol production. This cycle might already be employed by natural *K. marxianus* strains, since the FBP1 gene was up-regulated in the glucose-free xylose medium, as observed in RNA-seq data. In addition, it was proposed here that a highly active malic enzyme cycle, which effectively exchanges redox equivalents from NADH to $NADP^+$, would not only solve a redox imbalance under anaerobic, xylose fermenting conditions, but would draw the glycolytic flux forward due to forming an ATP sink. Finally, it is evident that ATP should not be thought of as always in demand. Depending on the condition, it might become excessive, especially in the case of pentose utilisation. The FBP/PFK substrate cycle and its regulation probably plays a key role in energy homeostasis, even in yeasts under selected conditions, which could determine phenotype switching. Alteration of the activity of the FBP/PFK substrate cycle and its effects on cofactor balances of ATP, NADH and NADPH may also be a mechanistic explanation for the recent discovery of the strong causative link between mutations affecting the FBP1 gene and renal cell carcinoma in humans.

References

1 Lertwattanasakul N, Kosaka T, Hosoyama A, Suzuki Y, Rodrussamee N, Matsutani M, et al. Genetic basis of the highly efficient yeast *Kluyveromyces marxianus*: complete genome sequence and transcriptome analyses. Biotechnol Biofuels. 2015;8(47). doi: 10.1186/s13068-015-0227-x.

2 Schabort DTWP, Letebele PK, Steyn L, Kilian SG, du Preez JC. Differential RNA-seq, multi-network analysis and metabolic regulation analysis of *Kluyveromyces marxianus* reveals a compartmentalised response to xylose. PLoS ONE 11(6): e0156242. doi:10.1371/journal.pone.0156242.

3 Klein CJ, Olsson L, Nielsen J. Glucose control in *Saccharomyces cerevisiae*: the role of Mig1 in metabolic functions. Microbiology. 1998;144 (1): 13-24.

4 Li B, Qiu B, Lee DSM, Walton ZE, Ochoki JE, Mathew LK, et al. Fructose-1,6-bisphosphatase opposes renal carcinoma progression. Nature. 2014;514: 251-255.

5 Runquist D, Hahn-Hägerdal B, Bettiga M. Increased expression of the oxidative pentose phosphate pathway and gluconeogenesis in anaerobically growing xylose utilising *Saccharomyces cerevisiae*. Microb Cell Fact. 2009;8(49). doi: 10.1186/1475-2859-8-49.

6 Hahn-Hägerdal B, Karhumaa K, Fonseca C, Spencer-Martins I, Gorwa-Grausland MF. Towards industrial pentose-fermenting yeast strains. Appl Microbiol Biotechnol. 2007;74: 937–953.

7 Voet D, Voet G. Biochemistry. 4th ed. New Jersey: Wiley; 2011.

8 Newsholme EA, Challiss RAJ, Crabtree B. Substrate cycles: their role in improving sensitivity in metabolic control. Trends Biochem Sci. 1984;9: 277-280.

9 Newsholme EA, Crabtree B, Higgins SJ, Thornton SD, Start C. The activities of fructose diphosphatase in flight muscles from the bumble-bee and the role of this enzyme in heat generation. Biochem J. 1972;128: 89-97.

10 Jarmuszkiewicz W. Uncoupling proteins in mitochondria of plants and some microorganisms. Acta Biochim Pol (Engl Transl). 2001;48: 145-155.

11 Luévano-Martínez LA, Moyano E, de Lacoba MG, Rial E, Uribe-Carvajal S. Identification of the mitochondrial carrier that provides *Yarrowia lipolytica* with a fatty acid-induced and nucleotide-sensitive uncoupling protein-like activity. Biochim Biophys Acta. 2010;1797: 81–88.

12 Luévano-Martínez LA. Uncoupling proteins (UCP) in unicellular eukaryotes: true UCPs or UCP1-like acting proteins? FEBS Lett. 2012;586: 1073-1078.

13 Fischer E, Zamboni N, Sauer U. High-throughput metabolic flux analysis based on gas chromatography–mass spectrometry derived ^{13}C constraints. Anal Biochem. 2004;325: 308–316.

14 Schilling CH, Edwards JS, Palsson BO. Toward metabolic phenomics: analysis of genomic data using flux balances. Biotechnol Prog. 1999; 15: 288–295. PMID: 10356245.

15 Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng. 2003;5(4): 264–276.

16 Blank LM, Lehmbeck F, Sauer U. Metabolic-flux and network analysis in fourteen hemiascomycetous yeasts. FEMS Yeast Res. 2005;5: 545–558.

17 Verduyn C, Postma E, Scheffers WA, van Dijken JP. Effect of benzoic acid on metabolic fluxes in yeasts: a continuous culture study on the regulation of respiration and alcoholic fermentation. Yeast. 1992;8: 501–517.

18 Alff-Tuomala S, Salusjärvi L, Barth D, Oja M, Penttilä M, Pitkänen J, et al. Xylose-induced dynamic effects on metabolism and gene expression in engineered *Saccharomyces cerevisiae* in anaerobic glucose-xylose cultures. Appl Microbiol Biotechnol. 2016;100(2): 969–985.

19 Fuhrer T, Sauer U. Different biochemical mechanisms ensure network-wide balancing of reducing equivalents in microbial metabolism. J Bacteriol. 2009;191(7): 2112–2121.

20 Naylor CE, Gover S, Basak AK, Cosgrove MS, Levy HR, Adams MJ. $NADP^+$ and NAD^+ binding to the dual coenzyme specific enzyme *Leuconostoc mesenteroides* glucose 6-phosphate dehydrogenase: different interdomain hinge angles are seen in different binary and ternary complexes. Acta Cryst. 2001;D57: 635-648.

21 Saint-Prix F, Bönquist L, Dequin S. Functional analysis of the ALD gene family of *Saccharomyces cerevisiae* during anaerobic growth on glucose: the $NADP^+$-dependent Ald6p and Ald5p isoforms play a major role in acetate formation. Microbiology. 2004;150: 2209-2220.

22 de Kok S, Kozak BU, Pronk JT, van Maris JA. 2012. Energy coupling in *Saccharomyces cerevisiae*: selected opportunities for metabolic engineering. FEMS Yeast Res. 2012;12: 387–397.

23 Basso TO, de Kok S, Dario M, do Espirito-Santo JCA, Müller G, Schlölg PS, et al. Engineering topology and kinetics of sucrose metabolism in *Saccharomyces cerevisiae* for improved ethanol yield. Metab Eng. 2011;13: 694–703.

24 Peksel A, Torres NV, Liu J, Juneau G, Kubicek CP. ^{13}C-NMR analysis of glucose metabolism during citric acid production by *Aspergillus niger*. Appl Microbiol Biotechnol. 2002;58: 157-163.

25 Papagianni M. Advances in citric acid fermentation by *Aspergillus niger*: biochemical aspects, membrane transport and modelling. Biotechnol Adv. 2007;25(3): 244-263.

26 Piper P, Mahe´ Y, Thompson S, Pandjaitan R, Holyoak C, Egner R, et al. The Pdr12 ABC transporter is required for the development of weak organic acid resistance in yeast. EMBO J. 1998;17: 4257–4265.

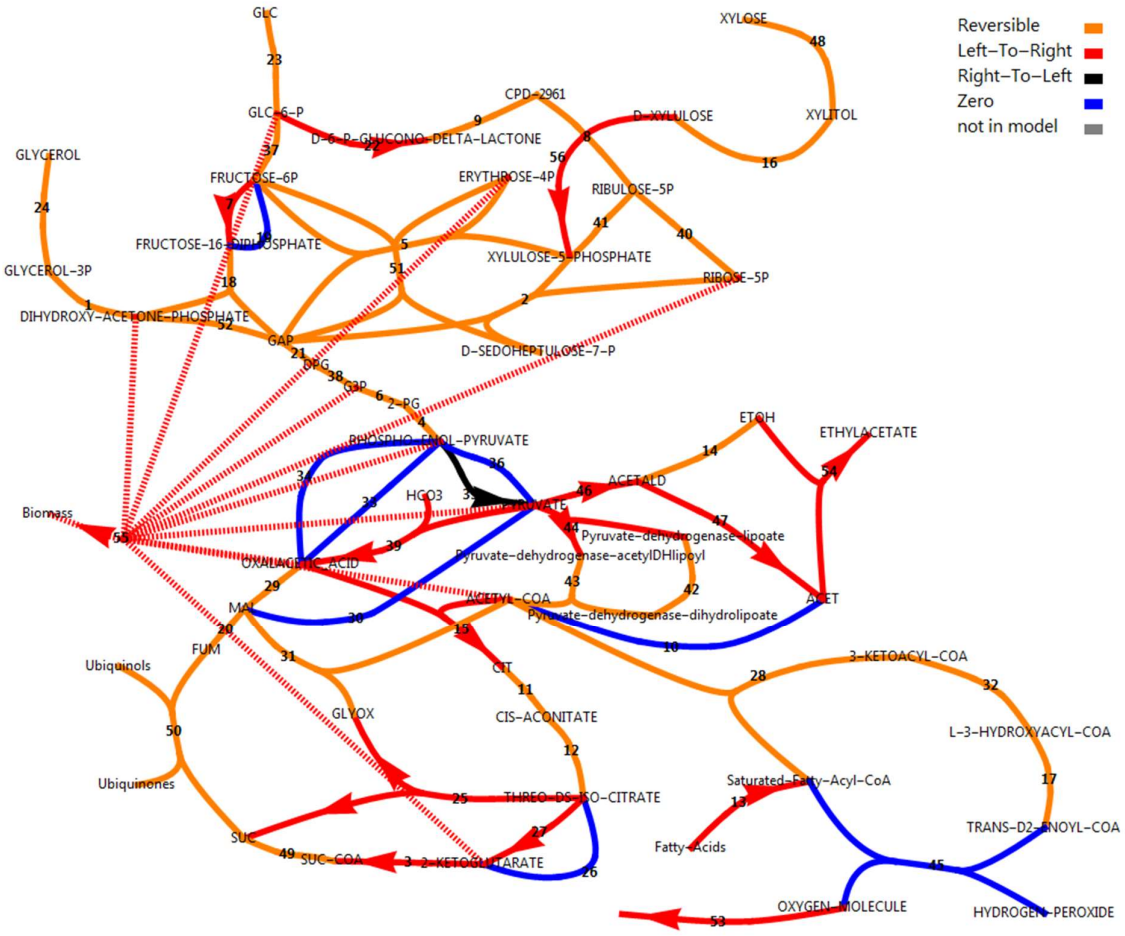
27 Abbott DA, Knijnenburg TA, de Poorter LMI, Reinders MJT, Pronk JT, van Maris AJA. Generic and specific transcriptional responses to different weak organic acids in anaerobic chemostat cultures of *Saccharomyces cerevisiae*. FEMS Yeast Res. 2007;7: 819–833.

28 Krasowska A, Łukaszewicz M, Bartosiewicz D, Sigler K. Cell ATP level of *Saccharomyces cerevisiae* sensitively responds to culture growth and drug-inflicted variations in membrane integrity and PDR pump activity. Biochem Biophys Res Commun. 2010;395: 51-55.

Supporting Information

This paper was published online in PLoS ONE on 27 May 2017. See supporting information S1 Figures online at https://doi.org/10.1371/journal.pone.0177319**.**

Schabort DWP, Kilian SG, du Preez JC. Elucidation of new condition-dependent roles for fructose-1,6-bisphosphatase linked to cofactor balances.



Supplementary Fig 1. Reaction directionality constraints applied in FBA simulations. No arrow indicates a reversible reaction. Arrows indicate the direction of unidirectional reactions. Black indicates reactions specified by the parameter set to operate from right to left while reactions in red operate from left to right. Reactions in blue are specified by the parameter set to be zero in both conditions. Note that often the default reaction

direction specified in databases is not the direction in which reactions are catalyzed even under the most common conditions. A visual representation of reaction direction greatly facilitates model construction and simulation. It is also helpful in situations when no steady state is reached, such that correct bounds could be set, and is complementary to reachability analysis and flux variability analysis.

Conclusions

In this work, a draft genome of a fast-growing, thermotolerant strain of *Kluyveromyces marxianus*, UFS-Y2791 was sequenced and assembled (Chapter 2 and Addendum 1). It was found that by using the cost-effective method of NGS and *de novo* assembly alone, the resultant draft genome already contained sufficient information to capture a detailed view of metabolic pathways in this non-model organism. After this reconstruction of a blueprint of cellular machinery and regulation, a large data set of differential gene expression was generated using RNA-seq, exploring the response of this yeast to glucose or xylose as the carbon source. In addition, the first genome-wide gene regulatory network for the yeast *K. marxianus* was constructed (Chapters 5 and 6). By analysing the network in the context of rich RNA-seq data, it was possible to reveal differentially active transcriptional regulators in a novel manner. This protocol of *in silico* network reconstruction was required to progress from the exploration of RNA-seq data (Chapter 3) to elucidating cause and effect in this non-model organism (Chapters 4-7) in an integrated systems biology manner.

Strikingly, all glycolytic genes were down-regulated in unison in the xylose medium compared to the glucose medium, as if by a single transcription factor (TF). The combined action of Gcr1 and Gcr2 may well be the mechanistic basis for this observation, of which the genes were both down-regulated (Chapters 5 and 6). Similarly, many of the peroxisomal genes were up-regulated, reminiscent of glucose derepression as was found in *Saccharomyces cerevisiae* [Young et al. 2003]. The heptamer frequency analysis in Chapter 4 suggested that Adr1 and possibly Mig1 were responsible for their up-regulation, as well as up-regulation of other genes for the utilisation of alternative carbon sources such as the inulinase gene INU1, alternative sugar transporters and those encoding the 2-methylcitrate cycle [Young et al. 2003]. This was supported both by the strong up-regulation of the Adr1 gene and the down-regulation of the Mig1 gene. Other TFs that control peroxisomal genes in *S. cerevisiae*, such as the oleate-responsive Oaf1 and Pip2 [Ratnakumar et al. 2010] were absent from the list of enriched TFs, hence Adr1 is the most likely candidate responsible for the strong peroxisomal response to the xylose medium containing no glucose.

Although the transcriptomic response in the xylose medium resembled derepression from glucose, it did not necessarily resemble gluconeogenesis. For instance, the genes SFC1 (succinate/fumarate mitochondrial transporter), PCK1 (phosphoenolpyruvate carboxykinase) and ICL1 (isocitrate lyase) are typical gluconeogenic enzymes, but were constitutively expressed with glucose or xylose as carbon

substrates. This suggested that at least one additional activator is required for up-regulation of gluconeogenic genes, which might be Rds2 and shown previously to activate genes of gluconeogenesis in *S. cerevisiae* [Soontorngun et al. 2007]. By including more such RNA-seq datasets, originating from cultivations in multiple nutrient combinations, these additional transcription factors might be identified. This approach is currently under investigation.

While various enzymes involved with alternative carbon source utilisation were up-regulated in the xylose medium, a number of genes involved in biosynthesis such as nucleotide and amino acid biosynthesis were moderately down-regulated, corresponding to the moderately down-regulated growth rate. The TF enrichment analysis based on the complete gene regulatory networks from likelihoods revealed that the TFs Gcn4, along with Arg81, Bas1 and Rtg3, controlled the down-regulation of these TFs.

It is notable that none of the major regulators associated with the down-regulated target gene sets were themselves down-regulated at the level of gene expression (Chapters 5 and 6). Clearly, post-translational modifications played a major role in regulating the activity of these TFs, of which phosphorylation is probably the most important. The apparent lower activity of Gcn4, combined with the many up-regulated genes of carbon source utilisation, suggested that Snf1 was the master kinase regulating the activity of Gcn4 via Gcn2 [Shirra et al. 2008] and other TFs, including Adr1 [Young et al. 2003] and Mig1 [Schuller 2003]. Snf1 is a key regulatory point in combining the signals resulting from the presence of fermentable carbon sources, as well as due to nitrogen limitation [Orlova et al. 2006], oxidative stress, a high salt concentrations and a high pH value [Hong and Carlson 2007]. Conversely, the strong up-regulation of the Adr1 gene rather suggested that the cAMP dependent RAS/PKA pathway was differentially regulated [Dombek and Young 1997]. Unfortunately, phospho-proteomics is technically very challenging and was, therefore, outside the scope of this investigation. RNA-seq is less technically challenging and can detect the differential expression of the genes encoding kinases. However, the vast majority of kinases were constitutively expressed. Their activities depend on their own phosphorylation states, which is generally the mode of their regulation. Thus, a method was developed to create an enriched subnetwork of the kinase signalling networks to elucidate the most differentially active kinases by their associations with enriched TFs, which were derived from enrichment statistics of the target sets of TFs using RNA-seq data. This analysis suggested the Pho85-Pcl5 cyclin dependent kinase, as well as the kinase Ssn3 (Srb10, Cdc8), to be important for the regulation of Gcn4. The idea was also extended to enrichment at any network depth, which could be termed *long-range enrichment*. Unfortunately, the longer network distances involve increasing

numbers of assumptions, making it unlikely to reveal differentially active upstream master kinases such as Snf1 and Ras/PKA that do not act directly on TFs. The possibility of improving this method is currently under investigation. Likely, phospho-proteomics is the single most suitable complement to RNA-seq for investigations of this type.

There was also substantial evidence in support of the yeast pheromone signalling system being differentially active (Chapter 6). The pheromone detector component genes of Ste2 and Ste3 were up-regulated in the xylose medium, as well as Gpa1 which receives signals from them. A few phosphorylation steps downstream in this pathway is Fus3, of which the gene was also up-regulated. Fus3 signals to Ste12, which binds to Tec1 (also up-regulated) and activates the gene expression of Phd1 (at least in *S. cerevisiae*), a major transcriptional activator of pseudohyphal growth [Broach 2012]. Both Phd1 and Ste12 were also shown to be significantly enriched and associated with mostly up-regulated genes (Chapter 6). Interestingly, it was found that Ste12 was under the gene regulation of Phd1 in the gene regulatory network, suggesting transcriptional rewiring between *S. cerevisiae* and *K. marxianus*. These observations could be investigated further, which would likely require the suitable phospho-proteomics.

The analysis of metabolism and its regulation at the genome scale is hampered by the fact that not all fluxes can be measured or calculated at the genome scale, since they may span subcellular compartments (Chapter 3). Thus, the network structure may not allow complete identifiability, despite the recent advances in ^{13}C-Metabolic Flux Analysis (^{13}C-MFA), practical experimental protocols [Dauner and Sauer 2000] and convenient software programs [Zamboni et al. 2005]. While improvements in flux analysis is an ongoing field of research, this will inevitably entail a stronger integration with the other omics. There also seems to be a strong case for constructing detailed enzyme kinetic models of selected pathways in which the fluxes cannot be resolved with ^{13}C-MFA. In particular, the pyruvate dehydrogenase bypass and alcohol dehydrogenase displayed isozyme switching of enzymes between the conditions, and the metabolic intermediates can cross the mitochondrial membrane (Chapter 3). A kinetic model might both resolve the fluxes among the compartments and reveal whether the reason behind isozyme switching might reside in alternative enzyme kinetic parameters.

In this work, metabolism was studied in several ways and each method discovered different aspects regarding metabolism. The use of RNA-seq was demonstrated as a particularly powerful starting point for the study of metabolism. Mapping differential RNA-seq data to central metabolic pathways

revealed a pattern of all genes in glycolysis being down-regulated, with the up-regulation of genes of the non-oxidative pentose pathway (PPP) and constitutive expression of genes of the oxidative PPP, as well as constitutive expression of the TCA cycle genes. Using flux balance analysis (FBA) to simulate the fluxes that were expected in the glucose and xylose media (differential FBA, Chapter 3), the differential flux pattern was found to be strikingly similar to the differential gene expression pattern. Hence, even reactions that are usually regarded as rapid equilibrium enzymes [Canelas et al. 2011] were differentially regulated. This observation suggested that the genetic level of regulation might be equally important to the metabolic level of regulation. The latter form involves the changes in flux brought about by changes in metabolite levels and their interaction with the enzymes. The theoretical framework of metabolic regulation analysis (MRA) was used as a practical approach to studying the contribution of the genetic (hierarchical) and metabolic levels of regulation. The genetic level of regulation might play a pertinent role in the regulation of fluxes, depending on the reaction and the conditions. MRA could be performed at the genome scale, if flux analysis could be expanded to a scale approaching that of functional genomics. It was also notable how well the transcript levels from RNA-seq corresponded to the simulated flux patterns.

The simulation framework of FBA was used in two instances to rationalise the transcriptomics response. By comparing the differential gene expression levels with the fluxes expected in the glucose and xylose media, incorrect expectations were revealed. Differential RNA-seq initially revealed that the genes of the oxidative PPP, which were expected to be up-regulated since the PPP produces the NADPH required by xylose dehydrogenase, were constitutively expressed in both these two culture media (Chapter 3). Comparative differential FBA revealed that, due to the lower growth rate on xylose, the cell did not require additional NADPH. FBA was also used to explore the potential role of the fructose-1,6-bisphosphatase (FBP) reaction, which is regarded as a typical gluconeogenic enzyme. The FPB1 gene was up-regulated in the xylose medium (Chapter 9). It was, however, demonstrated that especially in the xylose medium, up-regulation of the FBP1 gene might play an active role, serving in a futile substrate cycle with the phosphofructokinase (PFK) reaction, and thus likely plays no role in additional NADPH production under aerobic conditions. This is applicable to yeasts in which both of the oxidative reactions in the oxidative PPP produce a molecule of NADPH, whereas, if either of these produce NADH instead, the role for up-regulation of FBP1 may be for providing additional NADPH. In the case of cells grown on xylose, FBA showed that the ATP overproduction was substantially higher than when grown on glucose, supporting the potential role of the FBP/PFK cycle as a candidate replacement mechanism for a dedicated ATP uncoupling mechanism, which seems to be absent in yeasts. ATP should thus not be regarded as being always in demand. To the contrary, for the purpose

of engineering biofuel producing yeast strains, the utilisation of ATP dissipating strategies might prove advantageous, since ATP has a negative feedback on glycolytic flux [de Kock et al. 2012], which prevents an overflow metabolism to ethanol. The lack of an ATP hydrolysing mechanism was also linked to wasteful glycerol and xylulose production, which are often observed in engineered yeast strains and interpreted as representing constraints in oxidative metabolic capacity or oxygen limitation, but independent of possible ATP overproduction.

Considering this complex interplay of cofactors, a strategy was proposed for engineering a future Crabtree positive, xylose fermenting strain of *K. marxianus*, making use of a cytoplasmic malic enzyme cycle (Chapter 9). In addition, engineering strains for a constitutive high expression of Gcr1 or Gcr2 might result in an overflow metabolism towards ethanol. Conveniently, the pyruvate decarboxylase gene PDC1 also is under the regulation of Gcr2 (Chapter 6). The alcohol dehydrogenase isozyme responsible for ethanol production would need to be identified, among several ADH genes, and constitutively over-expressed. Alternatively, ADH1 from *S. cerevisiae* should be a suitable replacement for constitutive expression in *K. marxianus*.

Limitations of this study relate to the fact that only two conditions could be compared and that only RNA-seq data could be generated. Proteomics and specifically phosphor-proteomics could be a significant benefit, since it was revealed in Chapter 7 that many of the master regulators cannot be revealed by long-range enrichment analyses based on RNA-seq data alone. This prediction method, which uses the enrichment values of transcription factor based on their targets as a replacement for phospho-proteomics, is likely more relevant to close-range enrichment only, and phosporylations should likely be directly observed on the organism under study. However, the additional experimental complexity and significantly smaller scope of analysis in proteomics rendered these techniques outside the scope of this project. Another finding that might question the validity of the approaches used to reveal differentially active TFs, was that the putative zinc-finger-like TF involved in up-regulation of alternative carbon source utilisation genes was not found initially by the likelihood based network enrichment approach (Chapter 5). Chapter 4 revealed Adr1 or Mig1 as the most likely regulators based on the enumerative k-mer frequency approach, the consistent observations of differential expression of both their transcripts, and previous reports on the model species *S. cerevisiae*. After thorough investigation with multiple network construction functions, it was found that Adr1 could indeed be discovered from networks generated using the relevant function for incorporating the sources of evidence (Chapter 6). This not only implies that different modes of network construction should be used to reveal different TFs in a complementary fashion, but also that

the enumerative k-mer method, and the innovation of Occam's Razor motifs should be used in a complementary fashion to network-based enrichment approaches.

Although the work presented in this thesis could be considered as *integrative*, involving the gene regulatory level (Chapters 3, 5 and 6), signalling through kinases, (Chapter 7), the metabolic flux level (Chapters 3 and 9), as well as the chromosomal context of gene regulation (Chapter 8), our current understanding of eukaryotic gene regulation is still simplistic relative to the true complexity. The combination of high-throughput functional genomics datasets, enrichment statistics and likelihood based networks is set to become increasingly important in the information-rich future of the biosciences. We will likely see increased use of high-throughput datasets, while likelihood-based methods, such as was developed in Chapters 5 and 6, can combine multiple sources of evidence. Enrichment statistics is a useful method of making sensible deductions involving regulators that affect multiple targets. Importantly, the method is robust to potentially false positive assignments of some of the interactions. Ultimately, these methods could be combined into a machine learning algorithm that could optimise simultaneously the network structure as well as the motifs that define the binding sites, and also incorporate signal transduction pathways. This would not only give insight into the gene regulatory networks and signalling, but would also pinpoint our lack of knowledge regarding the DNA binding sites of certain TFs, revealing contradictions between the experimental data analysed and the information on the model species (Chapter 6). It is truly an exciting time for computational biology and bioinformatics.

References

Broach JR. Nutritional control of growth and development in yeast. Genetics. 2012;192: 73-105.

Dauner M, Sauer U. GC-MS analysis of amino acids rapidly provides rich Information for isotopomer balancing. Biotechnol Prog. 2000;16: 642-649.

de Kok S, Kozak BU, Pronk JT, van Maris JA. 2012. Energy coupling in *Saccharomyces cerevisiae*: selected opportunities for metabolic engineering. FEMS Yeast Res. 2012;12: 387–397.

Dombek KM, Young ET. Cyclic AMP-dependent protein kinase inhibits ADH2 expression in part by decreasing expression of the transcription factor gene ADR1. Mol Cell Biol. 1997;17(3): 1450-8.

Canelas AB, Ras C, Pierick A, van Gulik WM, Heijnen JJ. An *in vivo* data-driven framework for classification and quantification of enzyme kinetics and determination of apparent thermodynamic data. Metab Eng. 2011;13: 294-306.

Hong SP, Carlson M. Regulation of snf1 protein kinase in response to environmental stress. J Biol Chem. 2007;282: 16838–16845.

Orlova M, Kanter E, Krakovich D, Kuchin S. Nitrogen availability and TOR regulate the Snf1 protein kinase in *Saccharomyces cerevisiae*. Eukaryot Cell. 2006;5: 1831–1837.

Ratnakumar S, Young ET. Snf1 dependence of peroxisomal gene expression is mediated by Adr1. J Biol Chem. 2010;285(14): 10703–10714.

Schuller HJ. Transcriptional control of nonfermentative metabolism in the yeast *Saccharomyces cerevisiae*. Curr Genet. 2003;43(3): 139-60.

Shirra MK, McCartney RR, Zhang C, Shokat KM, Schmidt MC. A chemical genomics study identifies Snf1 as a repressor of GCN4 Translation. 2008;283(51): 35889–35898.

Soontorngun N, Larochelle M, Drouin S, Robert F, Turcotte B. Regulation of gluconeogenesis in *Saccharomyces cerevisiae* is mediated by activator and repressor functions of Rds2. Mol Cell Biol. 2007;27(22): 7895–7905.

Young ET, Dombek KM, Tachibana C, Ideker T. Multiple pathways are co-regulated by the protein kinase Snf1 and the transcription factors Adr1 and Cat8. J Biol Chem. 2003;278(28): 26146–26158.

Zamboni N, Fischer E, Sauer U. FiatFlux – a software for metabolic flux analysis from ^{13}C-glucose experiments. BMC Bioinf. 2005;6(209). doi:10.1186/1471-2105-6-209.

Addendum 1

Improved genome assembly and RNA-seq data analysis

Abstract

This addendum explains the improvement of the draft genome for *Kluyveromyces marxianus* UFS-Y2791 and improved RNA-seq data analysis. The use of the Velvet *de novo* genome assembler resulted in an improved draft genome assembly with double the N50 value as compared to the assembly using Abyss. By using a larger amount of RNA-seq data, an improvement was also made in the sensitivity of differential expression analysis. Moreover, it was found that the RNA-seq data from strain UFS-Y2791 could be effectively mapped directly to the complete genome of strain DMKU3-1042. This work was essential for the construction of a genome-scale gene regulatory network and for chromosome-based analyses.

Introduction

The ultimate goal of this study was to reveal the cause-and-effect pathways used by *Kluyveromyces marxianus* to control the differential response to glucose and xylose. Key to this analysis was the elucidation of differentially active transcription factors that controlled gene regulation. Two main approaches exist that could be employed for this purpose. Firstly, the enumerative method of k-mer frequency comparisons could be used, which reveals over-represented words (k-mers) in the regulatory regions of the differentially regulated genes. These k-mers can be matched to transcription factor (TF) binding motifs to reveal the identity of the TFs. The second approach is to construct a complete gene regulatory network and perform enrichment statistics using the differential RNA-seq data. Both these methods require a sufficiently complete draft genome or preferably a complete genome. The initial UFS-Y2791 draft genome is fragmented into 1094 contigs larger than 200 bp. Although the quality of the contigs may be considered to be very high, given the high per base sequencing quality and high coverage, studies in gene regulation could greatly benefit from using a more complete assembly for the strain. It might also be possible to map the RNA-seq data of strain UFS-Y2791 directly to one of the recently published complete genomes for different *K. marxianus* strains, independently from the genomic DNA data for the UFS-Y2719 strain (Chapter 2). This was tested and the results reported in this Addendum. Also, not all the RNA-seq data were used at the time of publication of Chapter 3, which could further be used to improve the analyses. Using a larger amount of RNA-seq data should result in better coverage of genes by RNA-seq reads, thus allowing better statistical significance of differential expression testing. The effect of using three datasets per sample instead of one dataset per sample is reported.

