Investigating uncharacterised genes in *Saccharomyces cerevisiae* using Robot Scientists

Erik Y. Bjurström (erikbju@chalmers.se)[1], Alexander H. Gower (gower@chalmers.se)[2], Praphapan Lasin (b.p.lasin@gmail.com)[1], Otto I. Savolainen (otto.savolainen@chalmers.se)[1,3], Ievgeniia A. Tiukova (tiukova@chalmers.se)[1,4], Ross D. King (rossk@chalmers.se)[1,2,5,6]

1. Department of Life Sciences, Chalmers University of Technology, Göteborg, Sweden
2. Department of Computer Science and Engineering, Chalmers University of Technology, Göteborg, Sweden
3. Department of Clinical Nutrition, University of Eastern Finland, Kuopio, Finland
4. Division of Industrial Biotechnology, KTH Royal University of Technology, Stockholm, Sweden
5. Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK
6. Alan Turing Institute, London, UK

Abstract

YGR067C is an uncharacterised open reading frame in *Saccharomyces cerevisiae*. We hypothesised that its biological function is to regulate respiration during the diauxic shift. We tested this hypothesis by predicting the effect YGR067C deletion would have on the transcriptomic and metabolomic profiles and then comparing it against empirical data. Predictions of the first-order effects was obtained by curating a list of pathways relevant to the hypothesis and higher-order effects was generated using simulation models based on the GEM Yeast9. Empirical data was generated from biological experiments performed in the Robot Scientist Eve where OD_{560}, transcriptomics, and metabolomics data were generated.

We found that *YGR067C* deletion led to downregulation of transcripts in some ethanol consuming respiratory pathways during the glucose phase. During the ethanol phase, we found that NAD+, NADP+ and NADH had increased accumulation, and several amino acid biosynthesis pathways were enriched for the *ygr067cΔ* strain, suggesting longer term consequences of *YGR067C* mediated regulation. Based on these observations we propose that the role of *YGR067C* during the diauxic shift is to regulate genes related to ethanol consumption and respiration in the glucose phase.

This study demonstrates hypothesis instantiation of previously uncharacterised ORFs combined with automated experimentation protocols for hypothesis verification.

Introduction

Saccharomyces cerevisiae is the most studied eukaryotic model organism. Despite this, there are many genes whose biological function are not understood, and a complete understanding of the interactions in the yeast cell is still far from being achieved[1]. The rate at which we progress toward this goal is limited by human capacity for experimentation and experiment design. To increase the effectiveness of systems biology, we can use computational techniques to select or refine hypotheses about function in a way to maximise information gain while minimising the time and economic costs of each, and we can use laboratory automation to increase the quality of empirical data, both important aspects for the wider goal of closed-loop automation of functional genomics. In this study we test and compare computational techniques for refining abstract hypotheses, applying them to investigate the biological role of an uncharacterised protein in *S. cerevisiae*.

YGR067C is an uncharacterised open reading frame (ORF) whose protein product functionality is unknown. What is known is that the gene product contains a zinc finger motif, similar to that of Adr1p, which is a respiratory transcription factor active during the diauxic shift in *S. cerevisiae*[2,3]. The diauxic shift is a metabolic network rewiring event in *S. cerevisiae*, in which the cell goes from fermentative consumption of glucose and production of ethanol, to respiratory consumption of ethanol once the glucose is depleted[4]. During the first growth phase, known as the glucose phase, respiratory genes are repressed by the transcription factor Mig1p[5]. As the glucose levels decrease, Mig1p is phosphorylated by Snf1p, relieving the respiratory genes of its repression[5]. Adr1p, an activator of respiratory genes such as ADH2, is also activated by Snf1p as the glucose levels decrease[3]. Thus, based on previous studies[6] and the structural similarity between YGR067Cp and Adr1p, we hypothesised that YGR067Cp acts as a transcription factor that regulates respiratory genes during the diauxic shift.

To test this hypothesis against experimental data, we need to make a prediction of phenotype that can be compared with empirical data. As shown in Figure 1 to maximise information gain we desire these predictions to be directly comparable to the empirical data, in this case transcriptomics and metabolomics. The first approach we took in achieving this is to identify genes and pathways that we expect to be affected by such regulation and looked for changes in those genes and pathways, in a ~~*YGR067C* deletant~~ *ygr067cΔ* strain with respect to the wild type. This approach relies on the detectability of the first- and second-order effects of deletion, namely those at most one step removed from the deletion in the metabolic network. However, we anticipate the signature of any regulatory effects by *YGR067C* to be realised across more than just the immediately affected pathways, metabolites, and genes but across the whole organism. To predict these higher-order effects of the knockout we need to perform simulations. And to perform simulations we need a computational model of yeast. There are several

computational models available [7], and which to use depends on several factors, including that we require that inputs to and outputs from the model correspond to experimental variables that we can control or measure. The second and third approach we take to predicting phenotype rely respectively on a flux balance analysis (FBA) model and a first-order logic model, LGEM+[8].