In terms of improving the assembly of the UFS-Y2791 draft genome, two routes were investigated. Firstly, the alternative *de novo* assembly algorithm Velvet [Zerbino 2008], was tested. Secondly, DNA reads were mapped to the complete reference genomes DMKU3-1042 and KCTC-17555, a process known as read-mapping. Using this approach, one option is to take only the stretches of DNA that matched the reference genome, discarding the unaligned reads. This might miss long insertions in the UFS-Y2791 genome as compared to the reference. Alternatively, the insertions may be filled in by *de novo* assemblies of unaligned reads, or by using the insertions hidden in pileup files. The latter process is not trivial, however.

The results described here opened up an avenue of complete genomic analysis. The full genomic context would later improve the quality of gene regulatory networks (Chapter 6), and subsequently allowing RNA-seq analysis in the context of chromosomes (Chapter 8).

Materials and Methods

Processing of NGS data, *de novo* genome assembly and annotation

Genomic DNA and RNA-seq datasets were generated on an Illumina HiScanSQ instrument (described in Schabort et al. 2016). The data were quality assessed using FastQC and trimmed using Trimmomatic [Bolger et al. 2014] in Galaxy [Afgan et al. 2015] using a sliding filter of four bases with minimum average Phred score of 20. *De novo* assemblies were performed using Abyss and optimised for the k-mer length using a program developed in Python (Chapter 2) or by using the Velvet-Optimiser [Zerbino 2008] in Galaxy. N50, N75 and N90 values were calculated in algorithms designed for *Reactomica* and implemented in Mathematica. For reference mapping of genomic NGS data to complete genomes, Bowtie2 [Langmead et al. 2009, Langmead et al. 2012] in Galaxy was used, using default parameters. WebAugustus [Stanke et al. 2008] was used to find open reading frames.

RNA-seq data analysis

Reads were mapped to a variety of genome assemblies using TopHat2 in Galaxy [Trapnell et al 2009, Trapnell et al. 2013, Kim et al. 2013]. The same read-mapping parameters were used throughout. These were as follows.

The constraint of the number of base pairs that may be different between any read and the genome (--read-edit-dist = 2), was set to two, and the final read mismatches was also set at two (--read-mismatches = 2). The maximum insertion and deletion lengths were both set at 3 (--max-insertion-length = 3; --max-deletion-length =3). These parameters are important if the quality of the genome was uncertain, or if the genome was not actually from the same strain and single nucleotide polymorphisms, insertions and deletions were to be expected. CuffDiff [Trapnell et al. 2013] was used to test for differential expression. In all cases, a GFF3 file was used to specify the gene windows in which CuffDiff calculated the differential expression [Trapnell et al. 2013]. This file either originated from annotations of draft assemblies of strain UFS-Y2791, or from the NCBI for strain DMKU3-1042. CuffDiff reports p-values as the statistical significance and well as the q-value, which is the p-value after accounting for multiple comparisons. Genes were only considered to be significantly differentially expressed when q-values were below 0.05.

Results and Discussion

Improved RNA-seq data analysis

Figure 1 shows the excellent per base Phred scores obtained, demonstrating the high quality of Illumina NGS data. A Phred score of 30 indicates that one in a thousand bases in a sequence may be incorrect (error rate = 1/(10^(phred/10))). A phred score of 20 is the default threshold, which could be assumed sufficient for RNA-seq. The decrease in quality with the number of base pairs sequenced is a common feature with most methods of sequencing. Figure 2 shows the quality of the trimmed reads. The majority of untrimmed reads had a quality of 37 (Figure 3), indicating one error expected in 5 012 bases read. Thus, the data were of outstanding quality.

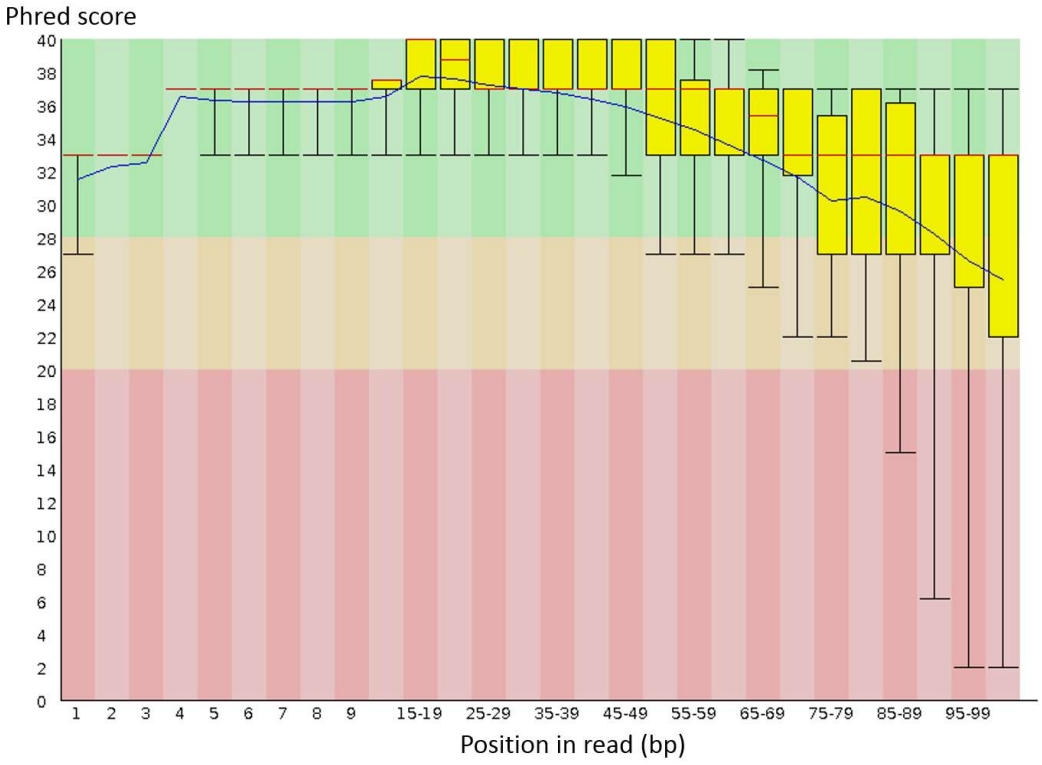


Figure 1. Box-and-whiskers plot of per base quality assessment of RNA-seq data before trimming, visualised using the FastQC tool. Orientation of reads is 5'-3', as sequenced on the Illumina instrument.

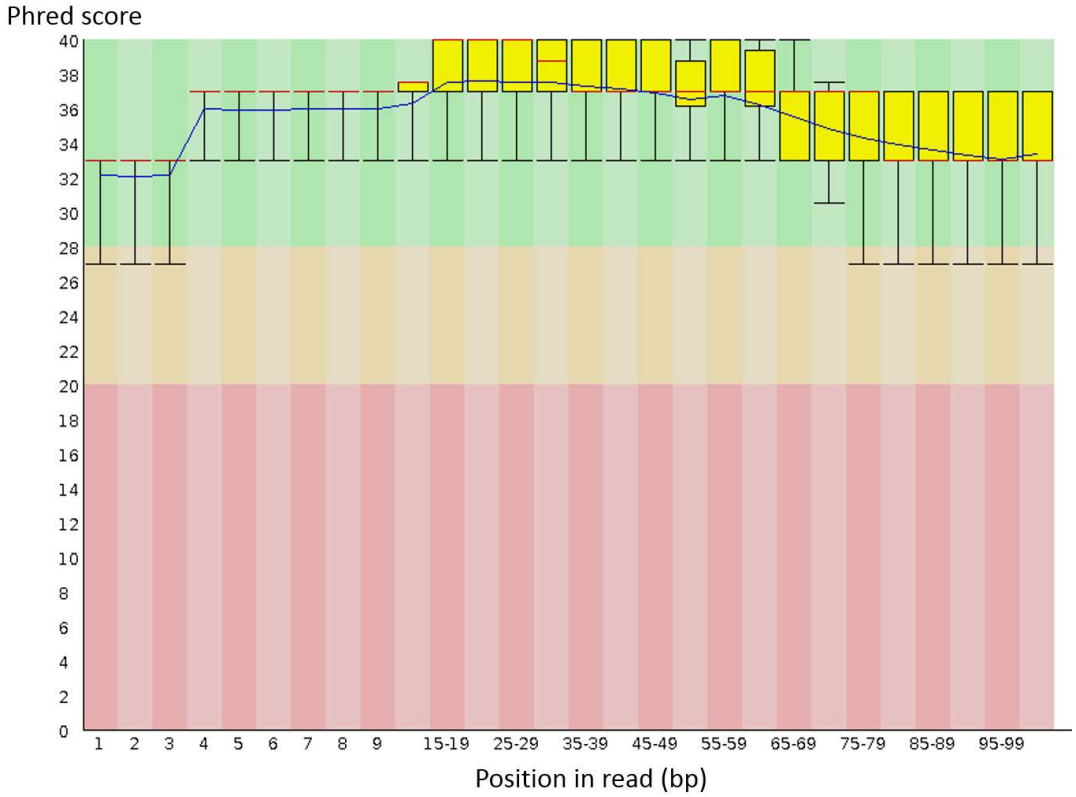


Figure 2. Box-and-whiskers plot of per base quality assessment of RNA-seq data after trimming, using the Trimmomatic tool, and visualised using the FastQC tool, showing higher average base quality. Illumina adapters were removed, followed by a sliding filter of four bases (from the 3' end), testing each read for a local minimum average Phred score of 20. The number of bases removed differs for each read. Orientation of reads is 5'-3', as sequenced on the Illumina instrument.

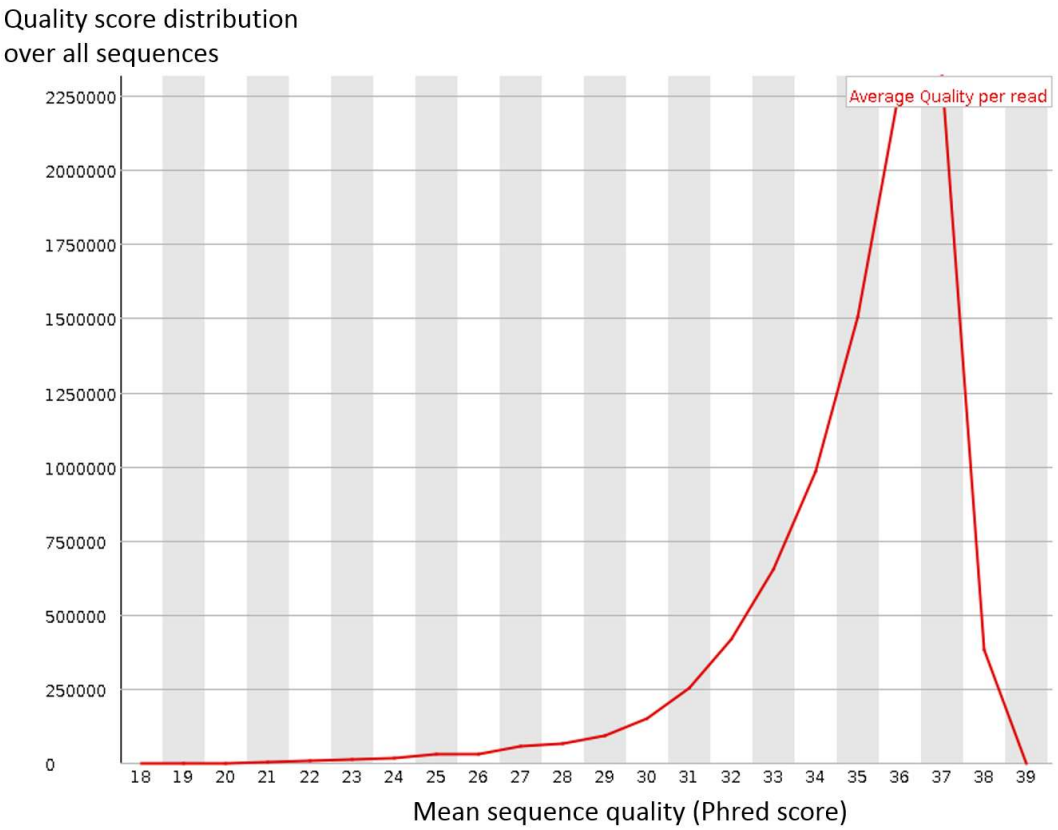


Figure 3. Mean quality score per read before trimming, as obtained from the FastQC tool. Phred scores were calculated over the length of each read. The median Phred score was 38, with a similar mean value.

Overall, excellent agreement was found between the fold changes obtained using the previous read-mapping and one dataset per sample [Chapter 3] and the results obtained with three datasets per sample (Figure 4). In the previous analysis, a few genes were found to have infinite fold changes, due to no reads aligning in one of the two samples. Due to practical reasons, the fold changes in such cases were changed to a value that was twice the maximal fold change that could be calculated. The infinite Log 2 fold changes in the original dataset were changed to 12.03 (2×11.03). For the analysis using more RNA-seq data, the infinite Log 2 fold changes were changed to 12.48 (2×11.48). Five cases of infinite fold changes were found for the original analysis, but only two when more data were used. Thus, the lower limit of detection was decreased (improved sensitivity) by using more NGS data. The same was done for negative fold changes. While two occurrences were found where no reads were mapped to a gene in the xylose condition, no such cases occurred with the improved analysis. Figure 4 shows that the previously infinite fold changes were changed to significantly lower and more realistic values.

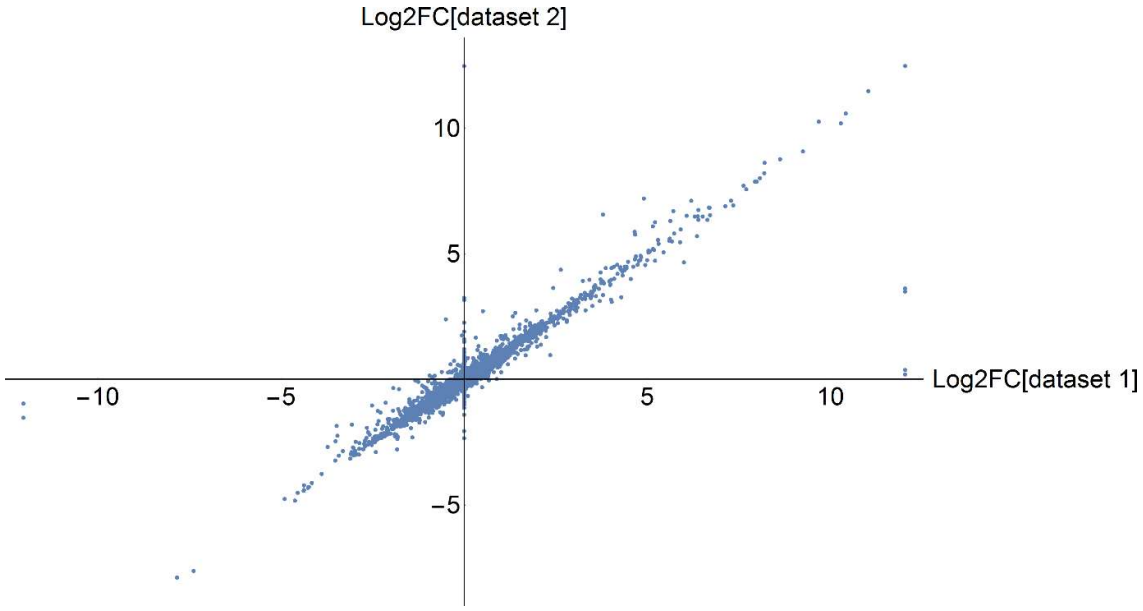


Figure 4. Correlation between RNA-seq analyses in terms of fold changes (Log_2 of FC) from glucose to xylose as the carbon source, using one and three datasets per sample. Dataset 1: previous analysis, using one dataset per sample (Chapter 2); Dataset 2: improved analysis, using three datasets per sample.

An improved *de novo* assembly using the Velvet-Optimizer

When using the Abyss assembler in constructing the first draft genome, a script was developed that allowed optimisation of the assembly based on only one parameter, the k-mer length (see Chapter 2). An optimal value for the k-mer length was found to be 46. In Galaxy, the Velvet-Optimizer assembler was recently added, which contains an automatic optimisation routine that optimises various parameters simultaneously. The automated Velvet Optimizer was used to determine if a better *de novo* assembly could be obtained. Figure 5 shows the distribution of contig lengths for the previous and the new draft genomes. The N50, N75 and N90 values are useful statistics to determine the quality of an assembly. These are summarised in Table 1. Whereas the previous genome was 10 695 063 bp in 1 094 contigs equal or longer than 200 bp, the new genome had 10 700 656 bp in 633 contigs equal or longer than 200 bp. The Velvet-Optimizer assembly thus consisted of an additional 5 593 bp.

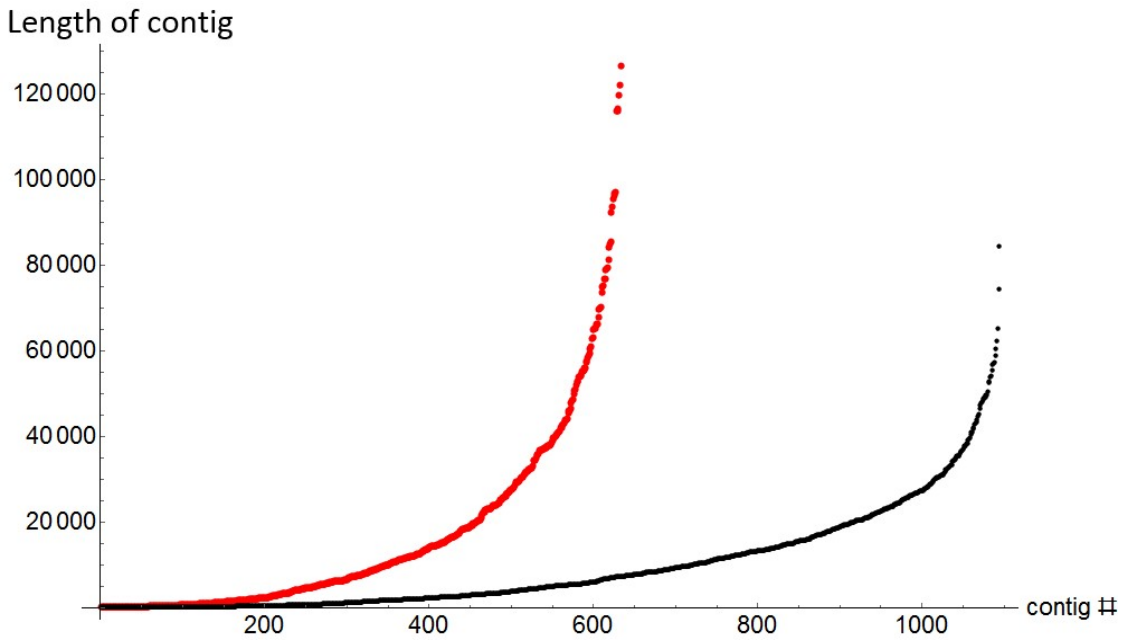


Figure 5. Distribution of contig lengths from *de novo* assemblies. Black: Abyss assembly. Red: Velvet-Optimizer assembly.

Table 1. Statistics for *de novo* assemblies based on N values.

	Total length	Contigs	N50	N75	N90
Abyss	10 695 063	1 094	21 248	12 075	5 789
VelvetOptimiser	10 700 656	633	39 193	20 399	10 088

In the improved draft genome, the longest contig was almost 50% longer, and the N50 value double the length of the previous. However, since the reason for the fragmentation with *de novo* short-read assembly was mostly due to repetitive regions, it may not be possible to improve the assembly using only these data. Longer reads may have to be generated on a different platform such as Illumina MiSeq, which generates 300 bp lengths, or on PacBio or another single-molecule sequencer that generates much longer reads, but which are not widely available yet. Alternatively, long mate-pair sequencing could be performed on an Illumina system.

Read-mapping of UFS-Y2791 against the complete genomes of DMKU3-1042 and KCTC-17555

A different strategy used for obtaining an improved genome for strain UFS-Y2791 was to perform read-mapping of genomic NGS data to a complete genome. Out of a total of the 14 495 422 paired-end genomic DNA reads used, 11 778 010 (81.25%) could be mapped to the DMKU3-1042 genome

using Bowtie2 in Galaxy. Similarly, 11 611 290 reads (80.10%) could be mapped to the KCTC-17555 genome. Tables 2 and 3 show that the two complete genomes have very similar chromosome lengths. However, the genome of strain DMKU3-1042 contains the mitochondrial chromosome in addition to the eight genomic chromosomes. As Lertwattanassakul et al. [2015] also stated, this is a more completely validated genome assembly as opposed to previously published genomes for *K. marxianus*. Upon inspection of resequencing output files, a low sequence coverage was found, suggesting that other routes of investigation were more promising.

Table 2. Chromosome statistics of the KCTC-17555 genome.

Chromosome	Length
JH924896.1_scaffold1	1 738 350
JH924897.1_scaffold2	1 700 508
JH924898.1_scaffold3	1 577 254
JH924899.1_scaffold4	1 410 702
JH924900.1_scaffold5	1 336 893
JH924901.1_scaffold6	1 198 968
JH924902.1_scaffold7	937 562
JH924903.1_scaffold8	909 243

Table 3. Chromosome statistics of the DMKU3-1042 genome and consensus of the UFS-Y2791 reference mapped to the DMKU3-1042 genome.

Chromosome	Length	Consensus length	Consensus %
AP012213.1_Chr.1	1 745 387	1 663 211	95.29
AP012214.1_Chr.2	1 711 476	1 642 304	95.96
AP012215.1_Chr.3	1 588 169	1 525 613	96.06
AP012216.1_Chr.4	1 421 472	1 375 132	96.74
AP012217.1_Chr.5	1 353 011	1 298 216	95.95
AP012218.1_Chr.6	1 197 921	1 141 744	95.31
AP012219.1_Chr.7	963 005	903 694	93.84
AP012220.1_Chr.8	939 718	874 987	93.11
AP012221.1_mitochondrial	46 308	38 525	83.19

Improved RNA-seq data analysis

Excellent mapping of sequence reads to the original *de novo* draft assembly was found, varying from 89.9% to 96.7% of reads successfully mapped to the genome, depending on the data file. This suggested a high per base quality of the first draft genome. Due to the fragmented nature of the draft genome, some of the reads might inevitably not have been mapped. Mapping the RNA-seq reads to the Velvet-Optimiser *de novo* assembly with more data made an improvement in that 96.7 to 97.4 of the reads mapped to the genome. When using the DMKU3-1042 genome as reference, the percentage of reads decreased as expected. However, 54.2% to 54.1% of reads could still be mapped to the DMKU3-1042 genome. This rate of mapping may still be considered to be sufficient, as it still amounted to a minimum of 4 153 414 reads mapped in one of the xylose samples and up to 10 244 755 in one of the glucose samples (Table 4).

The distributions of fold changes calculated when using different genomes as templates are shown in Figures 6 to 10. Limiting the differential calculations to only the coding regions marked as 'CDS' in the GFF3 files yielded indistinguishable results. Excellent agreement in terms of the distribution was also seen when using the DMKU3-1042 genome, compared to using other genomes.

Table 4. Read-mapping of paired-end data of strain UFS-Y2791 to the DMKU3-1042 genome.

Glucose sample 1		
paired-end direction	Left reads	Right reads
Input	16 557 856	16 557 856
Mapped	10 244 755	10 701 778
Overall read mapping rate	63.30%	
Glucose sample 2		
paired-end direction	Left reads	Right reads
Input	9 306 262	9 306 262
Mapped	5 737 682	5 927 154
Overall read mapping rate	62.70%	
Xylose sample 1		
paired-end direction	Left reads	Right reads
Input	7 991 575	7 991 575
Mapped	4 153 548	4 489 902
Overall read mapping rate	54.10%	
Xylose sample 2		
paired-end direction	Left reads	Right reads
Input	9 559 431	9 559 431
Mapped	5 295 989	5 639 053
Overall read mapping rate	57.20%	

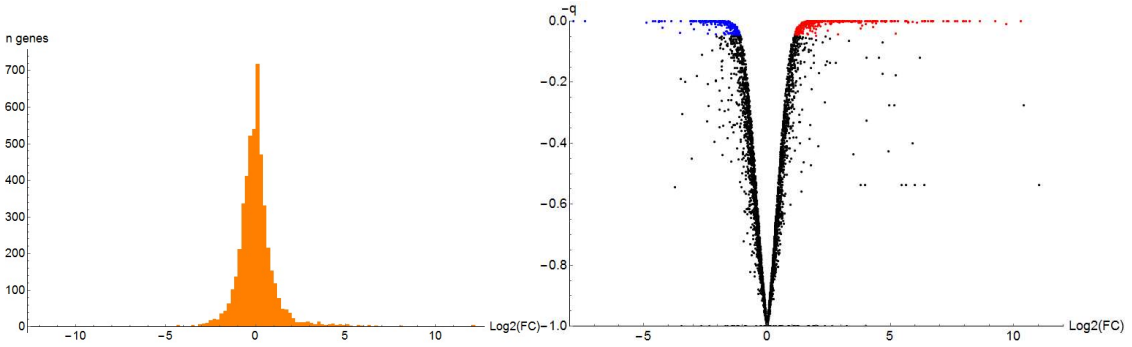


Figure 6. Distribution of fold changes in RNA-seq values with read-mapping against the Abyss assembly using one dataset per sample (from Chapters 2 and 3), in which the GFF3 annotation file was that obtained through annotation by the webAugustus web server. Left: Histogram of fold-changes. Right: Volcano plot of q-values (corrected p-values, as from CuffDiff) against Log_2 fold-changes.

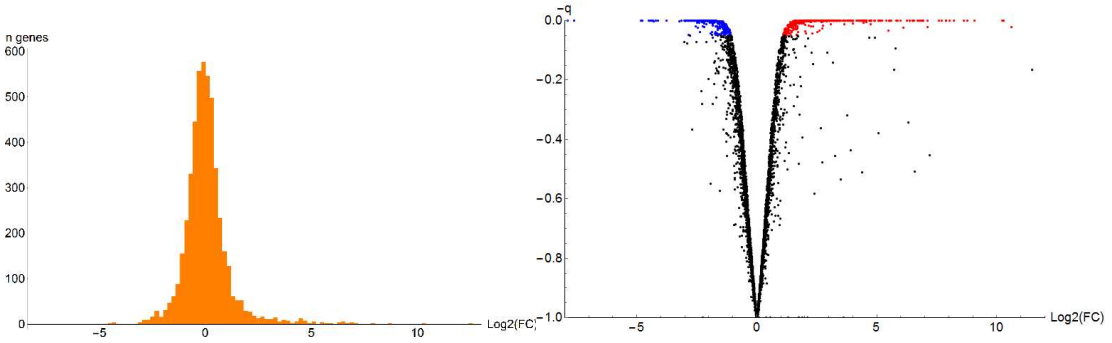


Figure 7. Distribution of fold changes in RNA-seq values with read-mapping against the Abyss assembly using three datasets per sample, in which the GFF3 annotation file was that obtained through annotation by the webAugustus web server.

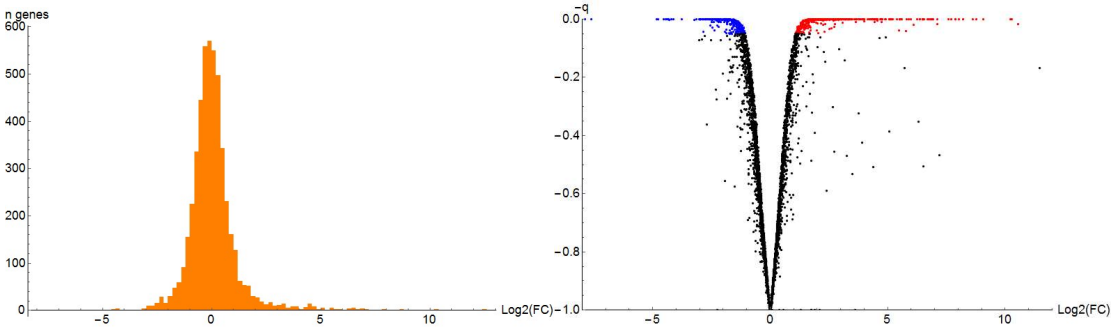


Figure 8. Distribution of fold changes in RNA-seq values with read-mapping against the Abyss assembly using three datasets per sample, in which the GFF3 annotation file was that obtained through annotation by the webAugustus web server, using only the coding sequences (CDS).

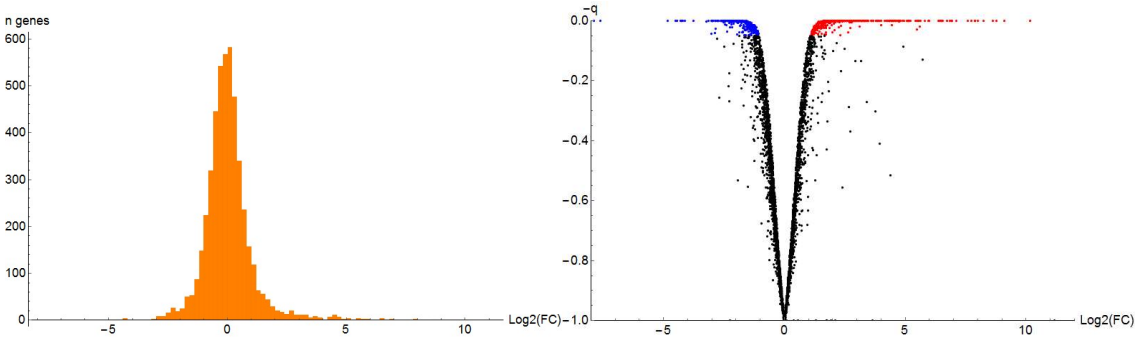


Figure 9. Distribution of fold changes in RNA-seq values with read-mapping against the Velvet-Optimizer assembly using three datasets per sample, in which the GFF3 annotation file was that obtained through annotation by the webAugustus web server.

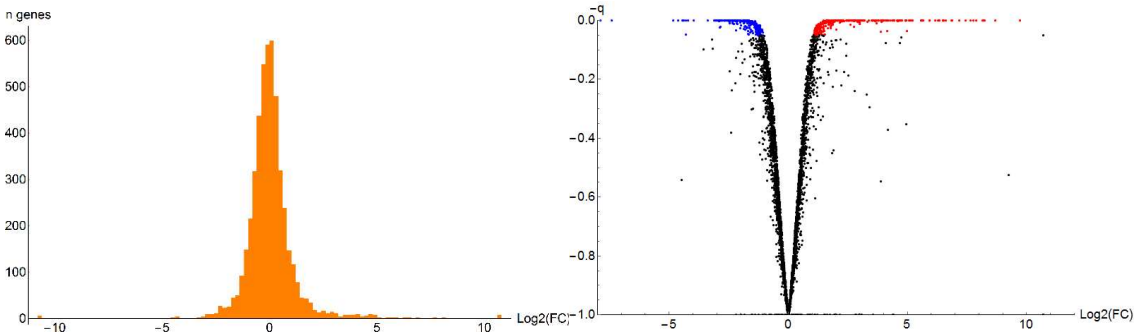


Figure 10. Distribution of fold changes in RNA-seq values with read-mapping against the DMKU3-1042 genome using three datasets per sample, in which the GFF3 annotation file was from the NCBI database, as annotated by Lertwattanassakul et al. [2015].

In terms of testing differential expression of individual genes, the highest fractions could be tested successfully in the Abyss and Velvet assemblies using all data, at 99.74 and 99.59% testabilities, respectively (Table 5). Notably, using the DMKU3-1042 genome allowed 97.16% testability, which was not only similar to the assemblies from the UFS-Y2791 strain using all RNA-seq data, but also a larger fraction compared to using the original Abyss assembly, using one dataset per sample, for which 94.81% were testable.

It was found that the latest version of the DMKU3-1042 genome on NCBI contained 4 963 protein coding genes, of which 4 952 have been annotated as distinct proteins against UniProt. There were 5 153 coding sequences (marked 'CDS'), thus some of the genes that encoded proteins consisted of more than one exon separated by introns. Introns were not included in the GFF3 annotation file of strain DMKU3-1042. A protein annotation file for Reactomica was created for the 4 952 unique genes by retrieving the UniProt annotation entries as a table from UniProt, using the 'protein_id' field that

links the entries to the UniProt identifiers. This would be used in further functional analyses (Chapters 5-8).

Table 5. Testability of differential expression of genes based on analyses using various genome assemblies.

assembly	number of ORFs	inf	neg inf	OK	NOTEST	% testable
KmaxOnAbyss full gtf 1 set	4953	5	2	4696	257	94.81
KmaxOnAbyss full gtf 3 sets	4953	2	0	4940	13	99.74
KmarxOnAbyss gene trans CDS 3 sets	4953	2	0	4940	13	99.74
KmarxOnVelvet gene trans CDS 3 sets	4861	1	0	4840	21	99.57
KmarxOnDMKU gene trans CDS 3 sets	4864	8	6	4726	138	97.16

Conclusions

In this addendum it was shown that the use of the additional RNA-seq datasets improved the accuracy of differential expression fold changes. The draft genome was also substantially improved by the Velvet-Optimizer *de novo* assembler. Notably, it was found that the complete genome of *K. marxianus* DMKU3-1042 could be used for read-mapping RNA-seq data, making RNA-seq data analysis of strain UFS-Y2791 in the context of a complete genome possible, and would also facilitate analysis of the gene regulatory programme (Chapters 5-8).

References

Afgan E, Sloggett C, Goonasekera N, Makunin I, Benson D, Crowe M, et al. Genomics Virtual Laboratory: A Practical Bioinformatics Workbench for the Cloud. PLoS One. 2015;10(10): e0140829. doi: 10.1371/journal.pone.0140829.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15): 2114-2120.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg S. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4): R36. doi:10.1186/gb-2013-14-4-r36.

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3): R25. doi:10.1186/gb-2009-10-3-r25.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Meth. 2012;9(4): 357–359. doi:10.1038/nmeth.1923.

Lertwattanasakul N, Kosaka T, Hosoyama A, Suzuki Y, Rodrussamee N, Matsutani M, et al. Genetic basis of the highly efficient yeast *Kluyveromyces marxianus*: complete genome sequence and transcriptome analyses. Biotechnol Biofuels. 2015;8(47). doi: 10.1186/s13068-015-0227-x.

Schabort DWP, Letebele PK, Steyn L, Kilian SG, du Preez JC. Differential RNA-seq, multi-network analysis and metabolic regulation analysis of *Kluyveromyces marxianus* reveals a compartmentalised response to xylose. PLoS ONE. 2016;11(6): e0156242. doi:10.1371/journal.pone.0156242.

Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics. 2008;24(5):637-644. doi: 10.1093/bioinformatics/btn013.

Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9): 1105-1111.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2013;7(3): 562–578. doi: 10.1038/nprot.2012.016.

Zerbino DR, Birney E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Research. 2008;18(5): 821–829. doi:10.1101/gr.074492.107].

Addendum 2

Multiple genome alignments and motif conservation

Abstract

The finding that RNA-seq data could be effectively read-mapped to the complete genome of strain DMKU3-1042 facilitated the construction of gene regulatory networks. The method of gene regulatory networks relies not only on the accuracy of motif finding or the availability of supporting data for transcription factor-DNA interactions, but also benefits from evolutionary conservation of a motif among closely related species. In this work, the genomes of several *Kluyveromyces* species were aligned using multiple genome alignment. To make use of this information for network construction, a method had to be developed for using multiple genome alignments in the calculation of conservation criteria at putative transcription factor binding sites. This addendum describes the software algorithms developed for this purpose, which was integrated into the likelihood framework described in Chapters 5 and 6.

Introduction

The development of a likelihood framework for the construction of gene regulatory networks was described in Chapter 5. This method required multiple genome alignments of species closely related to *K. marxianus* to calculate the degree of conservation for each TF binding site. Motifs that are conserved among sister species have a higher likelihood of being true binding sites. The progressiveMauve aligner [Darling et al. 2010] is a convenient program to use as it is fast and even is available as a desktop application in Windows. However, is was found that the format of the output files did not facilitate straightforward analysis in *Reactomica*. Therefore, a significant amount of software development had to be done to harness the outputs for calculation of conservation criteria to be used in the likelihood framework. The methods were developed here using the draft genome of

K. marxianus UFS-Y2791 as the reference strain of interest, although the procedures were later carried out with the complete genome of *K. marxianus* DMKU3-1042 as the reference (Chapters 6 and 8).

Multiple genome alignments for calculating the conservation ratio

Since non-coding DNA generally evolves rapidly compared to coding DNA, it should be best to use a variety of closely related genomes to allow sufficient similarity among genomes for proper alignment. Hence it was chosen to focus only on the *Kluyveromyces* genus. Seven complete or draft genomes of *Kluyveromyces* species were aligned using the Windows desktop version of progressiveMauve [Darling et al. 2010]. The complete genomes were for *K. lactis* and *K. marxianus* strains DMKU3-1042 and KCTC-17555, while draft genomes were available for *K. aestauri*, *K. dobzhanskii* and *K. wickerhamii*, as well as *K. marxianus* UFS-Y2791 (from this study). These are currently the only species in the *Kluyveromyces* genus for which either complete or published draft genomes are available. The order of upload of the genomes into progressiveMauve is important in that the first genome is assumed to be the reference genome. As the reference, the complete genome of DMKU3-1042 was uploaded first, which provided the correct order of contigs in other strains. As a test for multiple alignment, satisfactory alignment was found between the complete genomes of *K. marxianus* strains DMKU3-1042 and KCTC-17555 (Figure 1) with very few rearrangements.

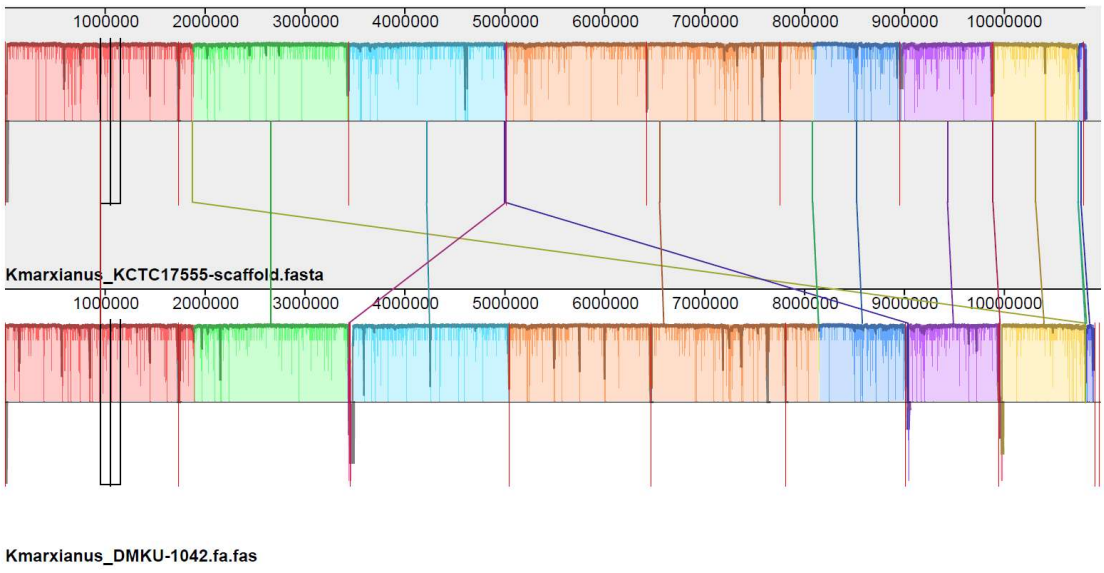


Figure 1. Alignment between the *K. marxianus* KCTC17555 and DMKU3-1042 genomes using progressiveMauve. Colours represent chromosomes and lines represent genome reshuffling during evolution.

Inclusion of all genomes resulted in a strongly crossed-over pattern, simply because the ordering and direction of the multiple contigs had not been established beforehand. Yet, the output contained alignment files organised into correctly alignment segments, which is what was of interest for the calculation of motifs.

Mapping between the segments of multiple genome alignments produced by progressiveMauve, and the positions in individual genomes was, however, a complex task. In addition, any of the genomes should be available as the reference sequence for constructing the regulatory network (the "gene regulatory reference"), even though the most complete genome should be used as the "alignment reference", which is the first genome uploaded into progressiveMauve. An efficient algorithm was required for finding segments of a particular type, which would allow straightforward analysis. For instance, an alignment segment may comprise of more than one contig for a given species. This presumably was allowed in the progressiveMauve software since it would, to an extent, assemble a more complete chromosome from contigs. However, assuming that two contigs were indeed situated next to one another on a chromosome in the gene regulatory reference, forming part of the same upstream regulatory region, while in fact this was only the case in some of the other genomes in the aligned segment due to rearrangements of the DNA between strains, may lead to incorrect results. Even if two contigs of a strain were adjacent to one another, there is no absolute certainty of the number of bases missing between the two contigs. Joining two contigs by their multiple genome alignment may even result in an artificially constructed regulatory motif. Moreover, the link between a gene and its upstream regulatory region should be clear; preferably both the gene and its upstream regulatory region must be in the same alignment segment. Various levels of stringency may thus be applied. The most strict criterion being that only those genes that have at least some upstream part of their transcript (or coding region), as well as their upstream regulatory region, on the same alignment segment for all genomes, should be considered to have multiple alignments of regulatory regions. Possible constraints applicable are: (*a*) only those parts of the multiple sequence alignment segments are useful which have ungapped alignments of the reference sequence, for which regulatory elements are sought; (*b*) multiple alignments must have at least some upstream part of their common transcript (or coding region), as well as their upstream regulatory region, on the same alignment segment for all genomes; (*c*) the upstream region needs to be long enough, preferably 1 000 bp from the translation start site; (*d*) a segment should not have fewer than a certain number of nucleotides, otherwise it becomes meaningless to use them as alignments. Out of the 2 275 segments in which *K. marxianus* UFS-Y2791 DNA featured, 1 584 (69.6%) represented segments that comprised a single

contig in *K. marxianus* UFS-Y2791 (Figure 2). From these, a total of 1 161 segments conformed to the criterion of having all seven genomic sections as DNA originating from only one contig.

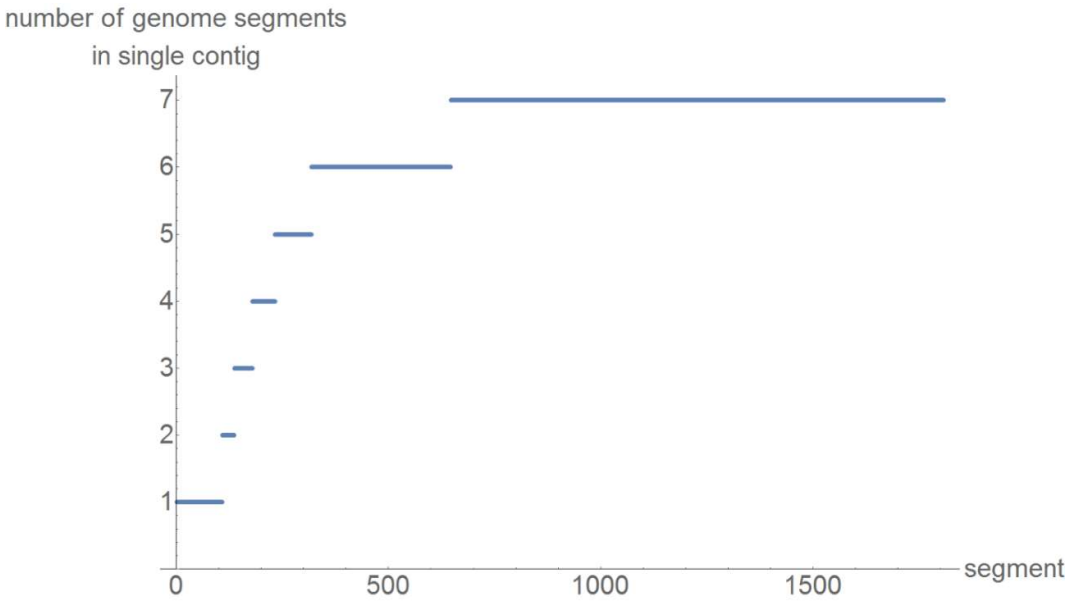


Figure 2. The distribution of the degree of non-interruption of each segment. The number of genome sections that originate from a single contig (as opposed to more than one contig brought together by multiple genome alignment) is on the Y-axis. A value of seven indicates that a segment (a multiple genome alignment between the isolates) contained only one contig per genome for all genomes.

Calculating the conservation score

From the list of segments that conformed to the criteria, alignment segment 87 was explored to develop the method. Figure 3 shows the output from a short section of segment 87, as viewed in the convenient alignment viewer designed for this purpose. The conservation score on the Y-axis was calculated as the number of bases from the seven genomes that were identical to the base in the *K. marxianus* UFS-Y2791 reference. A gap in the alignment was assumed to add zero to the score at a nucleotide position. A score of 0 arose when the reference base was a gap. A score of one was obtained when none of the other six bases corresponded. A score of two was, for instance, obtained when the reference character was C and others were -, -, -, C, C and -. Figure 4 shows the conservation scores of the complete segment 87. Peaks of high conservation are visible, with interspersed regions of low conservation (with respect to the strain UFS-Y2791 reference genome). To remove noise, a sliding filter (moving average) was applied at a window size of 10 bp (Figure 5). Small islands of conservation at this window size may correspond to conserved motifs, since the average width of DNA

binding sites is approximately 7 bp. Wider peaks of 100 bp or more should correspond to genes. The window size for the sliding filter was increased to 100 bp to filter out conserved motifs in divergent DNA to retain only peaks corresponding to genes (Figure 6).

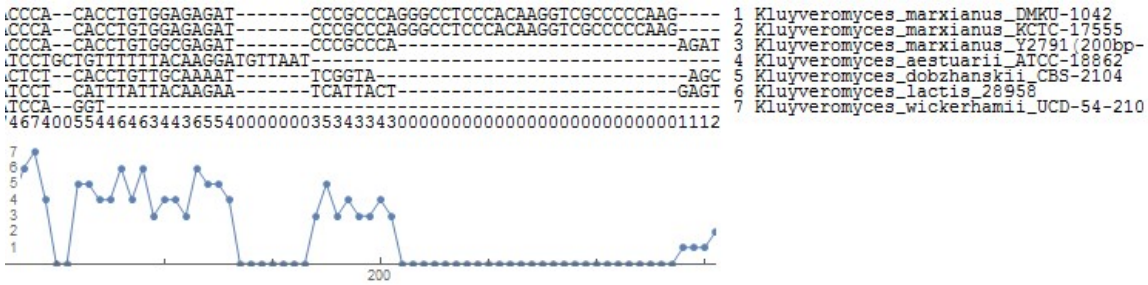


Figure 3. Alignment view of a section in segment 87 showing the conservation scores.

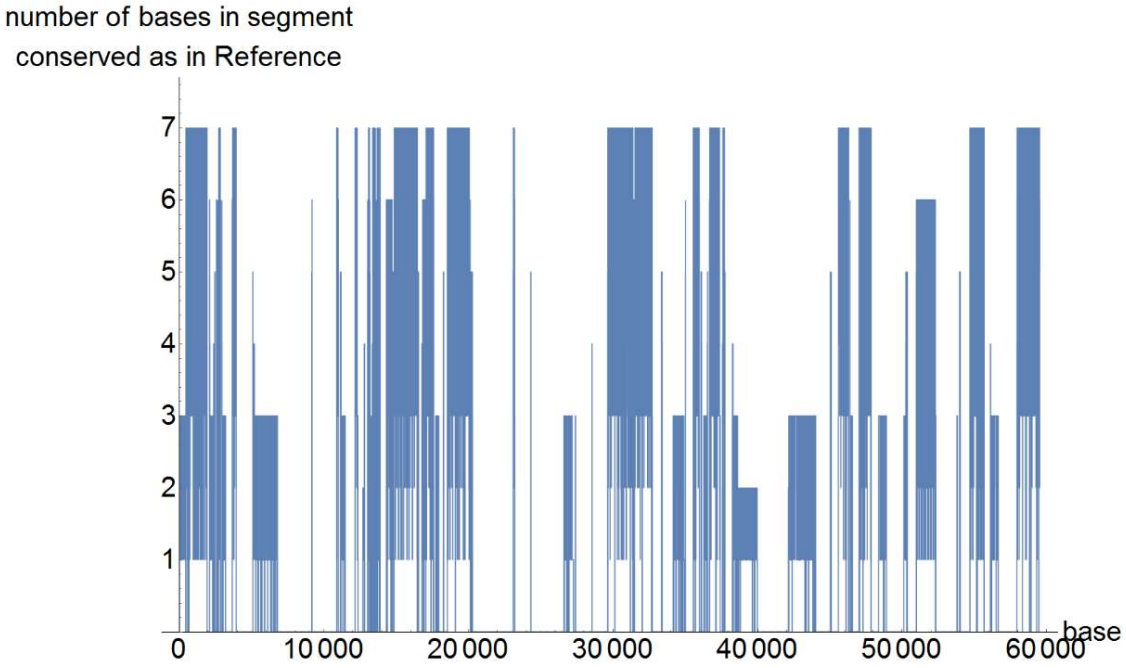


Figure 4. Conservation view of the complete segment 87 of the gapped alignment. The resolution is 1 bp.

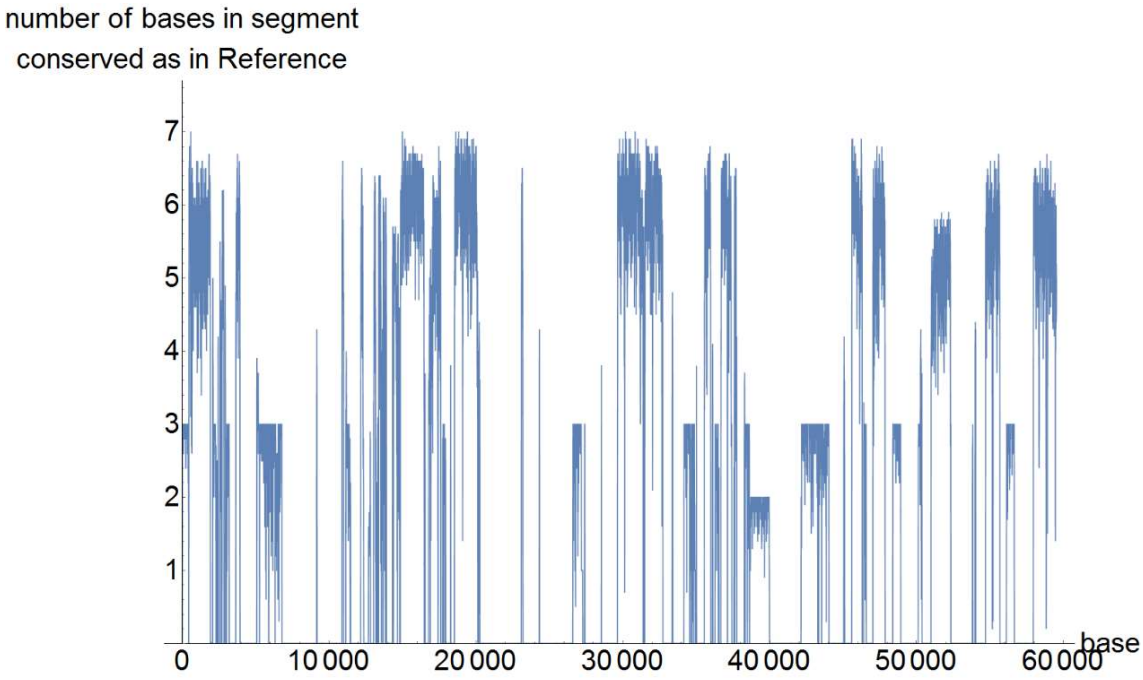


Figure 5. Conservation scores of the complete segment 87 of the gapped alignment. A 10 bp sliding window was used to remove noise. Genes and potential conserved regulatory elements are visible as peaks.

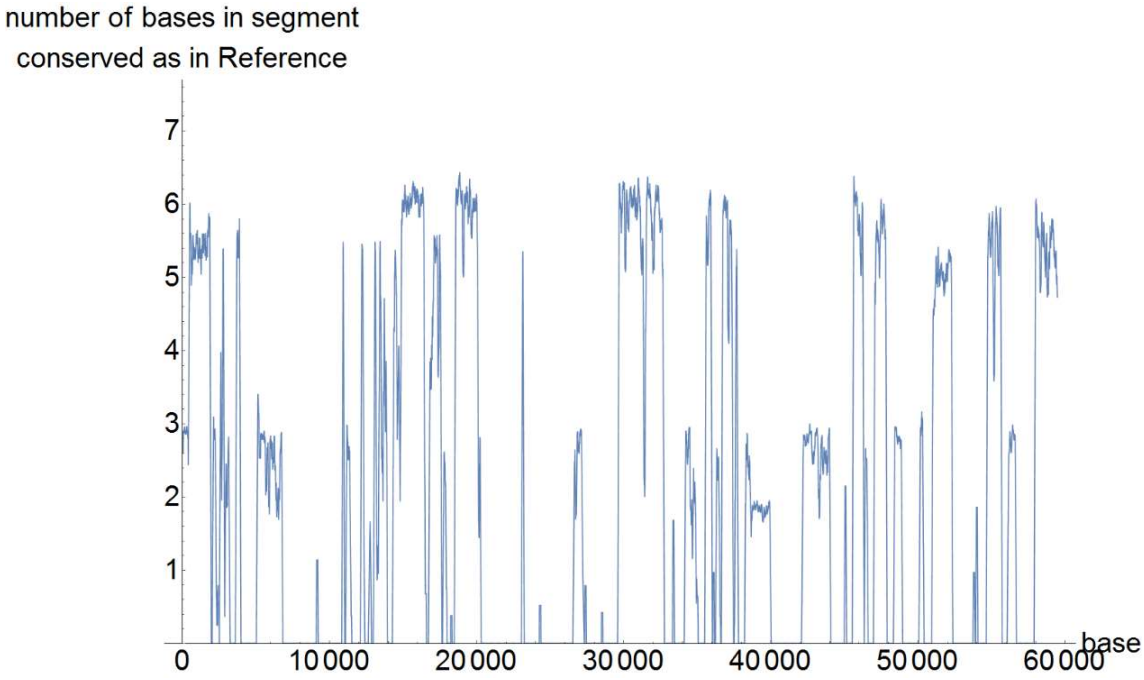


Figure 6. Conservation view of the complete segment 87 of the gapped alignment. A 100 bp sliding window was used to remove noise. Genes are visible as peaks.

A method for collapsing gapped alignment segments into ungapped alignments

Due to the complexities of mapping alignments to the reference genome, a simpler underlying data format was required. To achieve this, a gapped alignment was spliced into a continuous section in which the reference genome had no gaps, while gaps were allowed in other genomes in the alignment. Insertions in the sister genomes were removed in effect, and a suitable maker system for the deletions was introduced (see below). The marker system was introduced not only for the purpose of keeping track of removed insertions in sister genomes, but in particular to penalise conservation scores at the edges where a DNA section was deleted. One major advantage of the ungapped alignment format is that it could be viewed along with genome annotations of the reference strain in an "annotation track'-like fashion, as popular in genome browsers. The same resolution settings were applied for rendering Figures 7-9. In effect, the gapped alignment adopts the coordinate system of the reference genome and this renders obtaining the conservation score much simpler.

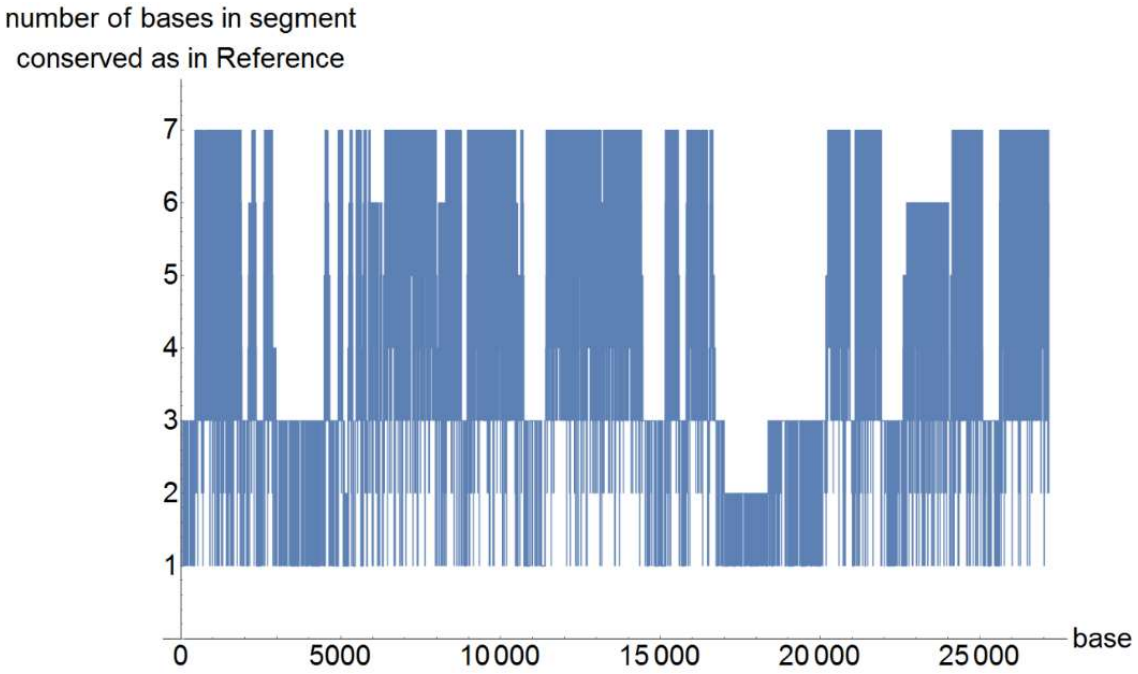


Figure 7. Conservation view of the complete segment 87 of the ungapped alignment, removing non-reference DNA sections where the reference genome had a gap, at 1 bp resolution.