Obtaining high quality empirical data to test the predictions is complicated by the nature of the hypothesis. *YGR067C* is hypothesised to be active during the diauxic shift, a transformative metabolic adaptation[9,10]. This necessitates working in dynamical systems, i.e. batch growth, as it would be impossible to observe this effect in stabilised systems such as chemostats[11]. Consequently, minor fluctuations in initial conditions and sampling time will lead to significant deviations in the outcome, which in turn affects the reproducibility of the experiment. To address this problem Robot Scientists (aka self-driving labs) have been developed to automate biological experiments to generate highly reproducible scientific results[6,12–14] . This study utilised the Robot Scientist Eve[15] to cultivate *S. cerevisiae* and generate multiomics data from a diauxic shift experiment.

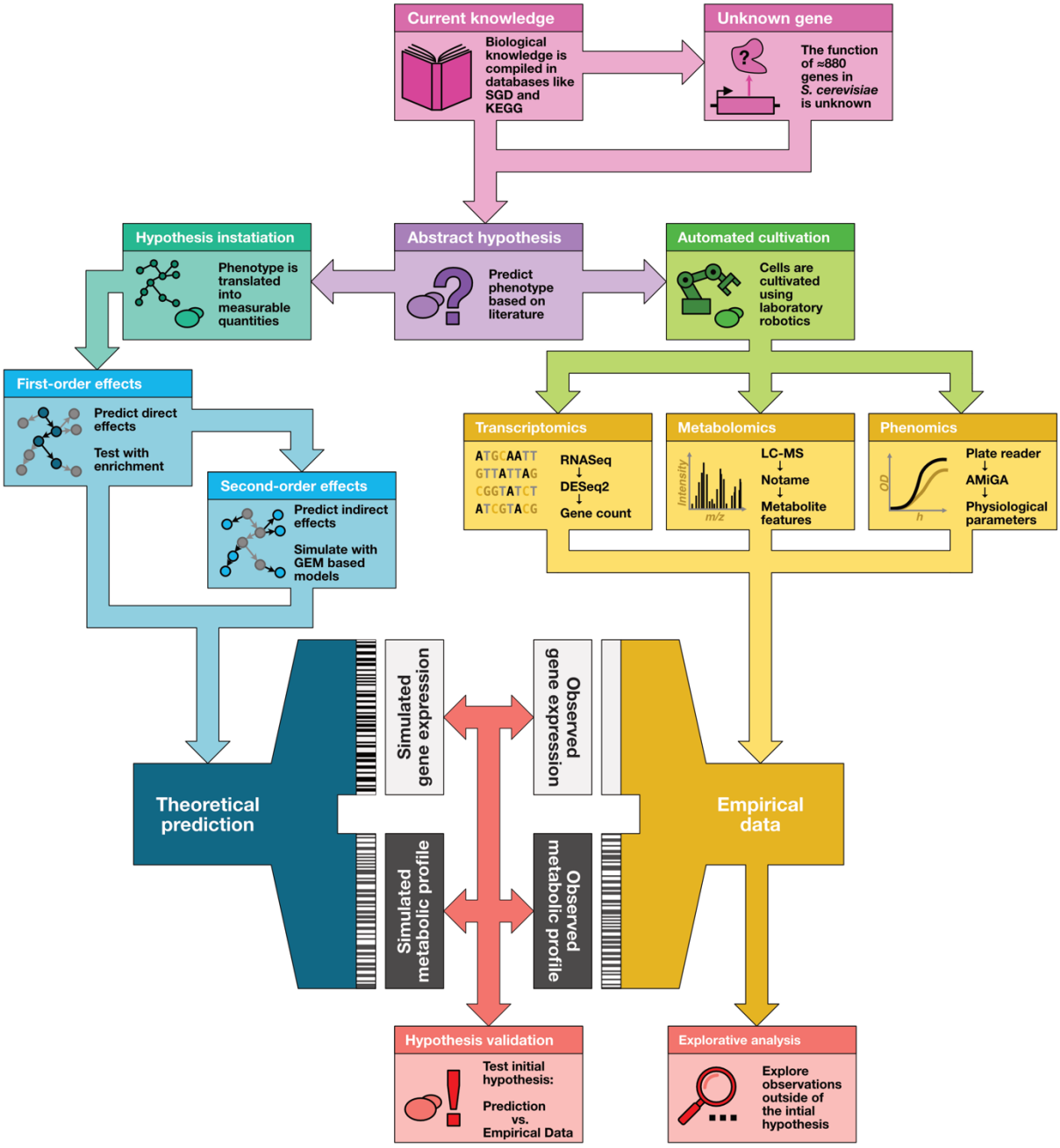

Figure 1: ***Instantiation of an abstract hypothesis.*** *To be able to test a hypothesis, we need to transform it into data that is comparable with empirical data. This instantiation is usually accomplished using a mathematical model; the choice of model, and of the parameters, will change the predictions. To obtain an interface between prediction and measurement where a statistical test can be used, it is most often necessary to transform both the raw model outputs and the raw measurement data. These additional levels of processing risk weakening the signal and amplifying the noise, leading to increased uncertainty in the test, but also enable the opportunity to compare predictions from different models across different measurements.*

Results

We hypothesised that the absence of YGR076C would disrupt respiratory pathways during the diauxic shift in *S. cerevisiae*. To generate the empirical data necessary for testing our hypothesis, a *ygr067cΔ* strain with a reference strain BY4741 was cultivated in the automated laboratory platform Eve. The cells were grown in minimal media with low glucose content (1.25 g/L) to ensure diauxic growth, see Methods section for

detailed composition. Glucose phase and ethanol phase samples were taken after 12 and 24 hours post-inoculation, respectively. Transcriptomic samples were obtained using RNAseq (transcriptomics) and metabolomic samples through liquid chromatography-mass spectrometry (LCMS), see Methods section for further details.

Pathway set prediction