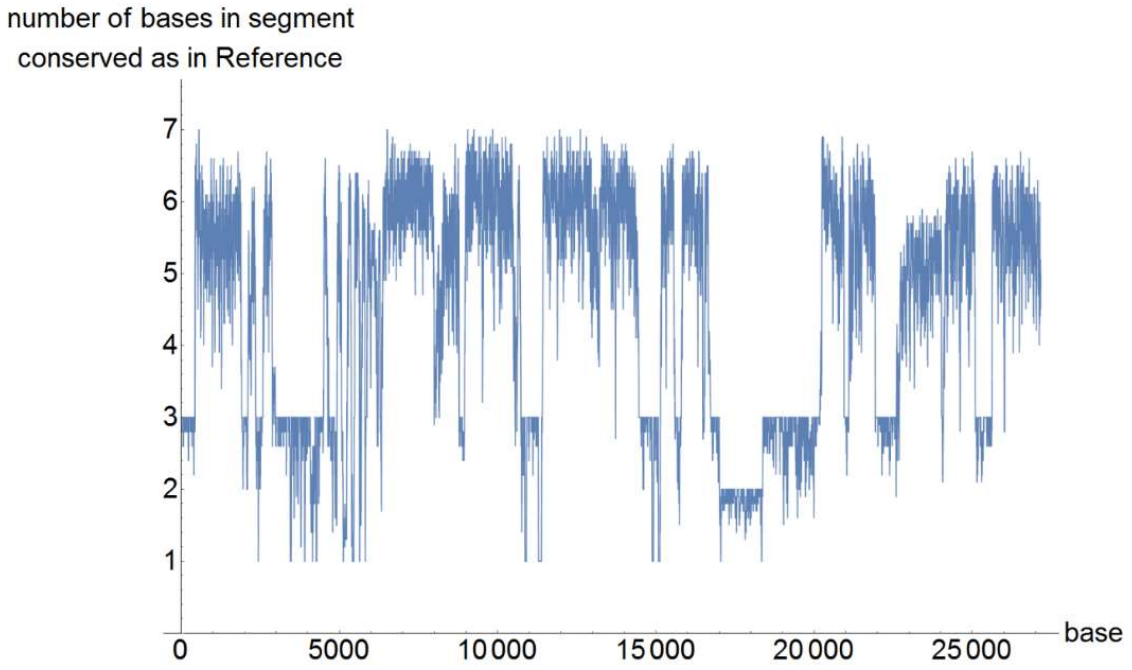


Figure 8. Conservation view of the complete segment 87 of the ungapped alignment, removing non-reference DNA sections where the reference genome had a gap, at 10 bp resolution.

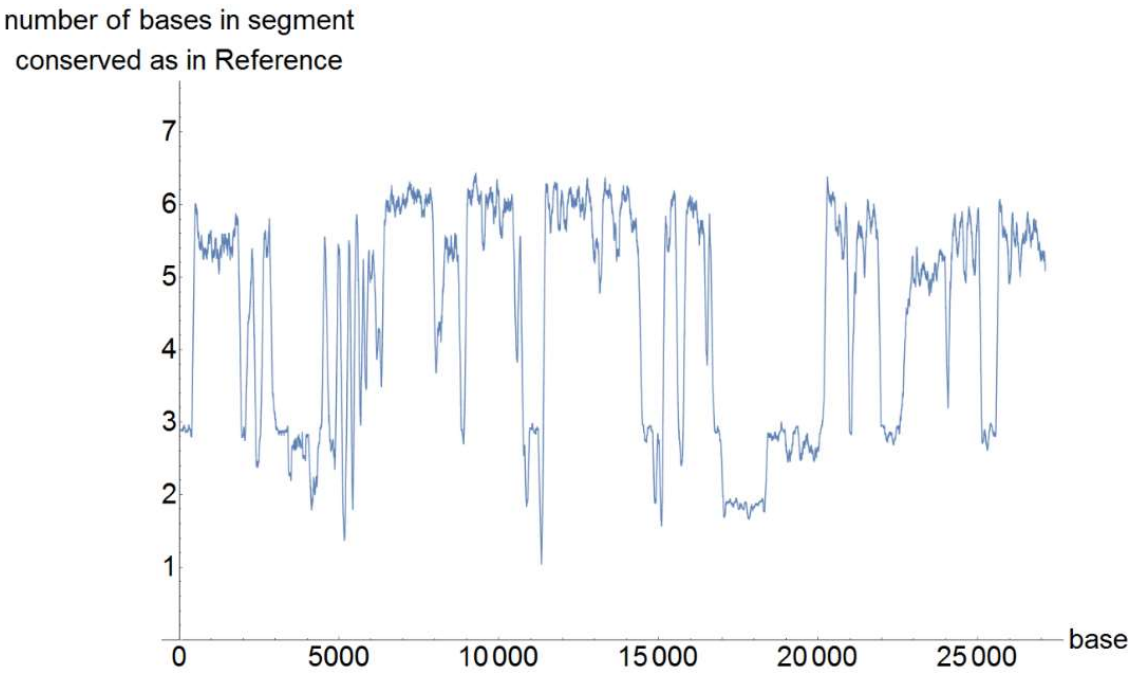


Figure 9. Conservation view of the complete segment 87 of the ungapped alignment, removing non-reference DNA sections where the reference genome had a gap, at 100 bp resolution.

An artefact that may arise is that which occurs when an average conservation score is calculated over a region (aligning to the reference genome) that spans a gap that has been removed from the non-

reference genomes. A conservative rule that may be applied for simplicity is that an averaged conservation score over the bases in a putative binding motif can be calculated as long as all sequences were ungapped in the sliding window. The marker system would signify these removed gaps.

However, it was observed that often a gap in the reference genome was introduced by single insertions of DNA sections in single, non-reference genomes, while the other genomes aligned perfectly with the reference genome. An identical long gap in both the reference genome and five of the homologous non-reference genomes should only penalise for one of the genomes – the one with the insertion relative to the reference. An additional vector (marker system) was calculated to capture the level of uncertainty introduced into the conservation score by insertions in non-reference genomes. Various possibilities exist to calculate such a penalty score. The first (*a*) is to calculate the total number of nucleotides in all genomes existing in this inserted section. Another (*b*) is to take the latter number and normalise it by the total number of characters in the section, to counteract the penalty by the number of gaps in non-reference genomes (actually, the lack of insertions). In the latter scenario, the total length of the insertion is irrelevant. A long insertion in a single non-reference genome would effectively be converted to a score of 1/6 for six non-reference genomes. However, a poorly aligned section with many short insertions in multiple non-reference genomes may have the same result, which should be avoided. Yet another approach (*c*) is to count the number of non-reference genomes which contains at least one insertion. The latter is the most trivial to use in further analyses, since the meaning is easier to understand than with methods (*a*) and (*b*). A value of one means that only one of the six non-reference genomes contained an insertion and thus perhaps the insertion could be ignored. The higher this value, the more uncertain the sequence conservation was. Method (*c*) was the approach taken in the rest of this study. Figure 10 demonstrates the procedure of converting a gapped alignment into an ungapped alignment and capturing the number of sequences that had at least one gap along the edges of the deleted sections.

From Table 6 it is evident that the majority of insertions (compared to the reference genome) was introduced by only one or two genomes in an alignment segment (169 insertion sections), whereas only in one case (six non-reference genomes as insertion), the insertion should rather be interpreted as a deletion from the reference genome with reference to the common ancestor of the seven genomes.

A	A	T	G	C	C	C	T	G	C	C	*K. marxianus DMKU*
A	G	T	G					G	A	C	*K. marxianus KCTC*
A	**G**	**T**	**G**					**G**	**C**	**C**	***K. marxianus UFS-Y2791***
A	A	T	G					G	C	C	*K. aestauri*
A	A	T	G	C	C	A	T	G	C	C	*K. dobzhanskii*
A	A	T	G	C	C		T	G	A	C	*K. lactis*
A	A	T	G	C	C	A	T	G	C	C	*K. wickerhamii*
7	2	7	7	0	0	0	0	7	5	7	conservation score

A	A	T	G	G	C	C	*K. marxianus DMKU*
A	G	T	G	G	A	C	*K. marxianus KCTC*
A	**G**	**T**	**G**	**G**	**C**	**C**	***K. marxianus UFS-Y2791***
A	A	T	G	G	C	C	*K. aestauri*
A	A	T	G	G	C	C	*K. dobzhanskii*
A	A	T	G	G	A	C	*K. lactis*
A	A	T	G	G	C	C	*K. wickerhamii*
7	2	7	7	7	5	7	conservation score
0	0	0	4	4	0	0	gap score marker

Figure 10. Method for collapsing gapped alignment segments into ungapped alignments. Top: gapped segment where the reference genome (*K. marxianus* UFS-Y2791) contained a gap. Bottom: ungapped alignment with the insertion removed. In this case, the gene regulatory reference was the UFS-Y2791 draft genome, even though the multiple genome alignment reference was the complete genome of strain DMKU3-1042. Note the gap score marker at the bottom which captures on both sides of the insertion the number of genomes that contained DNA in this section (an insertion relative to the reference genome, which was removed in the ungapped alignment).

Table 6. Gap statistics for segment 87. Distribution of the number of sequences present as insertions compared to reference sequence. The total length of the reference DNA in the segment was 26 766 bp. The majority of insertions (which were removed in the ungapped alignment) was introduced by one or two genomes, whereas only for very few such cases these insertions should rather be interpreted as deletions from the reference genome, compared to the reference genome (UFS-Y2791).

Number of non-reference genomes in insertion section present as insertion	Insertion sections with 1-6 genomes present as an insertion
1	87
2	82
3	18
4	10
5	3
6	1

The total length of the *K. marxianus* UFS-Y2791 DNA as uninterrupted segments was 9 085 954 bp, whereas the total length of the draft genome was 10 695 463 bp. Thus, 84.95% of the *K. marxianus* UFS-Y2791 draft genome was associated with a multiple alignment, even though the draft genome was separated into 1 096 contigs longer than 200 bp. This suggested that for a large majority of genes, a conservation score should be available to improve the certainty of assigning DNA binding sites. The distribution of the lengths of stretches of DNA in the reference draft genome in alignment segments is shown in Figure 11. These were only those segments that contained DNA from a single contig in the reference strain UFS-Y2791. The ungapped alignment, along with conservation scores and gap marker scores, was converted to string format, effectively compressing the large file 37.74-fold.

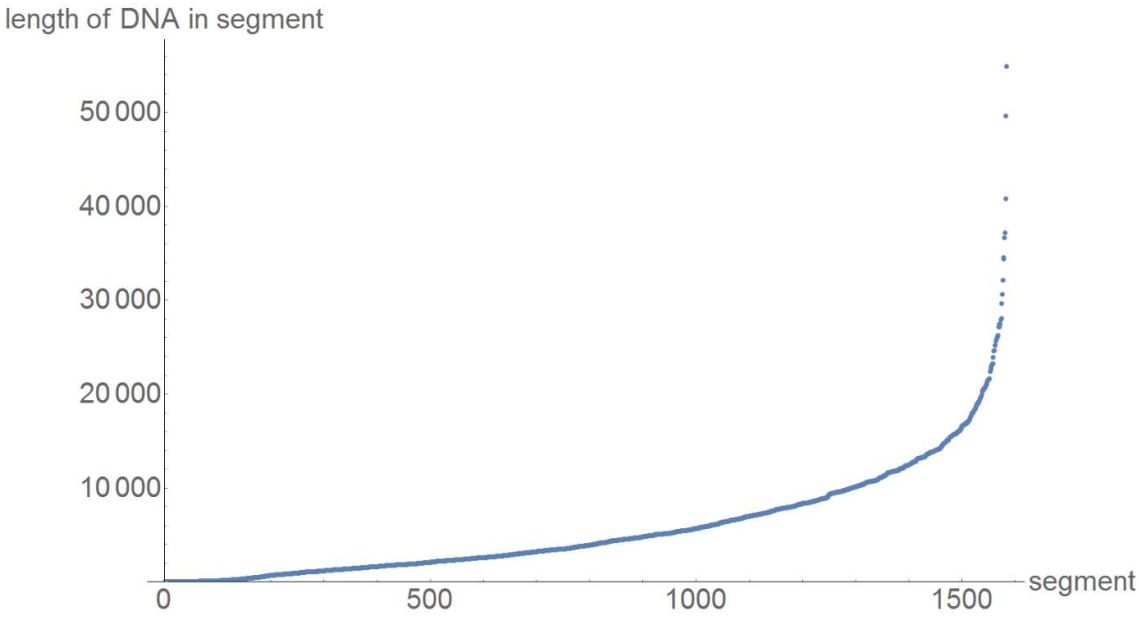


Figure 11. Length distribution of the *K. marxianus* UFS-Y2791 DNA in processed, ungapped multiple genome alignment segments. Only those segments spanning single contigs for *K. marxianus* UFS-Y2791 were included in this analysis. The total length of this aligned genomic DNA was 9 085 954 bp.

Conclusions

A method and accompanying software was also developed to harness the output from the progressiveMauve multiple genome aligner by converting data files into a format adopting the coordinate system of the reference species and calculating sequence conservation scores for the whole genome. This was instrumental in the likelihood method of constructing gene regulatory networks used in Chapters 5 and 6.

Reference

Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PloS ONE. 2010;5(6): e11147. doi:10.1371/journal.pone.0011147.

Addendum 3

Addendum for the likelihood method for gene regulatory networks - Chapter 5

This addendum describes further details of methods from Chapter 5.

Exploring the relationship between average nucleotide score, motif length and the motif match likelihood, *Lm*

Lm was calculated for every PPM length in the JASPAR database, from 5 to 21 bp, and at various values for *Sn*, from 0.3 to 1 using the equation below.

$$Lm = \frac{\prod_i^n Sn}{\prod_i^n b}$$

Figure 1 shows this relationship. A 21 bp motif which only allows perfect matches (*Sn* = 1) would thus accumulate scores between to 10^{10} and 10^{11}. At the other end of the spectrum, only a PPM that is at least 17 bp long could result in an *Lm* value larger than 10 if it were very degenerate (*Sn* = 0.3). A cutoff for *Lm* at 10 seems reasonable, and means that the sub-sequence was 10 times more likely to match the PPM model than the background model. When *Sn* was increased to 0.4 (0.15 above the average background of 0.25), shorter motifs of 9 bp could be identified. Although this states some form of a basis to reason about the significance of a motif match, a strict cut-off should not be decided on at this stage, since (a) background frequencies are not 0.25 and are positionally biased, and (2) multiple other criteria will be included to decide on the final inclusion of a TF-gene interaction. Inevitably, to save computer memory and speed up the algorithm, some cut-off had to be set for inclusion of a motif into a motif likelihood table as a potential binding site. It was found that when a cut-off lower than 10 was set, millions of potential motifs would have to be analysed further for the draft genome (see below).

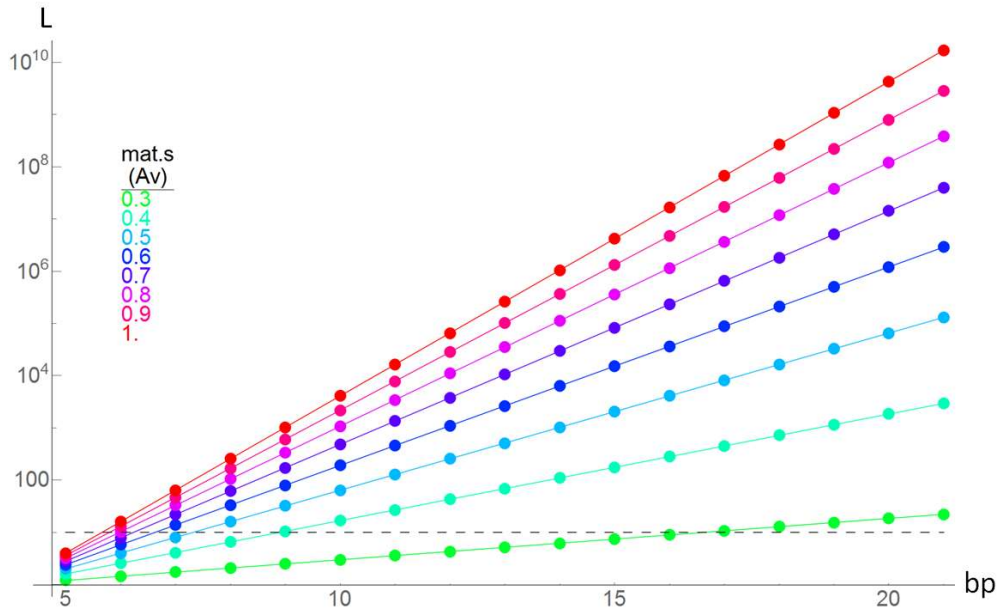


Figure 1. Exploring the relationship between average nucleotide score, motif length and the motif match likelihood, *Lm*. The X-axis reflects the length of a motif (PPM). Different lines represent different average nucleotide scores (*Sn* or mat.x), assuming a position independent background frequency of 0.25 for each nucleotide. Long motifs could accumulate high scores, compounded by high nucleotide scores.

The conservation score *Lc* from multiple genome alignments

Figure 2 shows the alignment of the reversed motif 9634 to the UFS-Y2791 genome as well as six other *Kluyveromyces* genomes. The sequence conservation scores and the gap marker scores were calculated as described in Addendum 2. The gap marker scores capture the penalties that could be applied to the background conservation scores due to deleting regions from sister species to convert the gapped alignment into an ungapped alignment (see Addendum 2 for details). The vector of conservation scores (indicated by 'c' in the alignment) was separated into a region corresponding to the motif (red) and two flanking regions of 20 bp. The latter was be used to estimate a background conservation score, which could be compared to that of the motif. The best length of the background flanking regions was not determined empirically at this stage, but limiting the length was required, since the multiple genome alignment only resulted in segments of alignment between genomes.

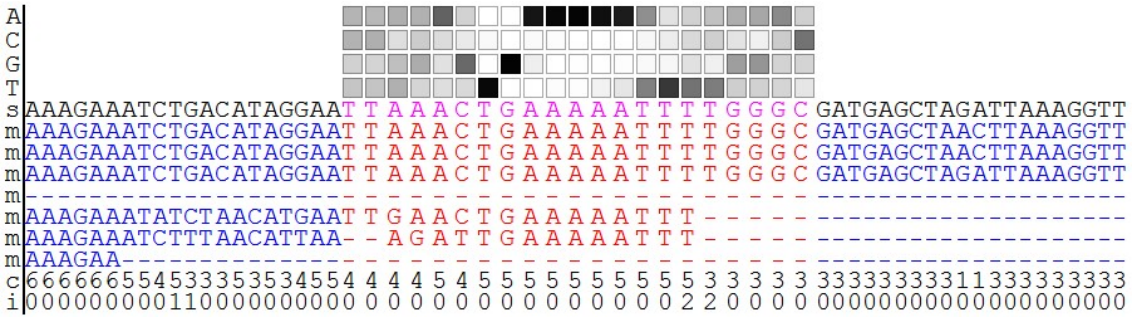


Figure 2. Multiple genome alignment view of a PPM matching to a candidate motif. The region in red indicates the frame corresponding to the PPM. The regions in blue indicate the flanking regions used to calculate the background conservation score. The row indicated by 's' specifies the genome of interest (*K. marxianus* UFS-Y2791). Rows specified by 'm' indicate sister species. The row specified by 'c' indicates the conservation score. The row specified by 'i' indicates the gap marker scores included by deleting regions from sister species to convert the gapped alignment into an ungapped alignment.

A possible concern was the presence of gaps in the alignment as depicted in Figure 2 (see Results and Discussion in Chapter 5). If the background conservation was to become very small, an unrealistically large *Lc* would be obtained. An important consideration was also that for genome alignment segments, significant stretches may occur in which the reference genome aligns with no other genome, where the conservation score would be calculated as zero, completely removing the motif from further analysis. Inability of the genome aligner to align phylogenetically distant sequences may thus have profound effects. It was considered to be a good strategy to treat this likelihood ratio partly as a quantifier and partly a qualifier, by not allowing *Lc* values to be lower than some cut-off value (0.5).

The confidence spectrum for the likelihood based on a common interaction, *Li*

A custom scoring system was designed to capture confidence in experimentally observed TF-target gene interactions of various types (Table 1). At the low end of the confidence spectrum, "microarray RNA expression" indicated correlations between expression of a TF gene and a target gene. Somewhat stronger evidence was found in "genome-wide gene expression regulator mutant expression profile" which indicates that by deleting a TF, there was a significant change in the expression of the target gene. A further improvement was found in "genome-wide gene expression regulator binding enrichment with conserved binding site", indicating not only a correlation between the expression of TF and target, but also that the relevant motif was present. The strongest evidence is likely found for interactions annotated as "chromatin immunoprecipitation-chip assay" and "chromatin immunoprecipitation-chip assay". In these assays, direct physical evidence was brought to light for an

interaction between the TF and the target gene upstream regulatory region. A value of one was assigned for no experimental evidence.

Table 1. Data from high-throughput experiments with *S. cerevisiae* capture in the SGD. The value *Li* is the likelihood statistic assigned here to capture the confidence in each type of experiment. Evidence of direct physical interaction was preferred over correlations in expression level. Data originated from a multitude of datasets found in SGD.

Annotation Type	Experiment Type	n	Li
None	None	1402710	1
binding enriched	genome-wide gene expression regulator binding enrichment	4053	2
expression activated	genome-wide gene expression regulator mutant expression profile	2573	2
	microarray RNA expression	2394	1.5
binding enriched with conserved binding site	genome-wide gene expression regulator binding enrichment with conserved binding site	2360	4
expression repressed	genome-wide gene expression regulator mutant expression profile	807	2
expression repressed	ethanol/glucose limitation	749	1.5
binding enriched	chromatin immunoprecipitation-chip assay	518	8
expression activated	ethanol/glucose limitation	379	1.5
activated	microarray expression profiling	123	1.5
expression repressed	microarray RNA expression	50	1.5
expression activated	microarray RNA expression	29	1.5
	chromatin immunoprecipitation-chip assay	14	8
	ethanol/glucose limitation	3	1.5

Calculation of likelihood rank ratio *L,* and construction of gene regulatory networks

The likelihood rank ratios *L* of all TF-target interactions was calculated using the *Lm*×*Lc*×*Lg*×*Li* likelihood function (see Materials and Methods in main text). The number *n* of top-scoring motifs were converted into *in silico* gene regulatory networks. Figure 3 shows the distribution of number of targets per TF for each of 28 gene regulatory networks. The number of large target gene sets increased as *n* increased. For the smallest network (the best 1 000 motifs), only 25 TFs were found, and the a little more than 100 target genes were included in the network. As the number of interactions increased, so did the number of target gene sets, which corresponded to more TFs in the network. Figure 4 shows the enrichment scores for each network and relates the scores to the names of TFs. Notably, the high-scoring TFs have small target gene sets, which was expected, since large gene sets would be an indication of incorrect assignments from poorly matched motifs. The TF Gcn4 was the top scoring TF

in most networks. It was also striking that an optimum in the enrichment statistic apparently existed, and this is also where Gcn4 was predominant. An optimum enrichment statistic for a TF may indicate that the true number of targets was found for that number of interactions in the network. With increasing numbers of motifs allowed, the true motifs are discovered, but as more motifs are allowed in the larger networks, the increasing rate of false assignments presumably lowered the scores and diluted the effect of the true positives. It was interesting that the same trend existed for Bas1 and Yap1 with an optimum size of the total gene regulatory network of about 4 000 motifs.

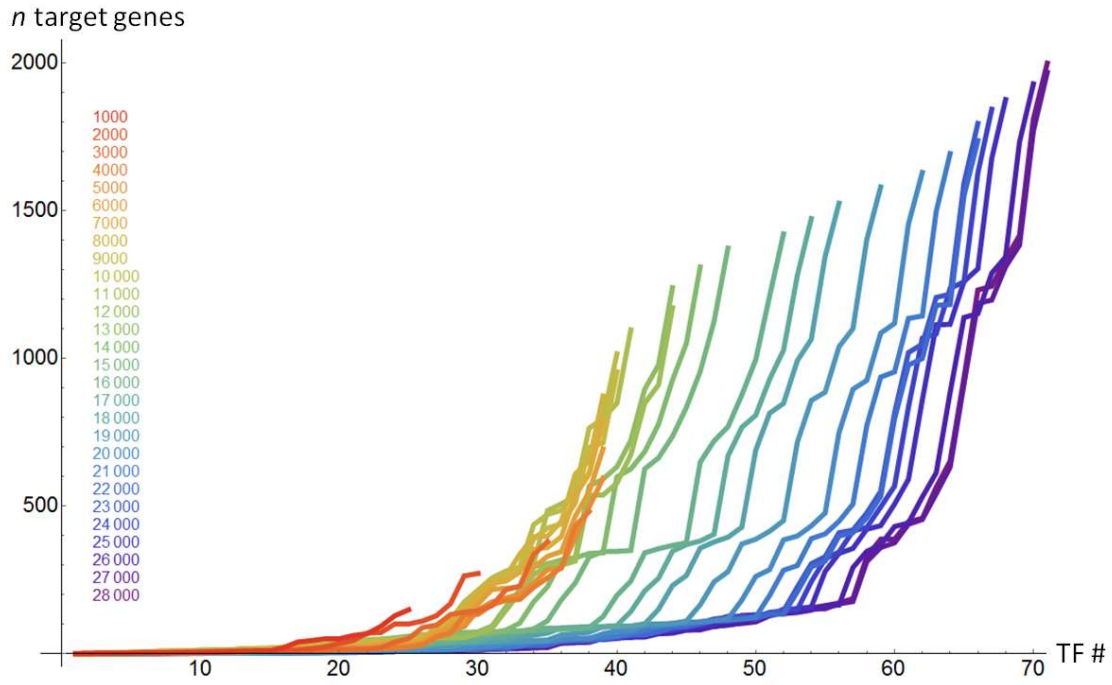


Figure 3. Distribution of number of targets per TF for each of 28 gene regulatory networks using *Lm*×*Lc*×*Lg*×*Li* as the likelihood function. Each line represents a network and the colours indicate the number of motifs considered.

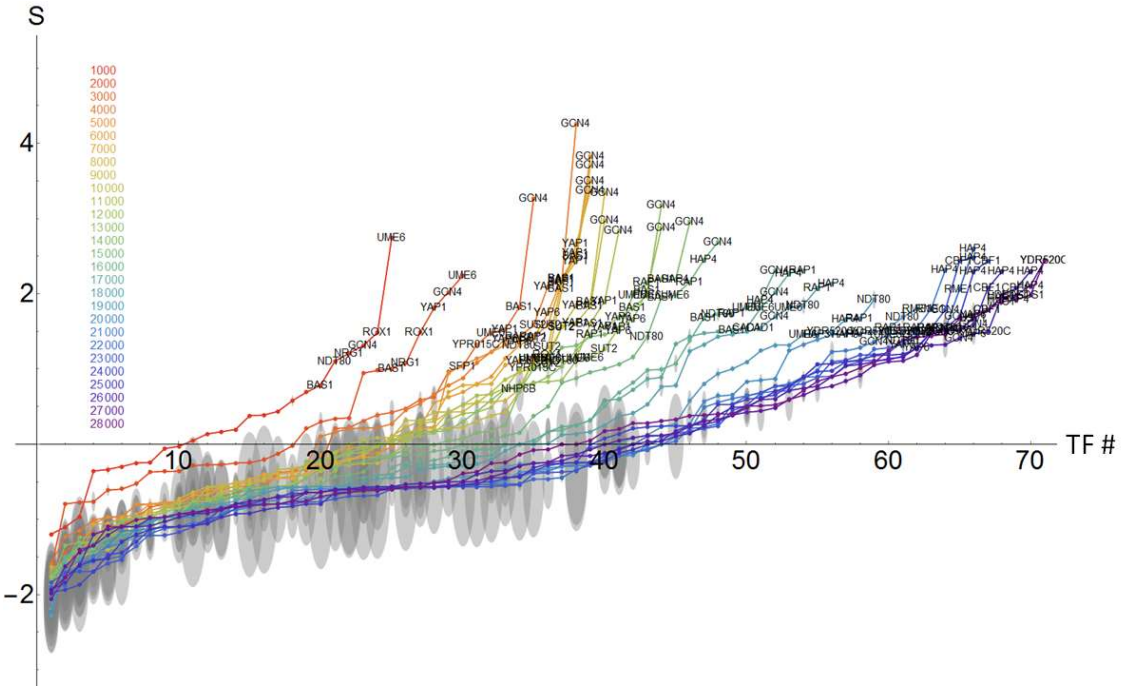


Figure 4. Distribution of enrichment score per TF for each of 28 gene regulatory networks using *Lm*×*Lc*×*Lg*×*Li* as the likelihood function. Each line represents a network and the colours indicate the number of motifs considered. A height of 1 in grey disks equals 1 000 targets. A strong optimum enrichment is present when including the best scoring 4 000 motifs.

The likelihoods of networks with sizes 1 000, 4 000 and 28 000 were investigated, especially the contribution of various sources of information to the final likelihood. Figure 5 shows a breakdown of the final likelihood, *L* (grey dotted line) of the 1 000 motif network into the contributing likelihoods for *Lm*, *Lc*, *Lg* and *Li*. It can be seen that the motif match likelihood *Lm* played a dominant role in the assignment of a high score. Hence, this scheme selects for long motifs and shorter motifs will only appear in the network if a larger network is chosen, as also shown in Table 2. This explains the observation of only a few TFs in the gene regulatory networks, since only 70 were found for the larger network. The distributions of likelihoods seem very similar between the 1 000, 4 000 and 28 000 motif networks (Figures 5, 6 and 7). In the 4 000 motif network, Gcn4, Bas1 and Yap1 could be considered as significantly enriched, assuming a cut-off at 1.67 (p = 0.05). Gcn4 seemingly regulated seven out of its 18 target genes and has a large motif of 21 bp. The 28 000 motif network included a different set of TFs, mostly only on the borderline of significance (YDR520C, Eds1, Hap4, Mac1, Cbf1, Yap6, Usv1), suggesting that the network contained too many false positives due to the large number of interactions.

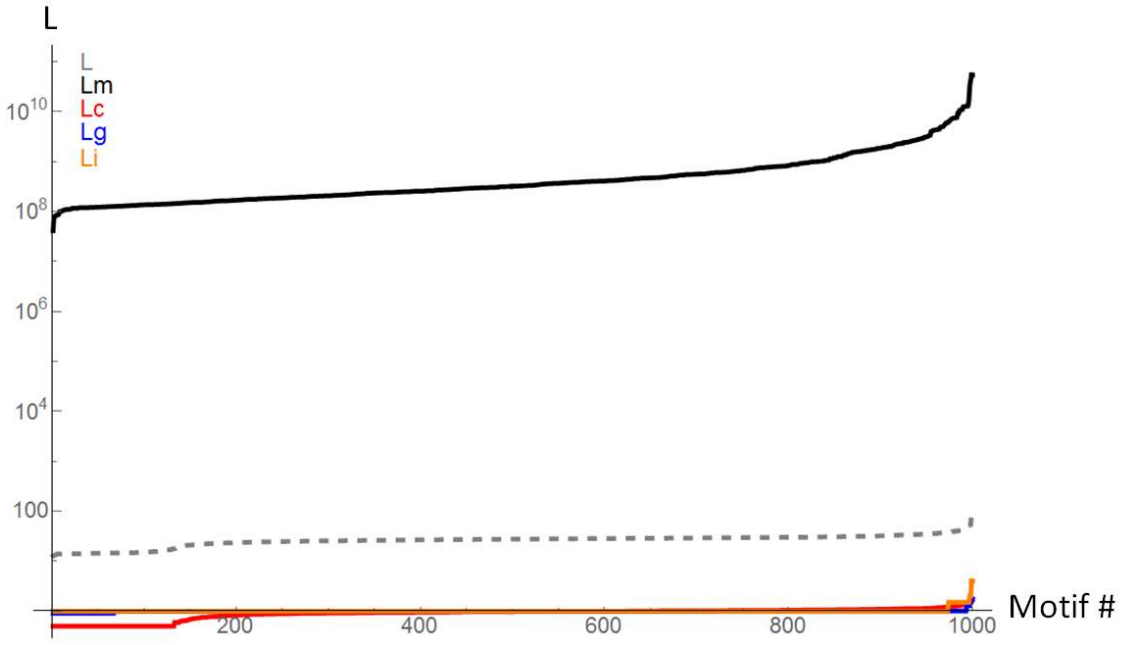


Figure 5. Contribution of various sources of data to the final likelihood for the 1 000 motif network using *Lm×Lc×Lg×Li*.

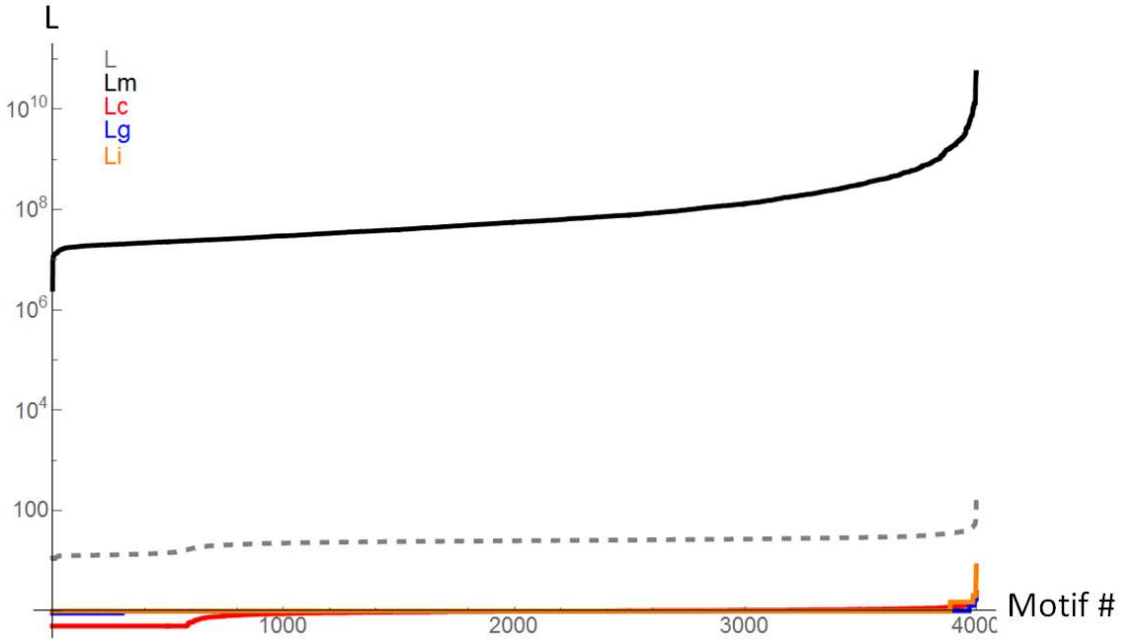


Figure 6. Contribution of various sources of data to the final likelihood for the 4 000 motif network using *Lm×Lc×Lg×Li*.

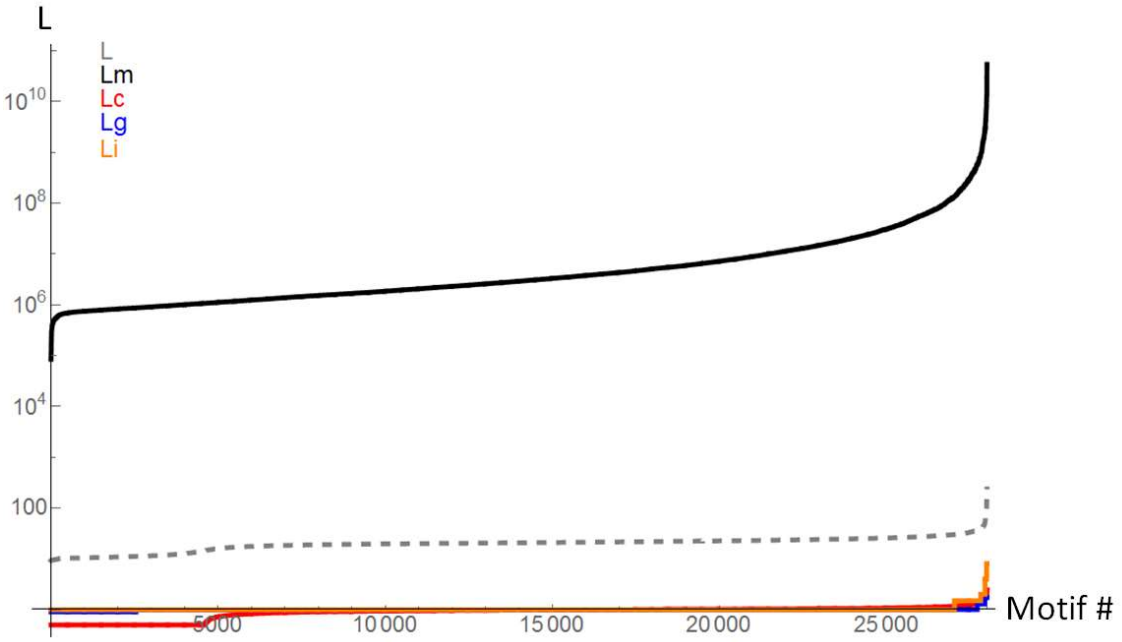


Figure 7. Contribution of various sources of data to the final likelihood for the 28 000 motif network using *Lm*×*Lc*×*Lg*×*Li*.

Table 2. Enrichment statistics for the top 10 TFs using a 4 000 motif network. Likelihoods were calculated as *Lm*×*Lc*×*Lg*×*Li*. Only long motifs were found as *Lm* was dominating the scoring scheme.

name1	name2	Length[Targets1]	Z(0)	nNo	nDown	nUp	length
9547	GCN4	18	4.27	11	4	3	21
9522	BAS1	14	2.21	10	2	2	21
9659	YAP1	11	2.10	7	0	4	20
9644	SUT2	19	1.58	16	0	3	20
9603	RAP1	1	1.43	0	0	1	10
9662	YAP6	17	1.40	13	0	4	20
9651	THI2	3	1.22	2	1	0	15
9656	UME6	9	1.08	6	1	2	13
9587	NDT80	31	0.88	25	3	3	21
9591	NRG1	16	0.78	14	1	1	20

The motif match likelihood *Lm* was clearly dominating the scoring system. An attempt should be made to suppress this effect and to also allow smaller motifs. Firstly, the effect of removing *Lm* completely was investigated. Motifs that were included all had *Lm* values above 10, but the likelihood function did not contain *Lm*, as shown below.

$$L = Lc \times Lg \times Li$$

Figure 8 shows the distribution of enrichment scores when *Lm* was removed, resulting in the inclusion of all TFs for the largest networks. An absence of a strong optimum enrichment was also evident.

Notably, even with this markedly different scoring system in which motif match likelihood played no role, the enrichment statistics again identified Gcn4 (general control protein 4) as significantly enriched with 11 out of 46 targets differentially regulated, mostly downward (Table 3). In addition, Gcr2 (glycolysis regulator 2) was the most significantly enriched TF, above Gcn4, and with a very significant enrichment score of 4.47 and 10 out of 22 targets regulated, almost exclusively downward. Other strongly enriched TFs (Rtg3, Swi4) and more weakly enriched TFs (Aro80, Hap1, Tye7, Spr2, Sok2, Opi1, Cup9, Rds1, Rds2, Ste12, Cin5) were also found. The motifs for Gcr2 may be inspected in Table 4 to demonstrate the method of calculation of final likelihood *L* as *Lc*×*Lg*×*Li*.

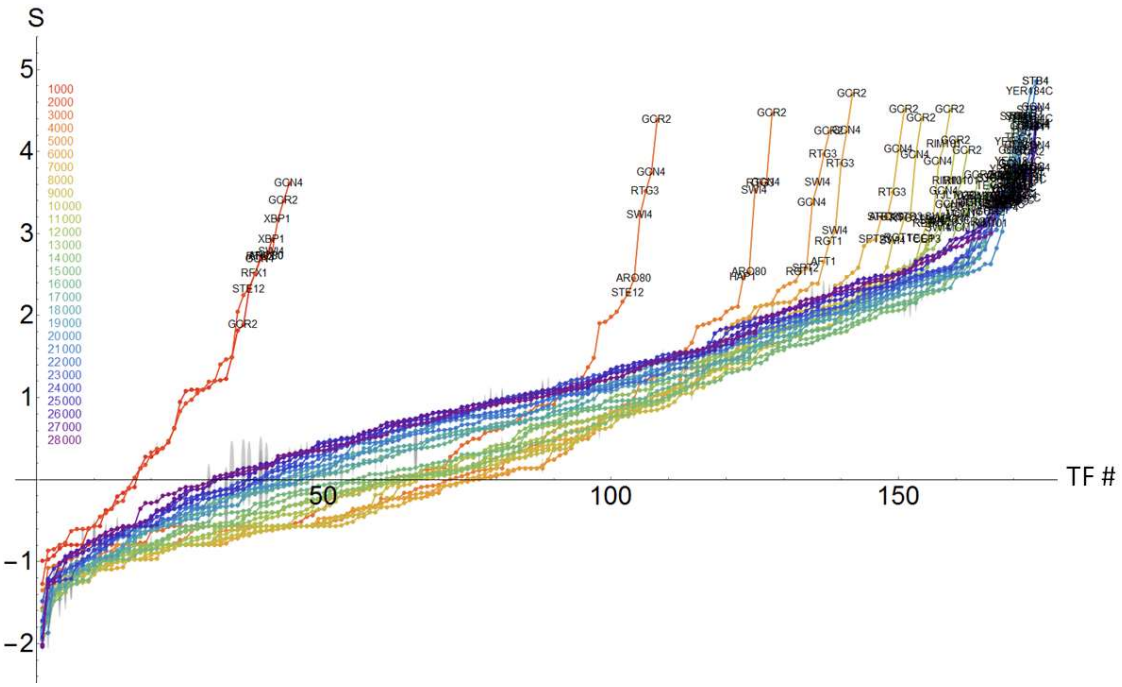


Figure 8. Distribution of enrichment score per TF for each of 28 gene regulatory networks using *Lc*×*Lg*×*Li*. A height of 1 of the grey disks equals 1 000 targets. No distinct enrichment optimum was present.

Table 3. Enrichment statistics for TFs using a 4 000 motif network. All motifs had *Lm* values of at least 10 but likelihood rank ratios *L* were calculated as *Lc*×*Lg*×*Li*. Both long and short motifs were found among enriched TFs. The numbers *n Up* and *n Down* refer to the number of target genes up or down-regulated from RNA-seq data.

motifID	TF	n Targets	Z(0)	n No	n Down	n Up	Motif length
9549	GCR2	22	4.47	12	9	1	7
9547	GCN4	46	3.62	35	8	3	21
9620	RTG3	23	3.61	14	8	1	20
9645	SWI4	21	3.53	13	1	7	8
9517	ARO80	7	2.54	4	0	3	21
9556	HAP1	31	2.47	22	6	3	8
9653	TYE7	19	2.11	14	3	2	7

motifID	TF	n Targets	Z(0)	n No	n Down	n Up	Motif length
9631	SPT2	14	2.10	11	2	1	10
9629	SOK2	30	2.05	23	3	4	11
9593	OPI1	4	1.96	2	1	1	7
9532	CUP9	2	1.95	1	1	0	9
9605	RDS1	1	1.91	0	1	0	7
9606	RDS2	4	1.89	3	1	0	7
9637	STE12	58	1.87	46	1	11	7
9528	CIN5	51	1.67	42	6	3	10

Table 4. Targets of TF Gcr2. All motifs had *Lm* values of at least 10 but likelihood rank ratios *L* were calculated as *Lc*×*Lg*×*Li*. Multiple occurrences of the same motif may occur in the same upstream regulatory region of a gene. During construction of a network, only one interaction was created for each TF target gene pair. Even though *Lm* was not included in the final calculation of likelihood *L*, it is included in the table for completeness.

TF	Target	Target name	test_id	log2FC	q	Signi-ficant.	Lm	Lc	Lg	Li	L
GCR2	GCR1	Glycolytic genes transcriptional activator GCR1	g1639.t1	-2.21	0.001	yes	196.3	1.09	1.74	2.00	3.79
GCR2	GCR1	Glycolytic genes transcriptional activator GCR1	g1639.t1	-2.21	0.001	yes	224.7	0.95	1.74	2.00	3.31
GCR2	GCR1	Glycolytic genes transcriptional activator GCR1	g1639.t1	-2.21	0.001	yes	389.8	0.95	1.74	2.00	3.31
GCR2	RAG2	Glucose-6-phosphate isomerase (GPI) (EC 5.3.1.9)	g1642.t1	-1.24	0.034	yes	158.3	1.17	1.00	2.00	2.35
GCR2	RAG2	Glucose-6-phosphate isomerase (GPI) (EC 5.3.1.9)	g1642.t1	-1.24	0.034	yes	68.9	1.11	1.00	2.00	2.22
GCR2	RAG2	Glucose-6-phosphate isomerase (GPI) (EC 5.3.1.9)	g1642.t1	-1.24	0.034	yes	956.9	1.06	1.00	2.00	2.13
GCR2	CAF40	Protein CAF40 (40 kDa CCR4-associated factor)	g1752.t1	0.94	0.159	no	48.7	1.09	1.01	4.00	4.41
GCR2	PDC1	Pyruvate decarboxylase (EC 4.1.1.1)	g1785.t1	-2.50	0.005	yes	221.4	1.60	1.00	4.00	6.42
GCR2	PDC1	Pyruvate decarboxylase (EC 4.1.1.1)	g1785.t1	-2.50	0.005	yes	183.0	1.16	1.00	4.00	4.63
GCR2	PHO11	Acid phosphatase PHO11 (EC 3.1.3.2) (P56)	g2192.t1	-0.61	0.521	no	62.5	1.05	1.00	2.00	2.11
GCR2	RFC5	Replication factor C subunit 5 (Replication factor C5)	g2595.t1	-0.44	0.635	no	206.2	1.00	0.90	4.00	3.59
GCR2	RFC5	Replication factor C subunit 5 (Replication factor C5)	g2595.t1	-0.44	0.635	no	141.1	0.85	0.90	4.00	3.06
GCR2	IST2	Increased sodium tolerance protein 2	g2598.t1	-0.38	0.617	no	252.5	1.05	1.00	4.00	4.22
GCR2	IST2	Increased sodium tolerance protein 2	g2598.t1	-0.38	0.617	no	196.4	1.04	1.00	4.00	4.19

TF	Target	Target name	test_id	log2FC	q	Signi-ficant.	Lm	Lc	Lg	Li	L
GCR2	IST2	Increased sodium tolerance protein 2	g2598.t1	-0.38	0.617	no	8007.8	1.02	1.00	4.00	4.08
GCR2	IST2	Increased sodium tolerance protein 2	g2598.t1	-0.38	0.617	no	157.2	0.90	1.00	4.00	3.61
GCR2	IST2	Increased sodium tolerance protein 2	g2598.t1	-0.38	0.617	no	2462.7	0.87	1.00	4.00	3.48
GCR2	IST2	Increased sodium tolerance protein 2	g2598.t1	-0.38	0.617	no	8018.4	0.81	1.00	4.00	3.25
GCR2	YCP4	Flavoprotein-like protein YCP4	g2686.t1	1.25	0.020	yes	26.1	1.05	1.00	2.00	2.11
GCR2	PHM7	Phosphate metabolism protein 7	g2979.t1	-0.76	0.258	no	16.3	0.92	1.00	4.00	3.69
GCR2	SUN4	Probable secreted beta-glucosidase SUN4 (EC 3.2.1.-)	g3018.t1	0.71	0.262	no	882.1	1.18	1.00	4.00	4.72
GCR2	SUN4	Probable secreted beta-glucosidase SUN4 (EC 3.2.1.-)	g3018.t1	0.71	0.262	no	103.6	1.00	1.00	4.00	4.01
GCR2	SUN4	Probable secreted beta-glucosidase SUN4 (EC 3.2.1.-)	g3018.t1	0.71	0.262	no	2289.4	0.81	1.00	4.00	3.24
GCR2	SUN4	Probable secreted beta-glucosidase SUN4 (EC 3.2.1.-)	g3018.t1	0.71	0.262	no	24994.6	0.71	1.00	4.00	2.86
GCR2	PFK1	ATP-dependent 6-phosphofructokinase subunit alpha (ATP-PFK 1)	g3288.t1	-2.27	0.001	yes	75.9	1.00	1.00	4.00	4.01
GCR2	PFK1	ATP-dependent 6-phosphofructokinase subunit alpha (ATP-PFK 1)	g3288.t1	-2.27	0.001	yes	15410.4	0.94	1.00	4.00	3.79
GCR2	CMK2	Calcium/calmodulin-dependent protein kinase II (EC 2.7.11.17)	g336.t1	0.09	0.937	no	659.6	1.18	1.00	2.00	2.36
GCR2	CMK2	Calcium/calmodulin-dependent protein kinase II (EC 2.7.11.17)	g336.t1	0.09	0.937	no	41.8	1.06	1.00	2.00	2.13
GCR2	TCM62	Mitochondrial chaperone TCM62	g3624.t1	-0.18	0.884	no	24.2	2.19	1.00	1.00	2.19
GCR2	GPM1	Phosphoglycerate mutase 1 (PGAM 1) (EC 5.4.2.11)	g367.t1	-2.35	0.001	yes	26.9	1.43	1.00	2.00	2.86
GCR2	CDC6	Cell division control protein 6	g3870.t1	0.84	0.273	no	68.6	1.00	1.00	4.00	4.01
GCR2	SKS1	Serine/threonine-protein kinase SKS1 (EC 2.7.11.1) (Suppressor kinase of SNF3)	g3.t1	-2.21	0.001	yes	203.5	1.05	1.01	2.00	2.13
GCR2	RPA34	DNA-directed RNA polymerase I subunit RPA34 (A34)	g413.t1	-1.86	0.074	no	1698.8	1.10	0.90	4.00	3.96
GCR2	RPA34	DNA-directed RNA polymerase I subunit RPA34 (A34)	g413.t1	-1.86	0.074	no	74.6	1.09	0.90	4.00	3.92
GCR2	RPA34	DNA-directed RNA polymerase I subunit RPA34 (A34)	g413.t1	-1.86	0.074	no	256.4	1.07	0.90	4.00	3.87
GCR2	RPA34	DNA-directed RNA polymerase I subunit RPA34 (A34)	g413.t1	-1.86	0.074	no	122.3	1.06	0.90	4.00	3.83

TF	Target	Target name	test_id	log2FC	q	Signi-ficant.	Lm	Lc	Lg	Li	L
GCR2	ERG25	Methylsterol monooxygenase (EC 1.14.13.72)	g4447.t1	-1.20	0.032	yes	2337.5	1.13	1.00	4.00	4.54
GCR2	ERG25	Methylsterol monooxygenase (EC 1.14.13.72)	g4447.t1	-1.20	0.032	yes	56.2	1.09	1.00	4.00	4.35
GCR2	ERG25	Methylsterol monooxygenase (EC 1.14.13.72)	g4447.t1	-1.20	0.032	yes	94.0	1.05	1.00	4.00	4.22
GCR2	ERG25	Methylsterol monooxygenase (EC 1.14.13.72)	g4447.t1	-1.20	0.032	yes	817.2	1.00	1.00	4.00	4.01
GCR2	HOP2	Homologous-pairing protein 2	g4625.t1	0.00	1.000	no	121.3	1.10	1.00	2.00	2.21
GCR2	ENO	Enolase (EC 4.2.1.11) (2-phospho-D-glycerate hydro-lyase)	g4751.t1	-1.74	0.039	yes	2141.6	1.06	1.00	2.00	2.13
GCR2	ENO	Enolase (EC 4.2.1.11) (2-phospho-D-glycerate hydro-lyase)	g4751.t1	-1.74	0.039	yes	173.3	1.05	1.00	2.00	2.11
GCR2	PHO89	Phosphate permease PHO89 (Na(+)/Pi cotransporter PHO89)	g719.t1	1.17	0.091	no	39.7	1.10	1.00	2.00	2.21
GCR2	PHO89	Phosphate permease PHO89 (Na(+)/Pi cotransporter PHO89)	g719.t1	1.17	0.091	no	10.5	1.07	1.00	2.00	2.15
GCR2	MET3	Sulfate adenylyltransferase (EC 2.7.7.4) (ATP-sulfurylase)	g887.t1	-1.31	0.016	yes	35.2	0.75	1.00	4.00	3.01

Improving the balance in the likelihood scoring system

The distribution of final likelihoods based on the likelihood function $Lm \times Lc \times Lg \times Li$ is shown in Figure 9. It is evident that the motif likelihood ratio *Lm* was dominant in the scoring, as the final distribution adopted the shape of the distribution of *Lm*. An improved scoring system may be to effectively weight the scoring scheme to include *Lm* in some suppressed form. Since *Lm* increases exponentially with the length of a motif, the range of *Lm* was extremely large. Therefore, a good strategy might be to include *Lm* as a Log values like $\text{Log}_2(Lm)$. Figure 10 shows a more equal contribution by all sources of evidence to the final likelihood by using the latter in the function. Figure 11 shows that all TFs were included for the larger networks, a trait that shows allowance for short motifs, which required support from additional sources of evidence. Again, some optimum in TF enrichment was seen for a network size around 16 000 motifs. The resulting distributions and networks were identical when $\text{Log}_{10}(Lm)$ was used, and subsequently, only $\text{Log}_{10}(Lm)$ was used as the output is easier to understand.

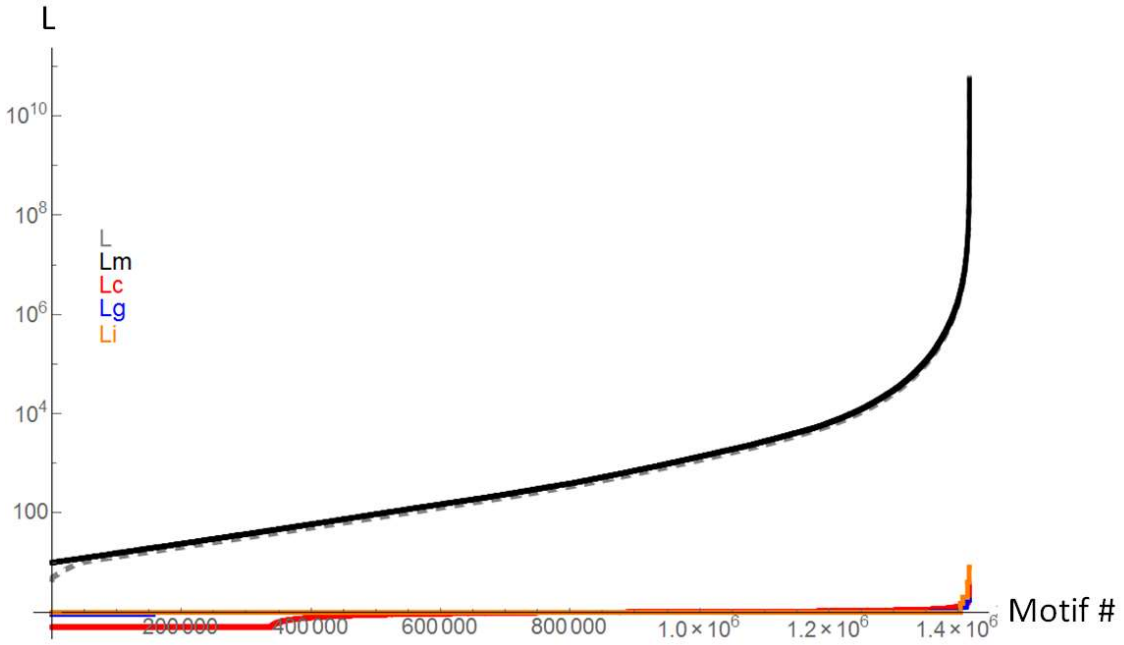


Figure 9. Distribution of the final likelihoods *L* based on the likelihood function *L* = *Lm*×*Lc*×*Lg*×*Li*. The final likelihood *L* adopted the shape of *Lm*.

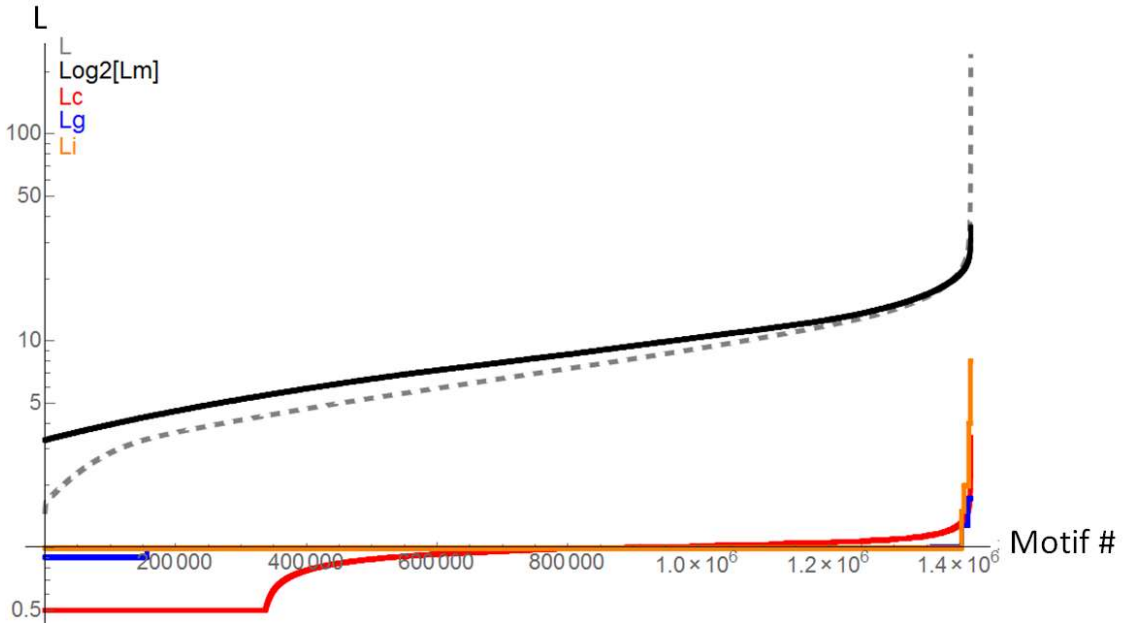


Figure 10. Distribution of the final likelihoods *L* based on the likelihood function *L* = $\mathbf{Log_2}$(*Lm*)×*Lc*×*Lg*×*Li*. The final likelihood *L* adopted the combined shapes of all likelihood distributions for the various sources of evidence.

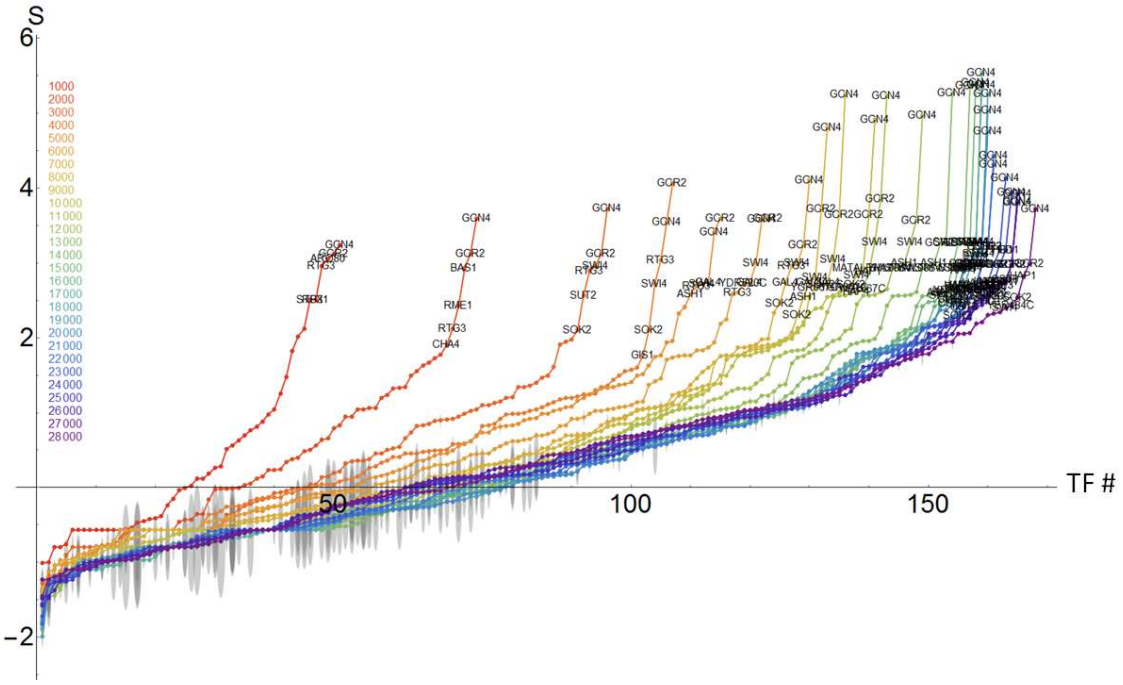


Figure 11. Distribution of enrichment score per TF for each of 28 gene regulatory networks using $Log_{10}(Lm)\times Lc\times Lg\times Li$. A height of 1 in grey disks equals 1 000 targets.

Using a network with 4 000 motifs again revealed Gcr2, Gcn4, Rtg3, Swi4, Sok2 and Gis1 as enriched, while the enrichment scores were somewhat lower than with only *Lc*×*Lg*×*Li*. In a 20 000 motif network (Table 5), strongly enriched TFs again were Gcn4, Gcr2, Stb4 and Swi4. Note that shorter motifs were included as opposed to the case when *Lm* was dominating the likelihood function. Since Stb4, Phd1 and Tfb1 only had 2 targets each, these should be ignored due to the small sample sizes, although the enrichment statistic does incorporate the standard deviation of the background enrichment.

Table 5. Enrichment statistics for a 20 000 motif network using the likelihood function $Li = Log_{10}(Lm)\times Lc\times Lg\times Li$. TFs with scores above 1.67 ($p < 0.05$) are shown.

name1	name2	Length[Targets1]	Z(0)	nNo	nDown	nUp	motiflength
9547	GCN4	168	4.761989	132	22	14	21
9549	GCR2	18	3.22025	11	6	1	7
9635	STB4	2	3.120459	1	1	0	7
9645	SWI4	25	2.97451	17	3	5	8
9556	HAP1	35	2.903399	25	6	4	8
9666	YDR520C	13	2.635549	8	4	1	10
9570	MAC1	11	2.434674	7	0	4	8
9659	YAP1	177	2.371953	145	13	19	20
9530	CST6	54	2.369175	42	8	4	9
9669	YGR067C	30	2.237415	23	3	4	14

name1	name2	Length[Targets1]	Z(0)	nNo	nDown	nUp	motiflength
9629	SOK2	30	2.175238	23	3	4	11
9520	ASH1	11	1.928312	7	0	4	10
9614	RME1	31	1.926014	24	3	4	10
9522	BAS1	60	1.923404	50	5	5	21
9674	YLR278C	12	1.909336	9	1	2	8
9656	UME6	19	1.905283	14	1	4	13
9612	RIM101	9	1.895591	6	1	2	7
9638	STP1	23	1.872487	19	2	2	8
9528	CIN5	63	1.851159	54	5	4	10
9653	TYE7	69	1.844238	55	10	4	7
9599	PHD1	2	1.801656	1	0	1	10
9647	TBF1	2	1.801656	1	0	1	8
9654	UGA3	3	1.781612	2	1	0	8

Since each run produced a slightly different result, a formalised optimisation criterion may be best to decide on the best network. A good optimization criterion to finding the best number of motifs might be one that optimises not only for the maximal enrichment statistic, but also for maximal distinguishing power between strongly enriched TFs and the majority of non-enriched TFs (background). A simple criterion is to use the difference between the highest enrichment score *Sa* and the lowest enrichment statistic *Sz*, or to increase the robustness, using the mean value of the three lowest enrichment scores *Sx*, *Sy* and *Sz*.

$$\text{Maximise}(n) \qquad (Sa - \text{Mean}[Sx, Sy, Sz])$$

Several other methods were also tested for inclusion of *Lm,* including raising *Lm* to various powers (Figure 12). The highest scoring network among all methods and all values for *n* was obtained by allowing 16 000 motifs, using the $\text{Log}_{10}(Lm) \times Lc \times Lg \times Li$ method. However, a substantially smaller network could be generated with a similar score *S* using 9 000 motifs, which would likely be more accurate and which resulted in 5 443 interactions and 136 TFs. Using simply the highest score *S* as the criterion resulted in the same deduction (Figure 13). The best network resulted in again finding Gcn4 and Gcr2 as most significantly enriched (see Chapter 5, Table 3).

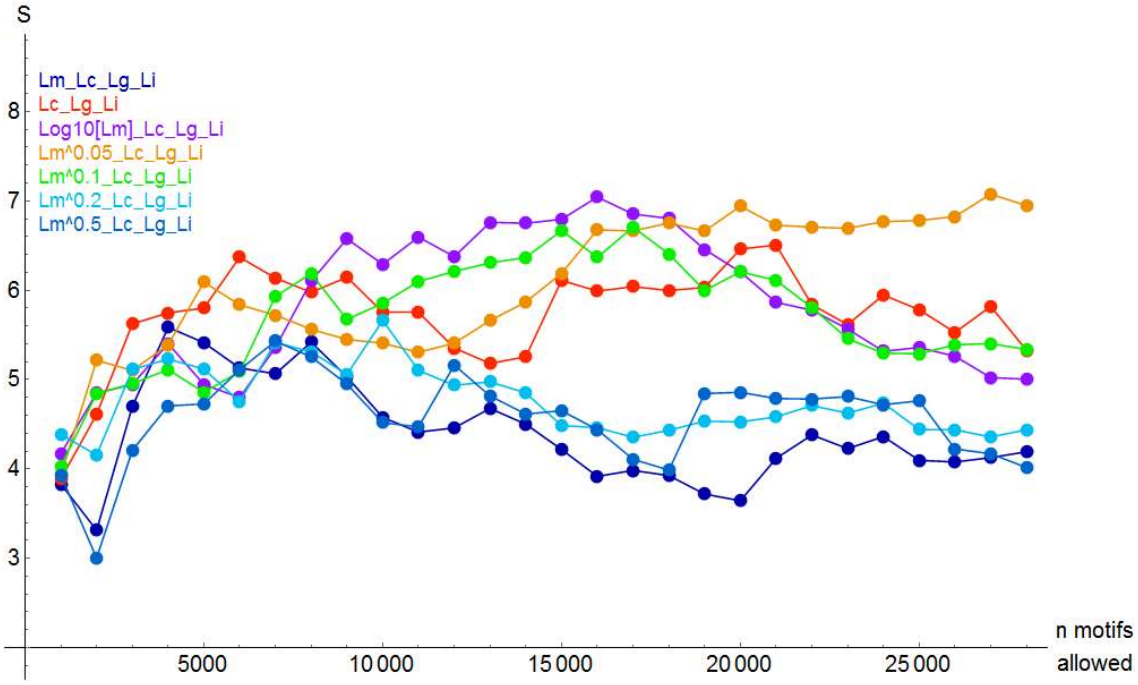


Figure 12. Optimisation of the network for the highest score minus the mean of the three lowest scores (*Sa* – Mean[*Sx*, *Sy*, *Sz*]).

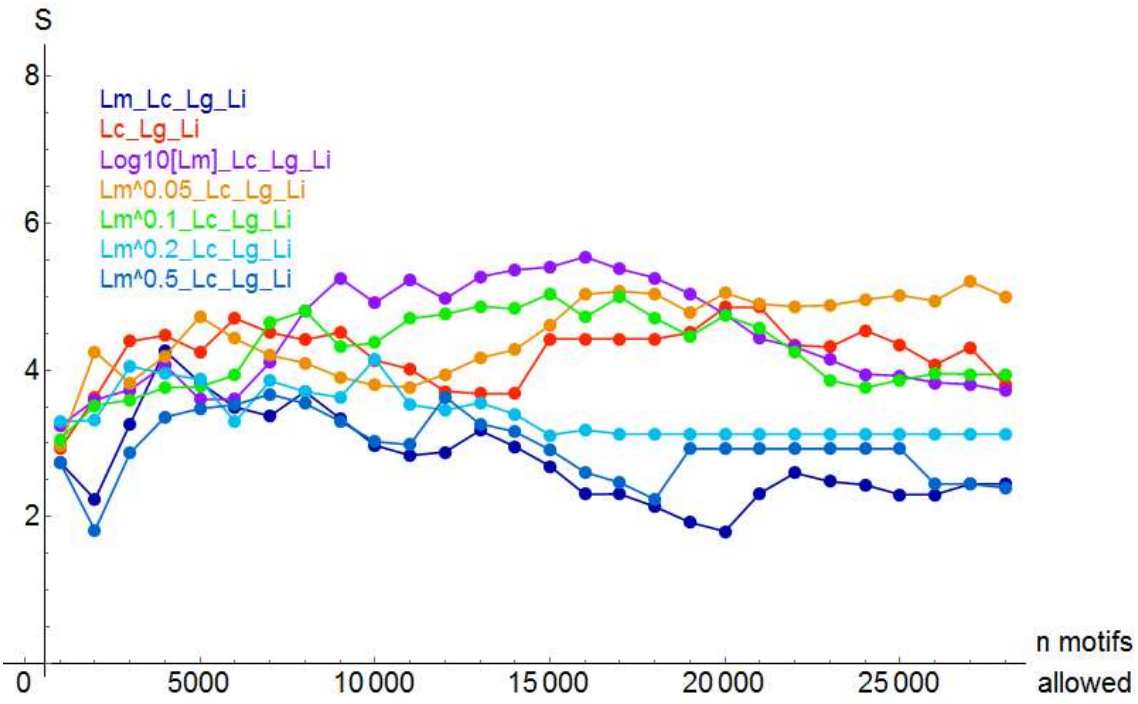


Figure 13. Optimisation of the network for the highest score *Sa*.

Addendum 4

Addendum for gene regulatory network based on the complete genome of *Kluyveromyces marxianus* - Chapter 6

This addendum describes further details of methods from Chapter 6.

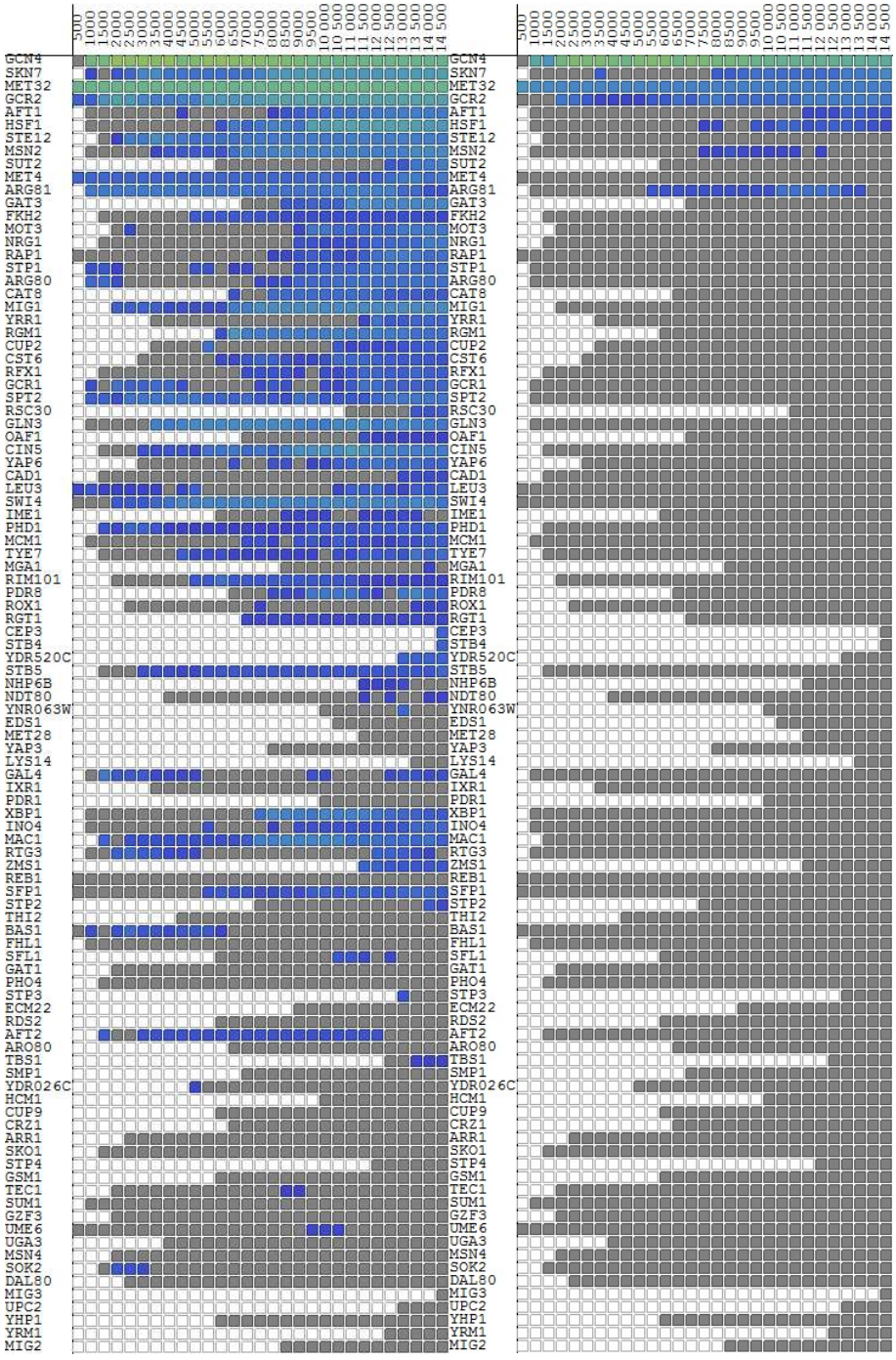


Figure 1. Enrichment statistics calculated for differential expression in gene sets using the Z-score method (left) or the hypergeometric distribution (right) for networks constructed using the function $Log_{10}(Lm)+Lc+Li$ for the final likelihood.

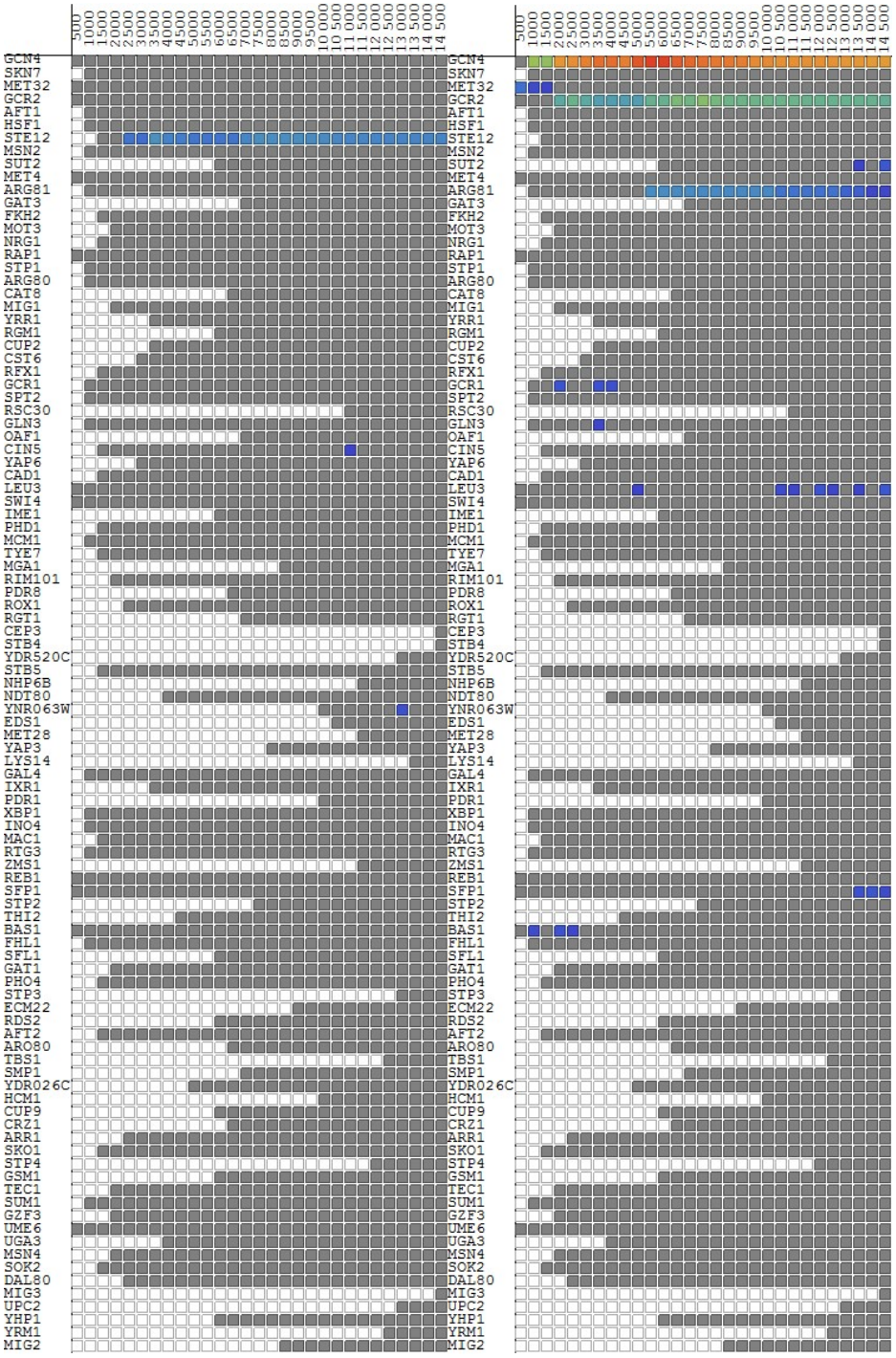


Figure 2. Enrichment statistics calculated for up-regulation (left) and down-regulation (right) using the hypergeometric distribution for networks constructed using the function $Log_{10}(Lm)+Lc+Li$ for the final likelihood.

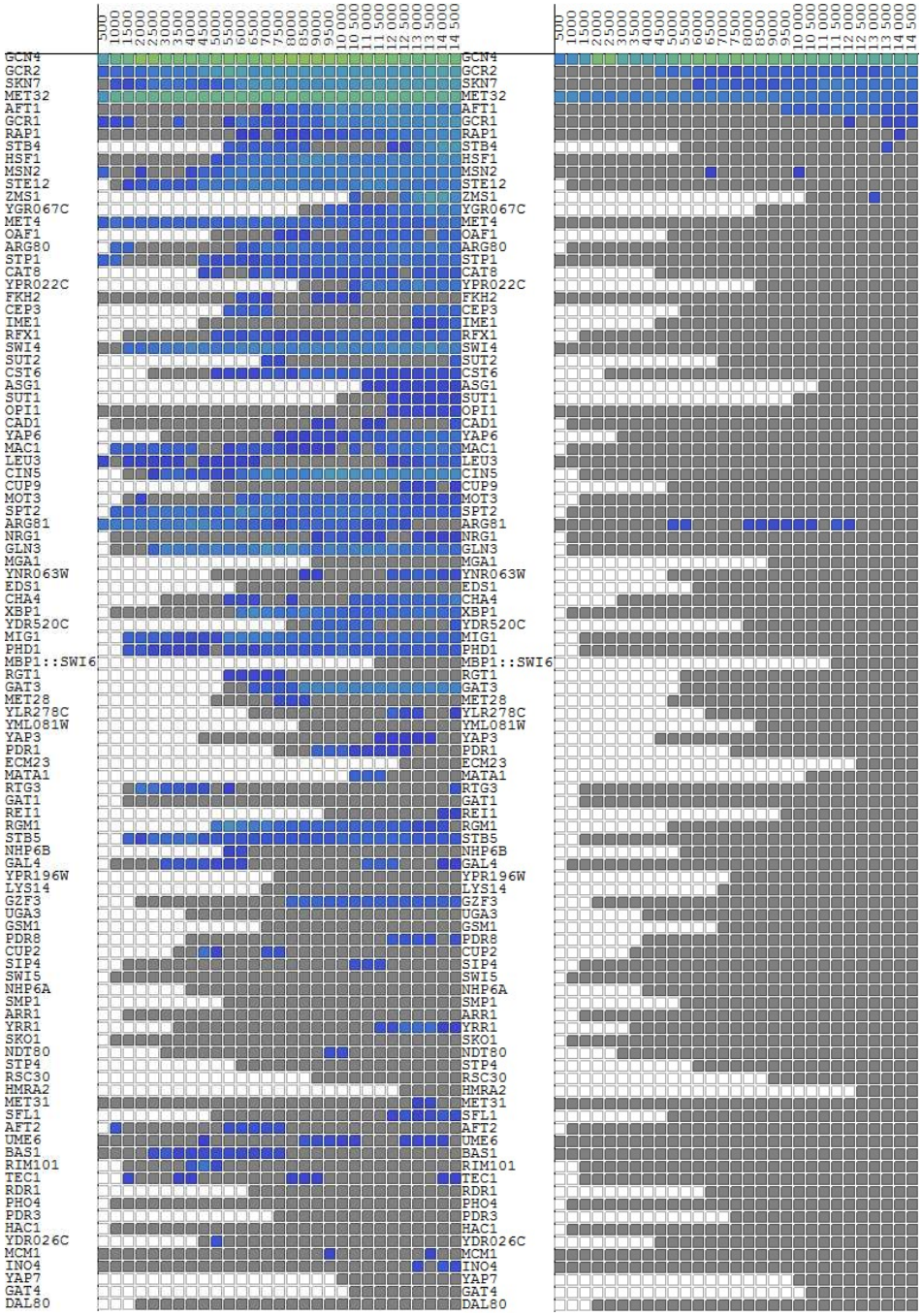


Figure 3. Enrichment statistics calculated for differential expression in gene sets using the Z-score method (left) or the hypergeometric distribution (right) for networks constructed using the function If[*Li*==0, $Log_{10}(Lm)\times Lc$, $Log_{10}(Lm)\times Lc\times Li$] for the final likelihood.

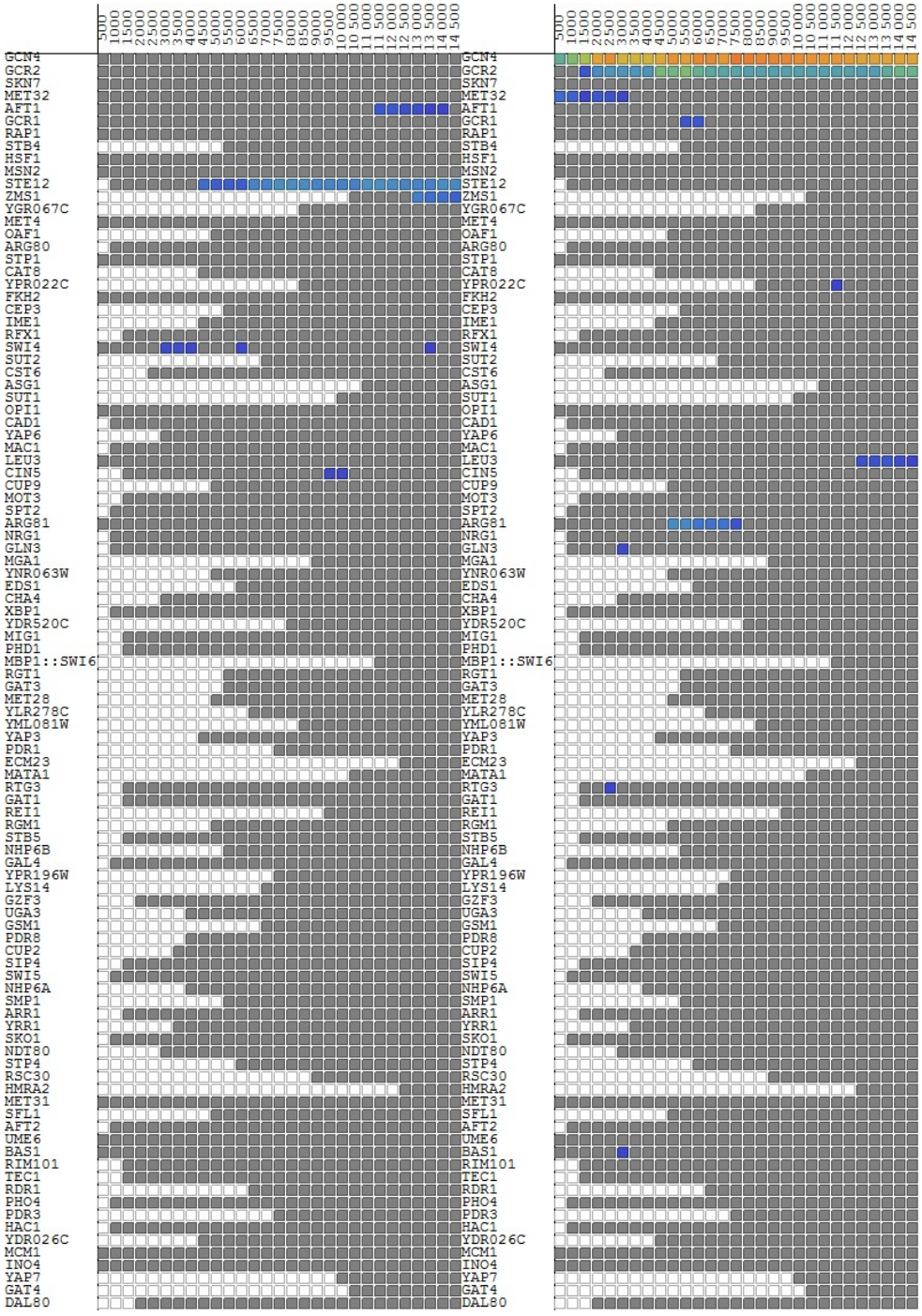


Figure 4. Enrichment statistics calculated for up-regulation (left) and down-regulation (right) using the hypergeometric distribution for networks constructed using the function If[*Li*==0, $Log_{10}(Lm)\times Lc$, $Log_{10}(Lm)\times Lc\times Li$] for the final likelihood.

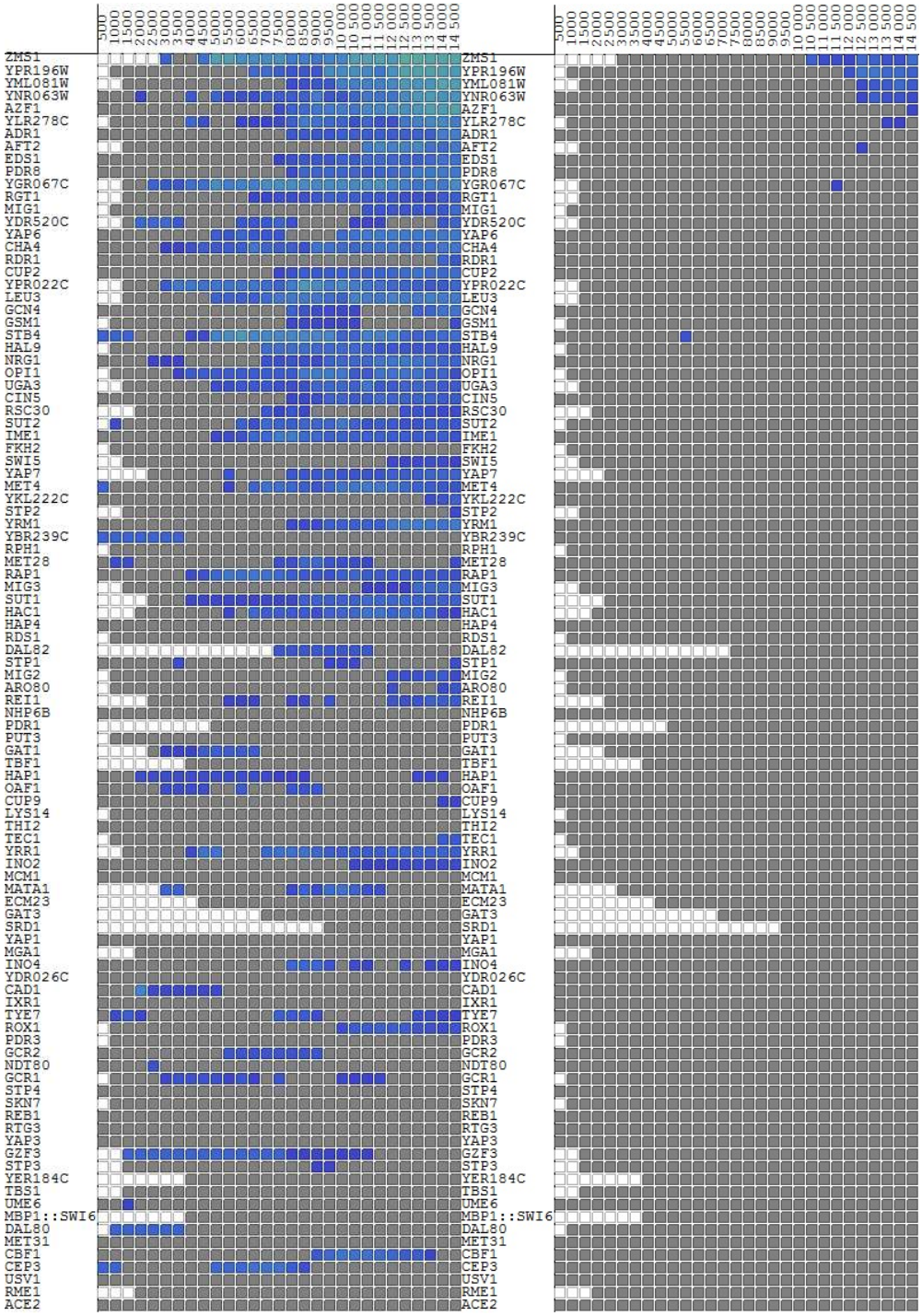


Figure 5. Enrichment statistics calculated for differential expression in gene sets using the Z-score method (left) or the hypergeometric distribution (right) for networks constructed using the function $Log_{10}(Lm)\times Lc$ for the final likelihood. Note that out of 174 TFs, Adr1 is seventh and Mig1 is nineteenth, while Gcn1 is fifteenth, ranked based on enrichment score for differential expression using the hypergeometric distribution.

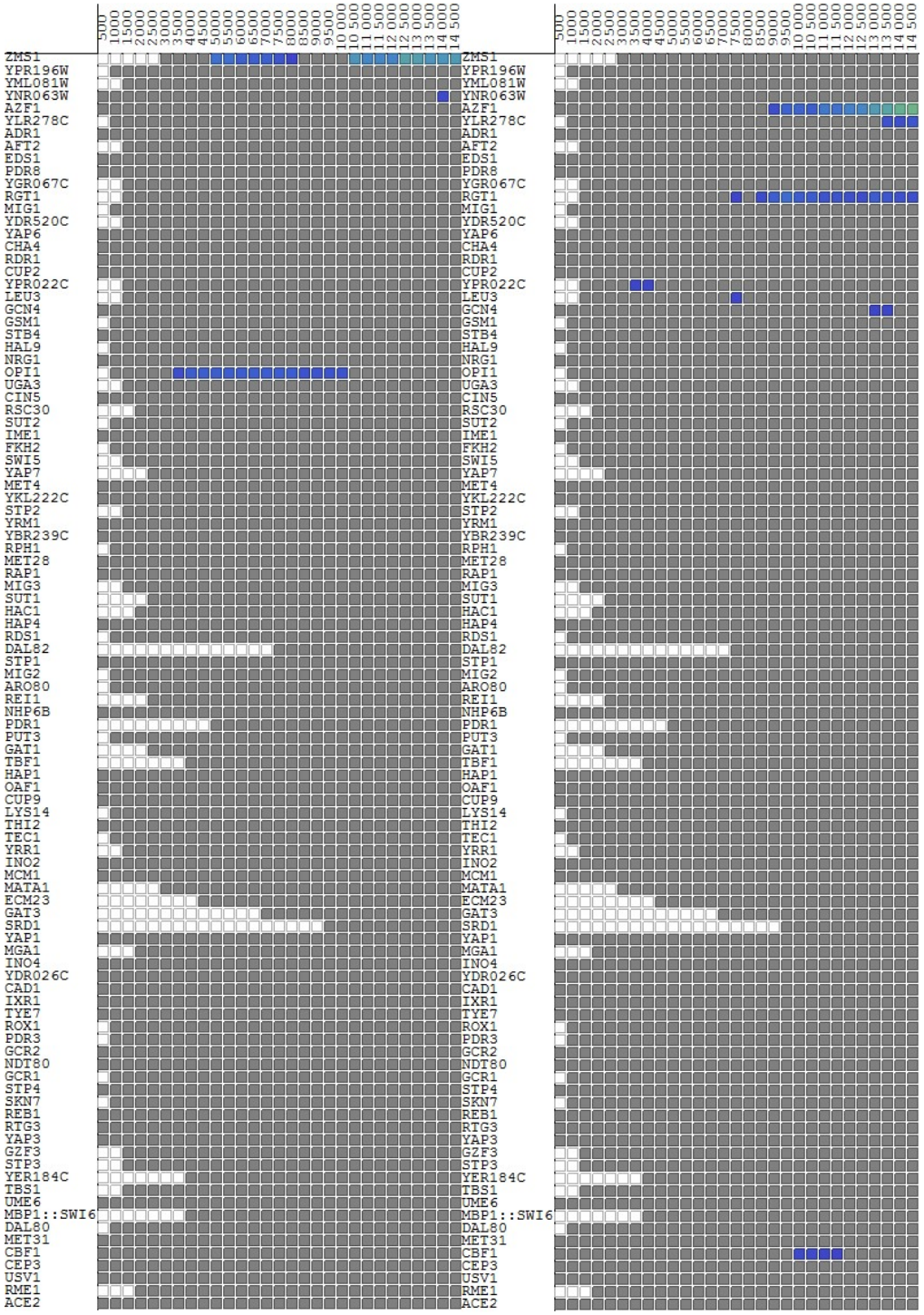


Figure 6. Enrichment statistics calculated for up-regulation (left) and down-regulation (right) using the hypergeometric distribution for networks constructed using the function $Log_{10}(Lm)\times Lc$ for the final likelihood.

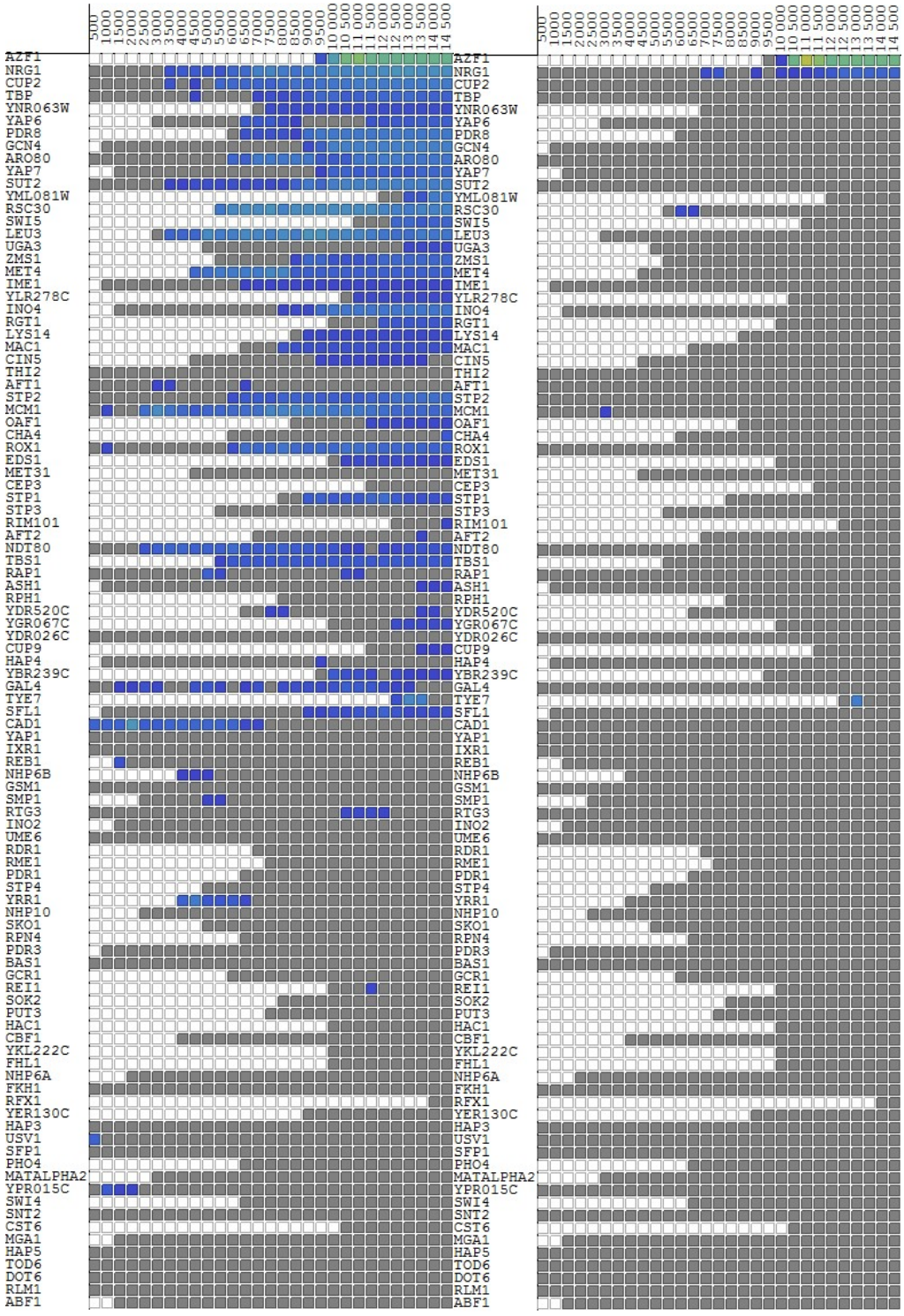


Figure 7. Enrichment statistics calculated for differential expression in gene sets using the Z-score method (left) or the hypergeometric distribution (right) for networks constructed using the function $Log_{10}(Lm)$ for the final likelihood.

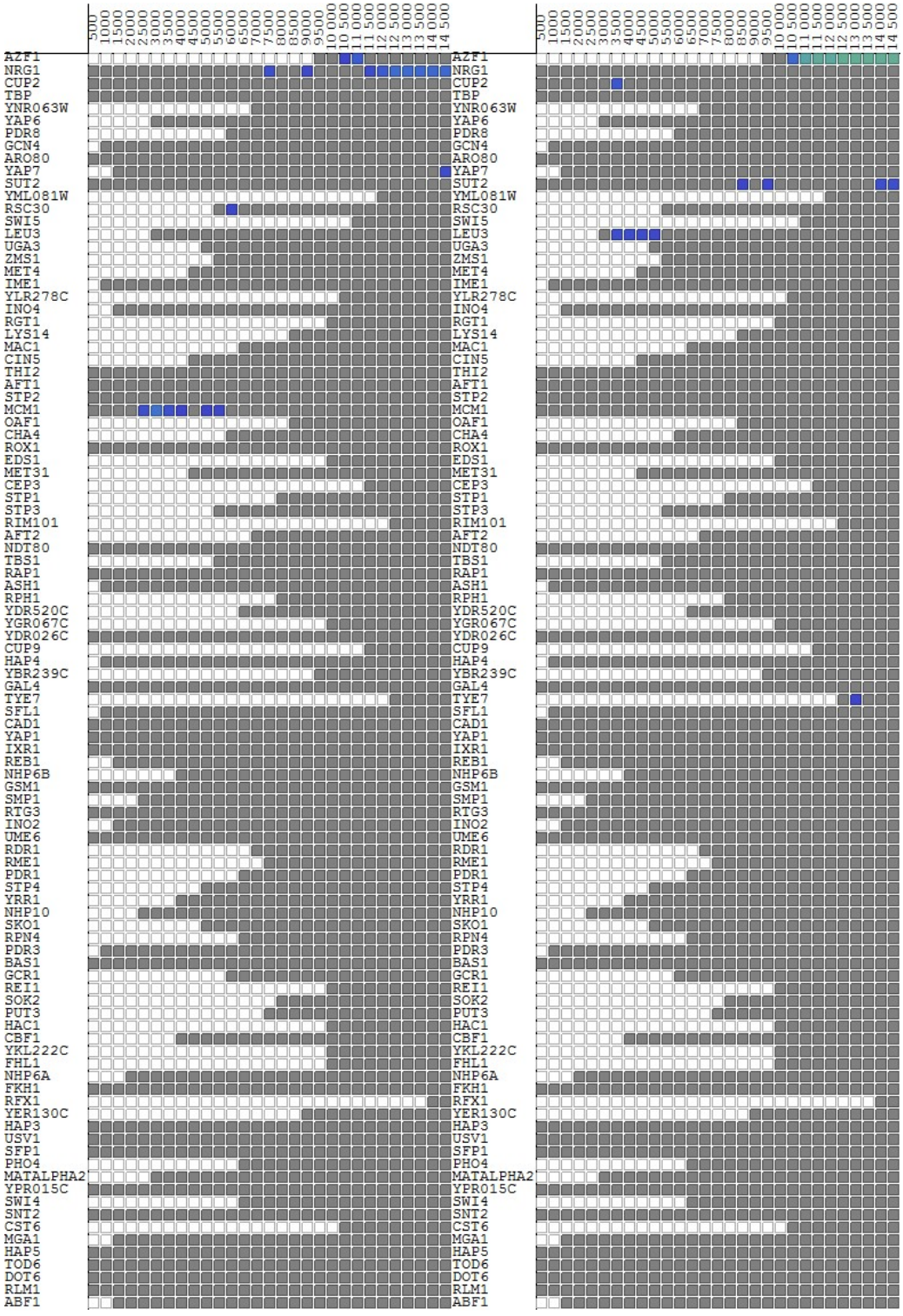


Figure 8. Enrichment statistics calculated for up-regulation (left) and down-regulation (right) using the hypergeometric distribution for networks constructed using the function $Log_{10}(Lm)$ for the final likelihood.

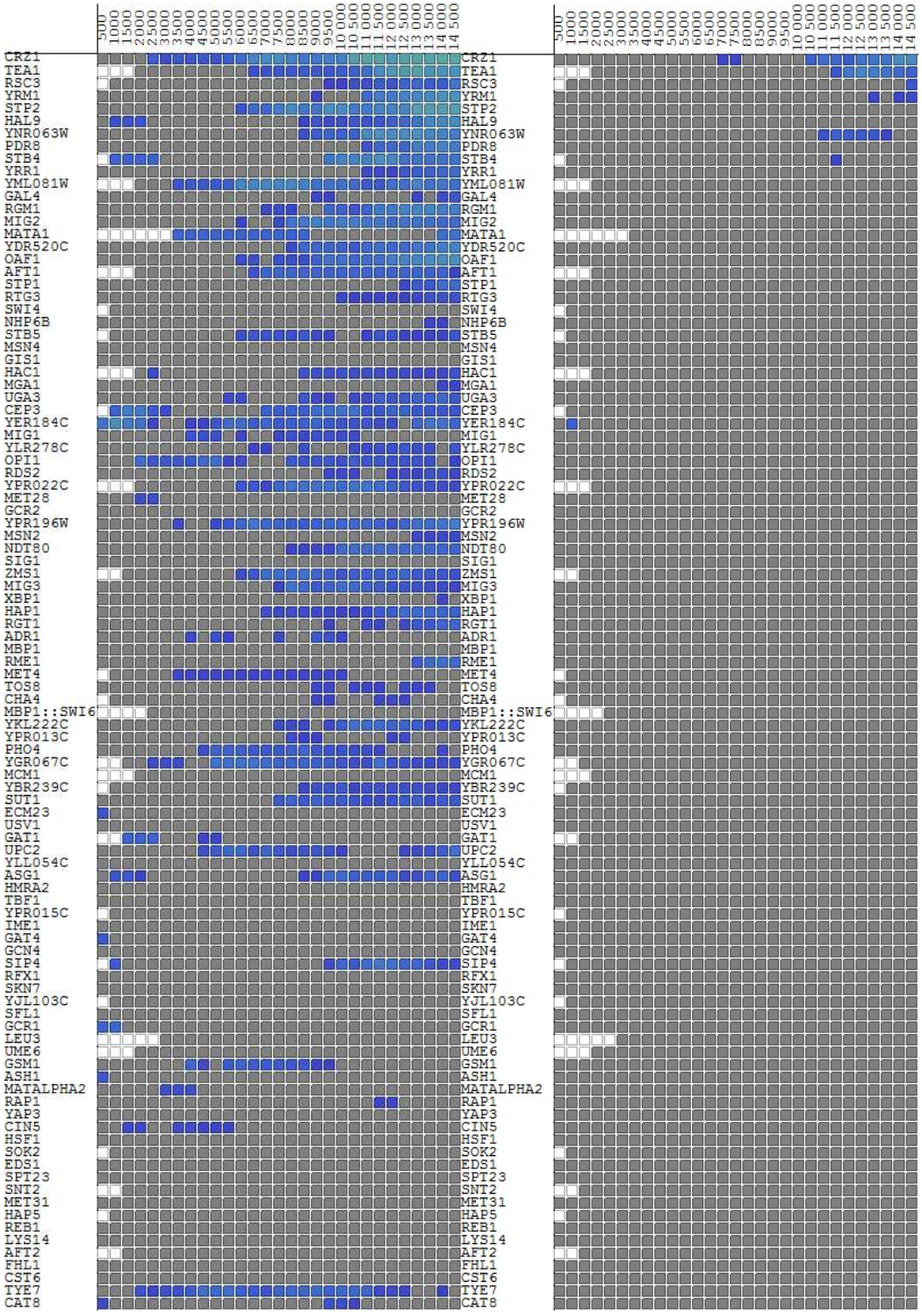


Figure 9. Enrichment statistics calculated for differential expression in gene sets using the Z-score method (left) or the hypergeometric distribution (right) for networks constructed using the function *Lc* for the final likelihood.

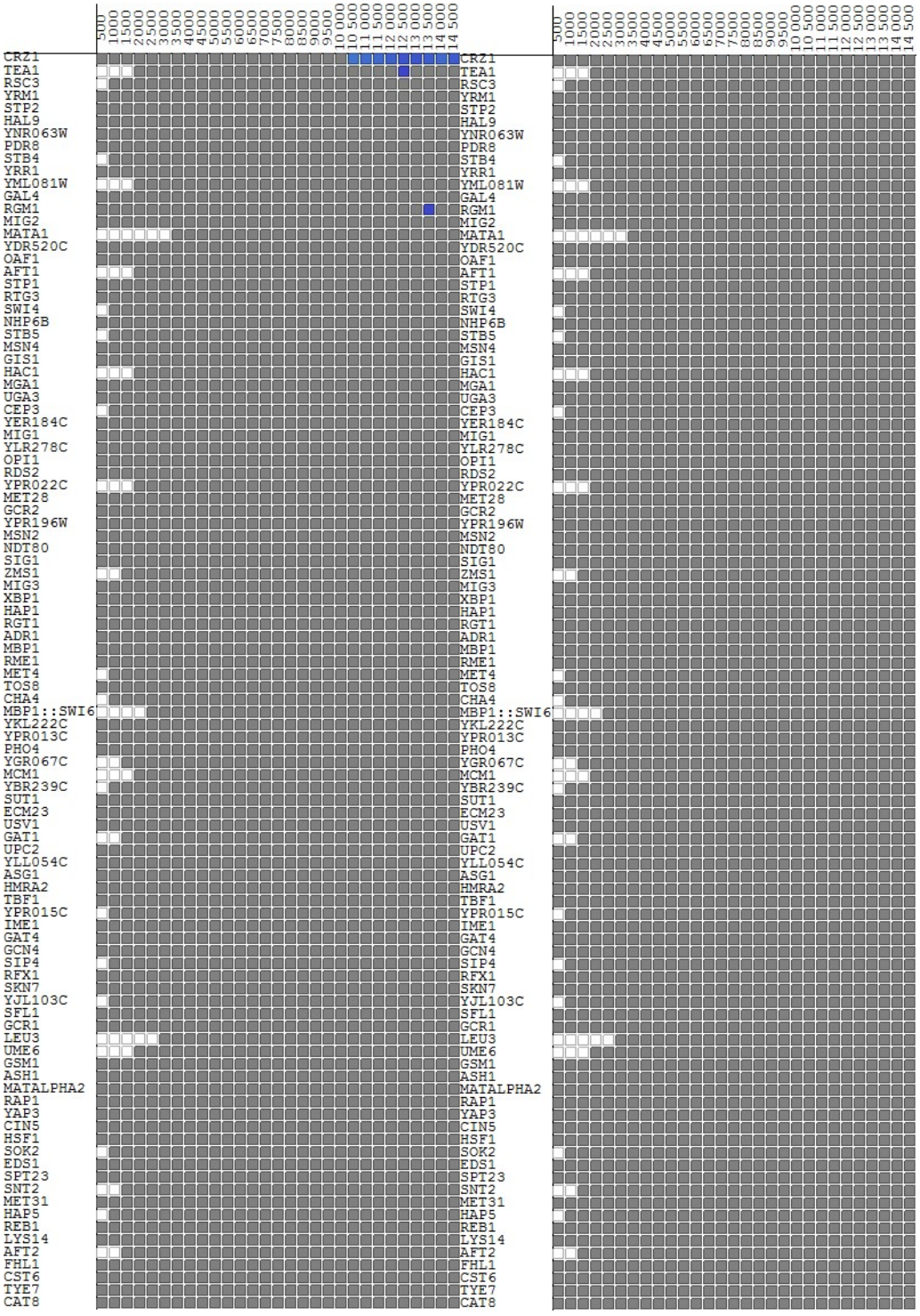


Figure 10. Enrichment statistics calculated for up-regulation (left) and down-regulation (right) using the hypergeometric distribution for networks constructed using the function *Lc* for the final likelihood.

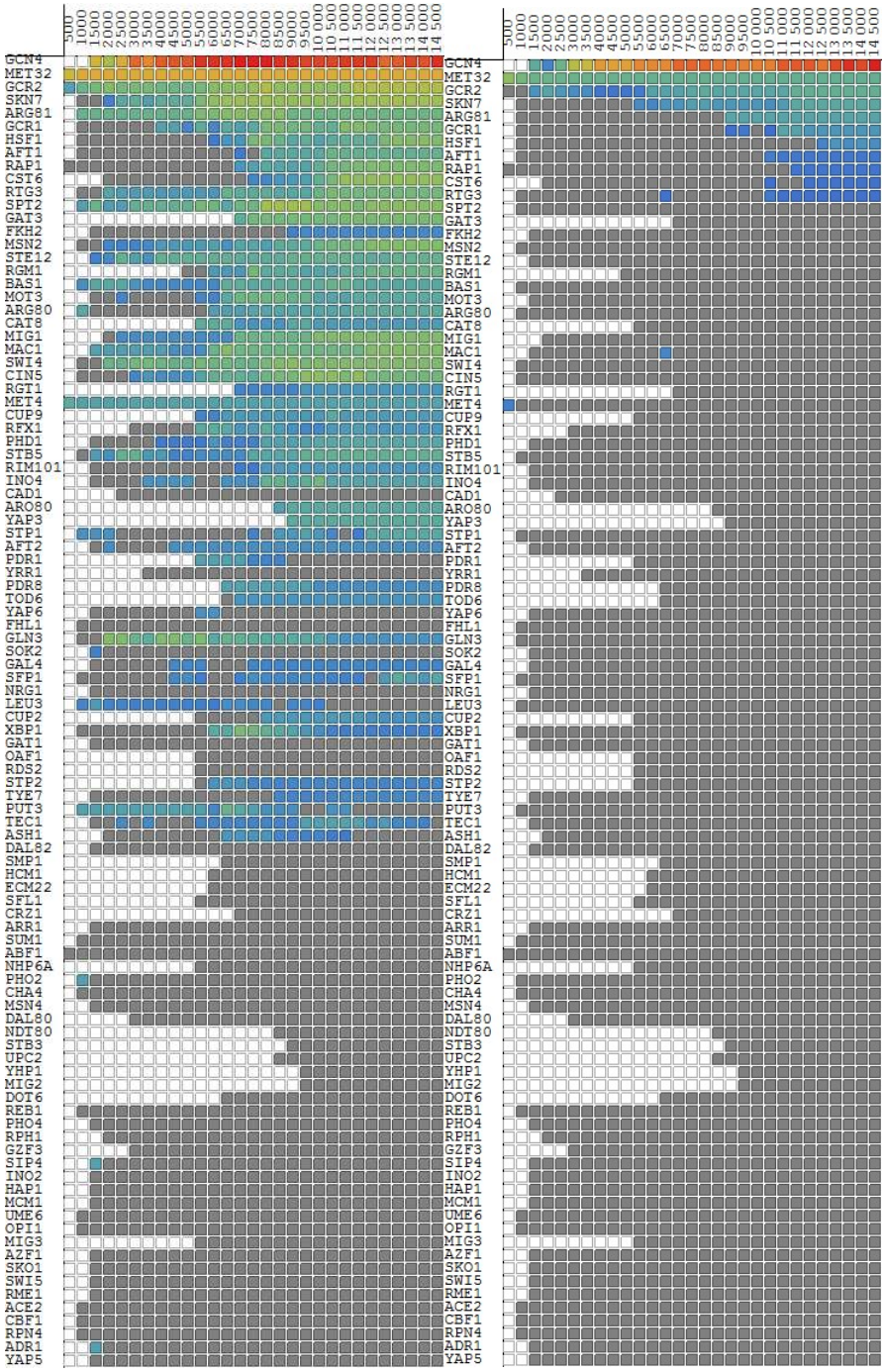


Figure 11. Enrichment statistics calculated for differential expression in gene sets using the Z-score method (left) or the hypergeometric distribution (right) for networks constructed using the function *Li* for the final likelihood.

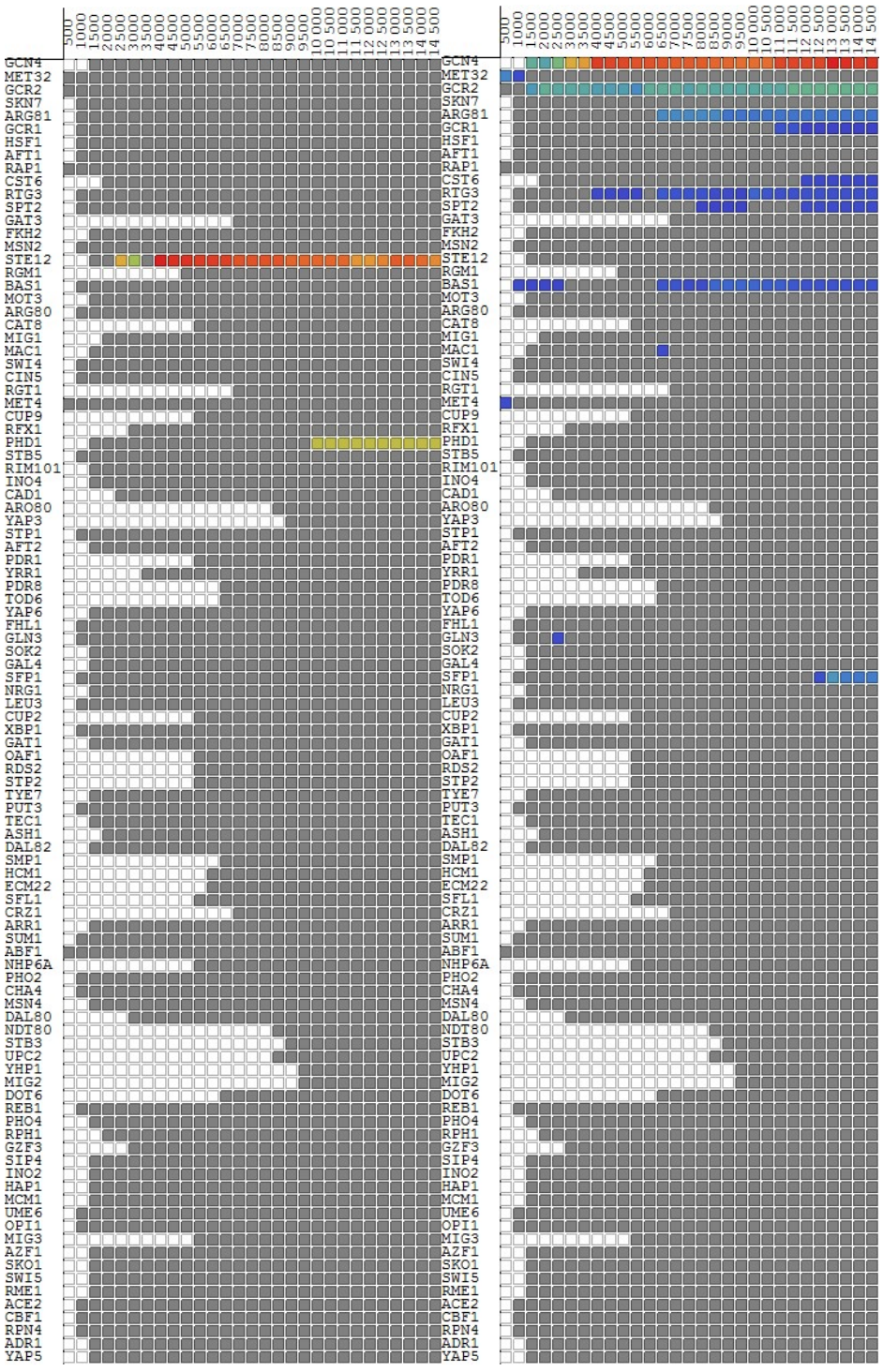


Figure 12. Enrichment statistics calculated for up-regulation (left) and down-regulation (right) using the hypergeometric distribution for networks constructed using the function *Li* for the final likelihood.

Table 1 shows the complete list of enrichment statistics and gene expression based on RNA-seq for all putative transcription factors. The gene for Adr1 had the most significant up-regulation of all transcription factors at 37-fold, consistent with its role as activator of the genes involved with the

utilisation of alternative carbon sources [Young et al. 2003]. The gene for Mig1, which has been suggested to play a lesser role compared to Adr1, was also down-regulated 4-fold, consistent with its role in repression. Genes for other regulators such as Tec1 involved in pseudohyphal development, and Kar4 involved with pheromone signalling, were up-regulated 4.3-fold and 7.1-fold, respectively, suggesting the basis for up-regulated genes in these processes in the xylose medium.

Table 1. Enrichment statistics using only biological prior information of all effectors of transcription in *S. cerevisiae*, regardless of the presence of a top-scoring motif in *K. marxianus* (method SGD). The results were sorted to the minimum of the q-values for either the differential expression, up regulation, or down regulation of a target gene set, as calculated using the hypergeometric distribution after correction for multiple comparisons.

name1	name2	K	Z(0)	nNo	nDown	nUp	qDiff	qUp	qDown	direction	ID	FC
9597	PDR3	1	1.5	0	1	0	0	11.51	0	NA		
	ACA1	2	3.6	0	0	2	0	0	17.05	NA		
9660	YAP3	2	3.6	0	1	1	0	0.72	0.41	NA	gene1726	1
9547	GCN4	169	**11.0**	113	39	17	**1.4E-13**	3.27	**4.9E-16**	down	gene3451	1
9549	GCR2	40	**6.0**	24	14	2	**5.8E-05**	84.92	**2.3E-08**	down	gene2613	0.4
	TUP1	184	**6.8**	134	30	20	**2.4E-08**	0.92	**5.9E-08**	down	gene1081	1
	SPT3	266	**7.2**	203	37	26	**7.1E-08**	1.46	**7.6E-08**	down	gene1396	1
9656	UME6	295	**8.1**	228	25	42	**1.4E-07**	**1.6E-05**	0.27	up		
	HFI1	160	**8.2**	116	26	18	**1.9E-07**	0.90	**8.7E-07**	down	gene1998	1
	SPT20	184	**7.1**	139	29	16	**7.1E-06**	12.30	**2.7E-07**	down	gene1462	1
	BUR6	252	**6.9**	195	17	40	**3.0E-06**	**1.3E-06**	9.63	up	gene3898	1
9578	MET32	205	**6.9**	158	23	24	**3.2E-05**	0.16	**5.6E-03**	down	gene4085	1
	SUA7	401	**3.7**	333	17	51	0.01	**3.2E-05**	116.99	up	gene3433	1
	SPT10	423	**6.2**	344	31	48	**6.1E-05**	**2.1E-03**	0.93	up	gene3729	1
9513	AFT1	68	**5.7**	46	11	11	**7.2E-05**	0.16	**1.2E-02**	down	gene643	1
	CDC73	29	**5.9**	16	8	5	**9.7E-05**	1.47	**8.4E-04**	down	gene635	1
	SIN4	200	**4.1**	164	26	10	0.17	131.89	**1.1E-04**	down	gene3684	1
9541	FKH2	106	**6.8**	77	16	13	**1.1E-04**	1.08	**1.5E-03**	down	gene4226	1
9620	RTG3	38	**4.1**	25	10	3	**4.7E-03**	38.97	**1.6E-04**	down	gene4377	1
9522	BAS1	54	**4.8**	39	12	3	0.03	80.97	**1.6E-04**	down	gene1050	1
9622	SFP1	1306	2.8	1148	94	64	12.2	182.14	**2.4E-04**	down	gene2869	1
	SIN3	73	**5.8**	51	12	10	**3.0E-04**	0.96	0.01	diff	gene262	1
9637	STE12	98	**3.8**	76	3	19	0.05	**3.3E-04**	127.24	up	gene3811	1
9579	MET4	177	**6.8**	137	21	19	**3.5E-04**	1.26	**4.8E-03**	down	gene4937	1
9602	PUT3	33	**5.3**	21	2	10	**4.0E-03**	**5.0E-04**	37.68	up	gene3852	1
9563	HSF1	231	**4.8**	184	24	23	**1.3E-03**	1.68	0.01	diff		
9548	GCR1	414	**3.1**	350	39	25	0.27	96.38	**1.5E-03**	down	gene759	0.2
9545	GAT3	5	**4.6**	1	2	2	**2.9E-03**	0.41	0.18	diff		
	SPT7	69	**4.7**	50	12	7	0.01	11.89	**3.0E-03**	down	gene4603	1
	HMS1	14	**4.4**	7	4	3	0.01	1.68	0.06	NA	gene3793	1
9611	RGT1	5	**3.2**	2	0	3	0.12	0.01	39.73	NA	gene4011	1
	WTM2	19	**4.1**	11	5	3	0.01	5.08	0.03	NA	gene1616	1
9567	IXR1	456	**4.4**	386	39	31	0.21	49.47	0.01	NA	gene4024	1
9530	CST6	308	**3.7**	257	29	22	0.16	39.22	0.02	NA	gene1355	1
9603	RAP1	445	**4.3**	372	34	39	0.02	2.03	0.32	NA	gene3662	1
9684	ZAP1	64	**3.8**	47	7	10	0.02	0.33	1.87	NA		
9599	PHD1	42	2.9	33	0	9	2.50	0.04	160.33	NA		
	KAR4	3	2.9	1	0	2	0.24	0.04	24.98	NA	gene28	7.1
9532	CUP9	7	**3.3**	3	2	2	0.05	1.29	0.59	NA	gene1995	1
9528	CIN5	86	**4.9**	66	8	12	0.05	0.44	3.67	NA		
9610	RGM1	18	**4.0**	11	2	5	0.06	0.10	9.36	NA		
9659	YAP1	330	2.2	285	29	16	9.32	153.65	0.07	NA	gene2340	3.0
	NCB2	53	**3.9**	39	6	8	0.07	0.93	2.18	NA	gene280	1
9650	TEC1	46	**3.5**	35	2	9	0.59	0.08	68.84	NA	gene4599	4.3
	SRB2	15	2.5	11	4	0	3.42	114.29	0.09	NA	gene2081	1
	SRB5	83	**3.1**	64	8	11	0.09	0.93	2.96	NA	gene3502	1
9626	SKO1	12	2.7	8	0	4	1.19	0.09	81.40	NA		
	CSE2	15	**6.0**	9	3	3	0.10	2.18	0.82	NA	gene4152	1

name1	name2	K	Z(0)	nNo	nDown	nUp	qDiff	qUp	qDown	direction	ID	FC
9631	SPT2	32	**4.1**	23	6	3	0.30	25.16	0.11	NA		
9612	RIM101	24	**4.4**	16	4	4	0.12	2.71	0.86	NA	gene1268	2.6
9625	SKN7	85	3.7	66	8	11	0.13	1.13	3.42	NA	gene500	1
9570	MAC1	20	**4.0**	13	3	4	0.14	1.21	2.42	NA	gene2791	1
	MED2	611	1.3	533	45	33	11.50	146.92	0.15	NA	gene1853	1
9568	LEU3	90	2.0	73	11	6	1.50	61.11	0.19	NA	gene4135	1
9584	MOT3	21	**3.7**	14	3	4	0.21	1.51	2.88	NA	gene4028	1
9516	ARG81	18	1.9	13	4	1	1.85	57.41	0.22	NA	gene1497	1
	WTM1	18	**3.4**	12	4	2	0.37	18.15	0.22	NA		
	MBF1	3	**3.8**	1	1	1	0.24	2.07	1.20	NA	gene3984	1
	FLO8	13	**3.2**	8	3	2	0.28	7.99	0.46	NA	gene5077	1
	CYC8	103	**3.1**	82	10	11	0.28	4.83	1.68	NA	gene4998	1
9651	THI2	19	2.3	13	4	2	0.53	20.60	0.29	NA		
9638	STP1	361	2.0	314	29	18	16.52	150.99	0.30	NA		
9531	CUP2	15	**3.2**	10	1	4	0.67	0.30	28.93	NA		
9544	GAT1	12	0.7	9	3	0	6.43	99.34	0.33	NA		
	SSN3	6	2.7	3	2	1	0.33	9.11	0.35	NA	gene4089	1
	HAA1	6	**3.0**	3	1	2	0.33	0.78	5.46	NA	gene2284	1
9524	CAT8	6	**3.3**	3	2	1	0.33	9.11	0.35	NA	gene4748	1
9609	RFX1	381	2.0	328	30	23	5.03	95.48	0.35	NA		
9515	ARG80	20	2.7	14	4	2	0.74	23.16	0.37	NA		
9639	STP2	5	**3.2**	3	0	2	2.06	0.41	39.73	NA	gene1087	1
9551	GLN3	74	2.6	61	9	4	5.32	91.46	0.44	NA		
	SRB8	14	**3.9**	9	3	2	0.44	9.72	0.63	NA	gene1229	1
9581	MIG1	10	3.7	6	2	2	0.46	3.85	1.82	NA	gene892	0.3
	GCN5	525	1.2	461	38	26	29.59	163.51	0.50	NA	gene2846	1
9606	RDS2	7	1.6	5	2	0	6.08	66.98	0.59	NA	gene1626	1
	SSN2	14	**4.3**	10	3	1	2.50	39.93	0.63	NA	gene3001	1
9517	ARO80	11	2.7	7	1	3	0.76	0.64	17.12	NA	gene2338	1
	NGG1	2	0.8	1	0	1	2.22	0.72	17.05	NA	gene3608	1
9621	SFL1	2	2.4	1	0	1	2.22	0.72	17.05	NA	gene1402	1
9543	GAL4	20	2.5	14	2	4	0.74	1.21	12.20	NA		
	RPD3	181	**3.0**	150	16	15	0.76	18.40	0.94	NA	gene1455	1
9593	OPI1	24	1.5	19	4	1	7.61	82.53	0.86	NA		
	GAL11	4	2.7	2	1	1	0.90	3.96	2.32	NA	gene1199	1
9613	RLM1	12	2.2	8	2	2	1.19	6.44	3.11	NA		
	RTF1	3	0.3	2	1	0	6.18	32.43	1.20	NA	gene2376	0.4
9598	PDR8	3	2.0	2	1	0	6.18	32.43	1.20	NA		
	ADA2	27	2.6	20	4	3	1.25	15.59	1.46	NA	gene2995	1
	UME1	7	1.3	5	0	2	6.08	1.29	53.11	NA		
	SWI6	76	2.4	62	4	10	3.10	1.31	53.80	NA	gene79	1
9636	STB5	17	**3.5**	12	2	3	1.36	3.44	8.07	NA	gene457	1
9592	OAF1	8	1.8	6	0	2	8.95	1.98	59.33	NA	gene2265	1
	HIR2	3	1.1	2	0	1	6.18	2.07	24.98	NA	gene691	1
	PGD1	4	0.9	3	1	0	11.47	41.92	2.32	NA	gene877	1
9594	TOD6	4	2.3	3	1	0	11.47	41.92	2.32	NA		
9539	FHL1	184	1.3	161	15	8	39.40	151.60	2.45	NA	gene1172	1
	TFC7	526	-1.4	472	35	19	126.26	183.36	2.63	NA	gene3923	0.4
9614	RME1	9	1.6	7	0	2	12.35	2.83	65.26	NA	gene2302	1
	PIP2	15	1.7	11	2	2	3.42	11.60	5.79	NA		
9527	CHA4	17	2.3	13	1	3	5.87	3.44	35.28	NA	gene2563	1
9600	PHO2	86	0.9	74	8	4	26.86	116.42	3.67	NA	gene1499	1
9627	SMP1	5	0.0	4	1	0	17.75	50.82	3.75	NA	gene786	1
	GTS1	5	0.8	4	1	0	17.75	50.82	3.75	NA		
9683	YRR1	5	1.1	4	1	0	17.75	50.82	3.75	NA		
9615	ROX1	34	1.1	28	4	2	13.40	66.03	3.83	NA		
	MSN1	13	0.7	10	2	1	8.50	35.62	3.90	NA		
9561	HCM1	4	0.3	3	0	1	11.47	3.96	32.53	NA	gene2170	1
9534	DAL81	4	0.7	3	0	1	11.47	3.96	32.53	NA	gene3855	1
9634	STB3	4	0.7	3	0	1	11.47	3.96	32.53	NA	gene3588	1
	TOA2	4	1.1	3	0	1	11.47	3.96	32.53	NA	gene4952	1
9536	ECM22	4	1.2	3	0	1	11.47	3.96	32.53	NA		
9529	CRZ1	4	1.9	3	0	1	11.47	3.96	32.53	NA	gene4817	1
	RGR1	229	2.0	198	17	14	16.50	85.77	4.27	NA	gene4290	1
9591	NRG1	48	1.7	39	5	4	6.10	32.92	4.56	NA	gene2521	1
9525	CBF1	135	0.9	116	11	8	18.57	86.91	4.75	NA		
9514	AFT2	40	2.0	32	3	5	4.92	6.67	22.01	NA		
9589	NHP6A	19	0.9	15	1	3	9.18	5.08	41.79	NA		
9535	DAL82	28	1.8	23	1	4	14.66	5.11	71.11	NA		
9533	DAL80	6	-0.2	5	1	0	24.74	59.16	5.46	NA		

name1	name2	K	Z(0)	nNo	nDown	nUp	qDiff	qUp	qDown	direction	ID	FC
9624	SIP4	6	1.9	5	1	0	24.74	59.16	5.46	NA	gene1834	5.7
9670	YHP1	7	2.2	5	1	1	6.08	12.23	7.40	NA		
	HIR3	5	1.4	4	0	1	17.75	6.33	39.73	NA	gene2903	1
9556	HAP1	52	1.1	45	5	2	38.20	117.90	6.47	NA	gene2706	0.3
9629	SOK2	21	1.5	18	0	3	35.66	7.11	117.86	NA	gene4217	2.5
	MSS11	7	0.9	6	1	0	32.20	66.98	7.40	NA	gene1468	1
9645	SWI4	99	2.7	83	7	9	7.45	16.39	17.79	NA	gene5080	1
9661	YAP5	248	-0.3	221	17	10	86.98	168.73	8.46	NA		
9642	SUM1	44	1.6	36	3	5	8.61	9.98	28.36	NA	gene2211	1
	RTG1	1	-0.5	1	0	0	20.25	11.51	8.73	NA	gene1850	1
9655	UPC2	1	-0.5	1	0	0	20.25	11.51	8.73	NA	gene3170	0.3
9657	USV1	1	-0.5	1	0	0	20.25	11.51	8.73	NA	gene4465	4.3
9582	MIG2	1	1.1	1	0	0	20.25	11.51	8.73	NA		
9542	FZF1	1	1.4	1	0	0	20.25	11.51	8.73	NA		
9586	MSN4	25	1.0	20	2	3	9.11	12.37	20.76	NA		
9553	GZF3	8	0.9	7	1	0	39.95	74.31	9.57	NA	gene2635	1
	SPT6	134	2.3	115	10	9	17.53	60.40	9.64	NA	gene4546	1
9575	MCM1	34	0.8	30	0	4	58.37	10.69	148.97	NA	gene2718	1
9511	ACE2	32	0.6	27	3	2	24.41	59.58	11.55	NA		
9585	MSN2	190	2.1	163	12	15	11.74	25.08	21.36	NA	gene2726	1
9518	ARR1	9	0.3	7	1	1	12.35	19.32	11.92	NA		
	BRF1	9	0.8	7	1	1	12.35	19.32	11.92	NA		
	ADF1	7	0.0	6	0	1	32.20	12.23	53.11	NA		
9523	CAD1	145	1.2	124	10	11	13.29	35.89	15.14	NA		
9601	PHO4	16	0.1	14	0	2	46.83	13.64	99.58	NA		
9607	REB1	618	-0.6	555	36	27	134.04	180.24	14.22	NA	gene4237	1
9509	ABF1	230	2.1	200	15	15	24.39	67.38	14.42	NA	gene2233	1
9554	HAC1	10	0.7	8	1	1	16.25	23.18	14.45	NA		
	MED4	469	0.5	413	28	28	40.91	102.67	14.94	NA	gene2585	1
9555	HAL9	8	0.1	7	0	1	39.95	15.65	59.33	NA	gene2243	1
9677	YOX1	8	1.7	7	0	1	39.95	15.65	59.33	NA	gene2991	1
	TFC6	17	0.5	14	1	2	19.95	15.82	35.28	NA	gene1272	1
9653	TYE7	50	2.3	42	4	4	16.98	36.97	15.94	NA	gene3913	1
9587	NDT80	2	-0.7	2	0	0	38.27	22.31	17.05	NA	gene3958	1
	SIP3	2	0.1	2	0	0	38.27	22.31	17.05	NA	gene119	1
9521	AZF1	2	0.3	2	0	0	38.27	22.31	17.05	NA	gene3930	1
9604	RDR1	2	1.2	2	0	0	38.27	22.31	17.05	NA		
9658	XBP1	764	0.6	678	43	43	69.25	138.79	17.15	NA	gene4846	2.6
	SPT8	29	0.4	25	1	3	38.36	19.16	74.24	NA	gene1143	0.4
	CTR9	9	0.3	8	0	1	47.84	19.32	65.26	NA	gene1458	1
	PPR1	12	1.5	10	1	1	25.31	31.37	19.92	NA	gene2075	1
	SPN1	55	1.3	47	4	4	26.64	47.73	21.73	NA		
9512	ADR1	10	-0.3	9	0	1	55.74	23.18	70.90	NA	gene952	36.9
	HIR1	3	-0.6	3	0	0	54.31	32.43	24.98	NA	gene3197	1
9616	RPH1	14	1.0	12	1	1	35.66	39.93	25.84	NA	gene1960	1
9577	MET31	15	0.1	14	1	0	92.65	114.29	28.93	NA		
9583	MIG3	4	-1.1	4	0	0	68.59	41.92	32.53	NA		
9595	DOT6	4	-0.2	4	0	0	68.59	41.92	32.53	NA	gene5043	1
	NDD1	32	-0.7	29	2	1	87.35	111.17	35.40	NA	gene3058	1
	URE2	5	-0.6	5	0	0	81.30	50.82	39.73	NA	gene3677	1
9662	YAP6	685	-0.9	615	36	34	135.45	170.14	40.14	NA		
	STB1	26	0.8	23	1	2	58.76	40.48	64.73	NA		
	PAF1	39	-0.7	35	1	3	79.34	41.45	102.87	NA	gene3127	1
9646	SWI5	69	0.4	61	4	4	63.79	80.08	42.06	NA	gene3551	1
9566	INO4	53	1.1	47	3	3	67.10	78.35	44.63	NA		
9596	PDR1	20	-0.5	18	1	1	70.01	66.03	45.07	NA	gene452	0.4
9617	RPN4	68	-0.1	61	2	5	87.81	46.16	116.58	NA	gene178	1
9520	ASH1	32	0.1	29	1	2	87.35	59.58	83.38	NA		
	RSF2	80	0.5	72	4	4	96.63	104.43	60.75	NA		
9540	FKH1	544	-1.4	494	26	24	168.28	177.96	79.59	NA		
	MED6	100	-1.4	91	3	6	124.15	80.33	130.04	NA	gene2107	1
	IFH1	54	-1.3	50	1	3	133.20	80.97	135.14	NA	gene960	1
9565	INO2	55	0.4	50	2	3	105.49	83.58	89.92	NA		
9559	HAP4	29	-1.5	28	0	1	155.43	101.19	139.26	NA		
9654	UGA3	265	-1.5	244	11	10	173.54	174.87	112.47	NA	gene3359	1
9573	MBP1	74	-1.5	70	1	3	169.66	126.89	160.74	NA	gene4327	1

Regulators associated with down-regulated target genes

Azf1

Large Azf1 target sets of 743 or 586 genes, respectively, were found by using the *Lm* or *Lm*×*Lc* functions. SGD interaction data proved detrimental to the enrichment score. Aft1 also mapped to poly A and poly T, suggesting that many of these may not be true binding sites. The strong enrichment found in the down-regulated set is in correspondence with the significant over-representation of poly A and poly T using the k-mer networking approach, both for the down-regulated and up-regulated gene sets. These observations, therefore, did not suggest any role for Azf1, but rather that genes with poly A or poly T were often differentially expressed.

Sfp1

For Sfp1, moderate enrichment was found with the method using SGD data only, but the very large gene set of 1 306 targets were halved by requiring a motif (with a score of above 100, function *Li*), and substantially improving the enrichment. Making use of all data (*Lm*×*Lc*×*Li*) resulted in more sites, but decreased the enrichment substantially, suggesting that the PPM of Sfp1 was not optimal for *K. marxianus*, and revealed a substantial number of false binding sites. Using motif strength alone (*Lm*) or *Lm*×*Lc* resulted in no significance, supporting the notion of a sub-optimal PPM. Yet, the fact that the *Li* function provided an improvement over the SGD data, suggested that significant transcriptional rewiring occurred between *K. maxianus* and *S. cerevisiae*, or that a large number of regulatory interactions with Sfp1 in SGD were due to secondary transcriptional effects. The Sfp1 PPM thus deserves attention for improvement.

Met32

Low enrichment was found for Met32 using the SGD set. However, when requiring a motif (*Li*), about half of the genes were found, with a substantial improvement in enrichment score. Emphasising the motif score resulted in lower enrichment, while it was insignificant when only *Lm* or *Lm*×*Lc* was used. This result suggested that the requirement of a motif removed some of the interactions in SGD that may have been derived from secondary effects, or which were not present in *K. marxianus* due to evolutionary differences. Since emphasis on the motif strength decreased the number of targets and the enrichment score somewhat, the motif likely did not discover all of the true targets, and the PPM may need refinement. Since the motif also was short, it may have been out-competed by longer motifs when motif strength was emphasised. There was also a striking similarity between the Met32 motif and the top heptamer in the k-mer network. Thus, the enrichment of the Met32 motif might have been due to differential activity of the zinc finger TF binding the top heptamers, which might likely be

Adr1 or Mig1. However, Adr1 or Mig1 should be associated with the up-regulated target set, suggesting that the potential role of Met32 deserves further evaluation, and that the PPM needs refinement

Bas1

The best enrichment of Bas1 was found using the SGD target set. The requirement for a motif (*Li*) revealed only 16 of the 54 targets, but with a similar score. Allowing motif discovery independent from SGD data (If[*Li*==0,*Lm*×*Lc*, *Lm*×*Lc*×*Li*] or *Lm*+*Lc*+*Li*) resulted in fewer motifs, suggesting the PPM could not result in a score sufficiently high to be included in the top 14 500, and would benefit from improvement. Although the Bas1 motif is similar to that for Gcn4 [Springer et al. 1996], only two of the targets overlapped between the two TFs (see below).

Rgt1

The Rgt1 gene set was discovered by motif strength and conservation between *Kluyveromyces* species (*Lm*×*Lc*), and inclusion of data from SGD was detrimental to the score. There was, however, only five genes in the target set from SGD that mapped to *K. marxianus* genes. The data suggested that there may either be a substantially higher number of targets in *K. marxianus* as opposed to *S. cerevisiae*, or that the true targets in *S. cerevisiae* have not been documented well, or that the Rgt1 motif may be very similar to the motif of another enriched TF. The fact that motif conservation among *Kluyveromyces* species improved enrichment over those discovered by motif strength alone, suggested that the sites were true binding sites for some transcription factor, which may be Rgt1.

Rtg3

Significant enrichment of the Rtg3 gene set was found, based on the SGD data. When motifs were required, 22 of the 38 sites were retained, with a slightly lower enrichment. As the importance of the motif score *Lm* was emphasised, the number of targets decreased and enrichment dropped. The motif is thus likely sub-optimal.

Sut2

A fairly high number of motifs (114) were found for Sut2 using motif score alone, but with borderline significant enrichment. Conservation criteria seemed detrimental to the score, and no significant enrichment was found with the SGD targets alone. Evidence was not sufficient to imply importance of Sut2 in the differential response. The SUT2 gene, which was only annotated in the UFS-Y2791 genome

as such, was 3.4-fold up-regulated, however, suggesting that further evaluation of its importance was required.

Leu3

While the Leu3 target set was not enriched in the SGD set, or by applying an additional requirement for a motif (*Li*), a small set of ten targets was found by motif strength alone. Including conservation selected for 35 targets, and by applying criteria that allowed decoupling of the final likelihood from SGD data (If[*Li*==0,*Lm*×*Lc*, *Lm*×*Lc*×*Li*] or *Lm*+*Lc*+*Li*), more targets were discovered and resulted in a higher significance. The result suggests that the target set in *K. marxianus* was somewhat different as opposed to the model species.

YLR278C

Discovery of the target set of YLR278C based on motif strength alone did not result in significance, while including conservation among *Kluyveromyces* species resulted in a significant enrichment. Inclusion of data from SGD was detrimental, since the targets of YLR278C have not been documented in SGD. Since the enrichment score was rather low, evidence was weak for its involvement in differential expression in *K. marxianus*.

Cup2

Discovery of the target set of Cup2 was based on motif strength alone, whereas including motif conservation among *Kluyveromyces* species was detrimental. No CUP2 gene was found. Evidence was not strong enough to imply Cup2 involvement in the differential response.

Fkh2

Moderate enrichment was found for the Fkh2 gene set from SGD. When the requirement for a motif was included (*Li*), insignificant results were found, as was found when the motif score was emphasised. The motif seemed incorrect for *K. marxianus* and possibly overly specific. Inclusion of pseudocounts in the PPM may improve the enrichment statistic.

YPR022C

YPR022C was of borderline significance. There was also a striking similarity between the YPR022C motif and the top heptamer in the k-mer network. Thus, the enrichment of the YPR022C motif may have been due to differential activity of the zinc finger TF binding the top heptamers, which may likely be Adr1 or Mig1.

Cst6, Mac1, Gln3, Spt2, Met4, Cbf1 and TYE7

Borderline significance was found for each of these TF target sets.

Spt3, Spt20 and Hfi1

Strong enrichment was found for each of these TF target sets. Only SGD data could be used, since these regulators do not have associated DNA binding motifs.

Sin4, Cdc73 and Spt7

Only SGD data could be used for assigning target sets, since these regulators do not have associated DNA binding motifs.

Regulators associated with up-regulated target genes

Ume6

The Ume6 gene set seems to be conserved between *K. marxianus* and *S. cerevisiae*, but the motif used may not be correct for *K. marxianus*. The UME6 gene was not included in the DMKU annotation but was included in the UFS-Y2791 annotation.

Crz1

Crz1 was discovered by sequence conservation among *Kluyveromyces* species alone. This suggested that the motifs matched poorly and that a related TF might be responsible for regulating the gene set. Rgm1, Crz1, and Tea1 all similarly have a preference for a stretch of four guanines or cytosines. Rgm1 mapped to three of the top heptamers in the up-regulated gene set. Since the motif match likelihood was detrimental to the score, the gene set was likely controlled instead by a different TF matching to these heptamers. Adr1 and Mig1 are candidates.

Mcm1

Mcm1 was discovered on motif strength alone. Both conservation criteria and SGD data were detrimental. Mcm1 also mapped to a top heptamer, with overlap by other TF motifs. Since SGD data on Mcm1 did not promote the enrichment, it is possible that the Mcm1 pattern is a proxy for a different TF. The detrimental effect of conservation also suggested that this regulatory set may be unique to the species, but is a less likely scenario as opposed to being a proxy for another TF.

Nrg1

The Nrg1 enriched set was discovered on motif strength alone. Both conservation criteria and SGD data were detrimental to the score. Since SGD data on Nrg1 did not promote the enrichment, it is likely that the Nrg1 pattern is a proxy for a different TF. The detrimental effect of conservation also suggested that this regulatory set may be unique to the species, but is a less likely scenario as opposed to being a proxy for another TF.

Put3

The Put3 gene set seemed to be conserved between *K. marxianus* and *S. cerevisiae*, but the motif used may not be correct for *K. marxianus*, since significant enrichment was only found using the SGD method.

Opi1

The Opi1 set of only six genes was discovered by motif strength and conservation between *Kluyveromyces* species alone, and inclusion of data from SGD was detrimental to the score. The data suggested that the target gene set had diverged between the two species. The OPI1 gene was absent from the DMKU annotation but was present in the UFS-Y2791 annotation.

Azf1

Large Azf1 target sets of 743 or 586 or genes were found using *Lm* or *Lm×Lc*. Other data proved detrimental to the score. Aft1 also mapped to poly A and poly T, suggesting that these may not be true binding sites. The strong enrichment found in the down-regulated set correspondended with the significant over-representation of poly A and poly T using the k-mer networking approach, both for the down-regulated and up-regulated gene sets. This observation did not suggest any role for Azf1, but rather that genes with poly A or poly T were often differentially expressed.

Aft1

All sources of evidence were required for discovery of the Aft1 gene set, using the functions *Lm×Lc×Li* or If[*Li*==0,*Lm×Lc*, *Lm×Lc×Li*]. Using If[*Li*==0,*Lm×Lc*, *Lm×Lc×Li*] obtained more target genes and a higher score as opposed to function A, which allowed the Aft1 binding sites to outcompete the binding sites of other TFs. The gene set from SGD alone, with 68 targets, did not indicate enrichment, suggesting transcriptional rewiring between *S. cerevisiae* and *K. marxianus*, or that the SGD target set contained a significant fraction of interactions resulting from secondary effects.

Swi4

All sources of evidence (*Lm*×*Lc*×*Li*) were required for discovery of the Swi4 gene set. The gene set from SGD alone did not indicate enrichment, and neither did requiring a motif (*Li*), suggesting transcriptional rewiring between *S. cerevisiae* and *K. marxianus*. More target genes and a higher score were obtained by using the *Lm*×*Lc*×*Li* function, as opposed to the If[*Li*==0,*Lm*×*Lc*, *Lm*×*Lc*×*Li*] function, which allowed other motifs with a higher likelihood to outcompete the binding sites for Swi4.

Yap7

Borderline significance was found for each of these TF target sets. A marginally enriched Yap7 gene set was discovered on motif strength alone. Both conservation criteria and SGD data were detrimental. Since SGD data on Yap7 did not promote enrichment, it is likely that the Yap7 pattern is a proxy for a different TF.

Rsc30

Borderline significance was found for each of these TF target sets. A marginally enriched gene set of Rsc30 targets was discovered on motif strength alone. Both conservation criteria and SGD data were detrimental. Since SGD data on Rsc30 did not promote enrichment, it is likely that the Rsc30 pattern is a proxy for a different TF.

Rgm1

Borderline significance was found for each of these TF target sets.

A marginally enriched gene set was discovered by sequence conservation among *Kluyveromyces* species alone. It suggested that the motifs matched poorly and that a related TF might be responsible for regulating the gene set. Rgm1, Crz1, and Tea1 all similarly have a preference for a stretch of four guanines or cytosines. Rgm1 mapped to three of the top heptamers in the up-regulated gene set. Since the motif match likelihood was detrimental to the score, the gene set was likely controlled instead by a different TF matching to these heptamers. Adr1 and Mig1 are strong candidates.

Tea1

Borderline significance was found for each of these TF target sets. The enriched Tea1 target set was discovered by sequence conservation among *Kluyveromyces* species alone. It suggested that the motifs matched poorly and that a related TF might be responsible for regulating the gene set. Rgm1, Crz1, and Tea1 all similarly have a preference for a stretch of four guanines or cytosines. Rgm1 mapped to three of the top heptamers in the up-regulated gene set. Since the motif match likelihood was

detrimental to the score, the gene set was likely controlled instead by a different TF matching to these heptamers. Adr1 and Mig1 are strong candidates.

YNR063W and Cin5

Borderline significance was found for each of these TF target sets.

Bur6 and Sua7

Strong enrichment was found for each of these TF target sets. Only SGD data could be used for assigning interactions, since these regulators do not have associated DNA binding motifs.

Spt10

Only SGD data could be used for assigning interactions, since Spt10 does not have associated DNA binding motifs.

References

Young ET, Dombek KM, Tachibana C, Ideker T. Multiple pathways are co-regulated by the protein kinase Snf1 and the transcription factors Adr1 and Cat8. J Biol Chem. 2003;278(28): 26146–26158.