The first order effects of the hypothesised disruption should be observable in the transcription of genes and the metabolism in pathways associated with respiration. A first order prediction was generated by curating a list of genes and metabolites recorded in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database which were associated with respiration, see Methods section. Univariate statistical tests were performed DESeq2 for transcriptomic data and notame for metabolomic. The hypothesis was then tested on the transcriptomic data using consensus gene set enrichment, which included Fisher's exact test, Boschloo's test, and Fast Gene Set Enrichment Analysis (FGSEA). Individual genes and metabolites within the KEGG pathways were also examined.

Transcriptomic analysis indicated that the gene transcription in the tricarboxylic acid (TCA) cycle, oxidative phosphorylation, and glyoxylate pathways were enriched during the glucose phase, see Table 1. Furthermore, FGSEA analysis showed that gene transcription in these pathways was downregulated in the glucose phase, suggesting that transcription is induced by the protein product of *YGR067C* during glucose rich environments. The transcripts of the fermentative glycolysis pathway, pyruvate dehydrogenase complex, and mitochondrial pyruvate importers were not enriched however, see Supplementary Table 4.1. Genes in the gluconeogenesis were statistically enriched during the glucose phase using methods with significance cutoffs (Fisher's and Boschloo's tests) but not FGSEA which considers all genes within a set, see Table 1. The differentially expressed genes (DEGs) within gluconeogenesis, *PYC1* (carboxylase responsible for conversion of pyruvate to oxaloacetate) and *FBP1* (phosphatase responsible for conversion between F1,6BP to F6P), were both downregulated while the genes with an adjusted p-value > 0.05 were upregulated. The transcription of the genes within the predicted pathways were generally downregulated in the deletion mutant during the glucose phase. Only *VMA9* and *VMA10*, both subunits of the H^+-ATPase complex, were both upregulated and part of the KEGG pathway oxidative phosphorylation (sce00190).

Table 1. ***Consensus gene set enrichment of respiratory pathways using data from differential gene expression analysis.*** *Columns describe the statistical test while rows describe phase and pathway. Fisher's test and Boschloo's test are two-tailed hypothesis tests where genes with a FDR < 0.05 were considered significantly differentially expressed. Only distinct regulation was considered for FGSEA and the directions were split into two separate metrics (up- and downregulation). Enrichment analysis with a p-value < 0.05 were considered statistically significant and are marked with bold.*

		Fisher's test	Boschloo's test	FGSEA up	FGSEA down
Glucose phase	TCA cycle	**7.11E-09**	**5.22E-09**	NA	**2.22E-03**
	Oxidative phosphorylation	**4.89E-05**	**4.27E-09**	NA	**1.73E-02**
	Glyoxylate shunt	**3.45E-06**	**2.38E-06**	NA	**2.22E-03**
	Glycolysis (fermentation)	1	1	NA	7.60E-01
	Glycolysis (respiration)	1	1	NA	1.06E-01
	Gluconeogenesis	**4.48E-02**	**3.94E-02**	NA	1.35E-01
Ethanol phase	TCA cycle	1	1	NA	9.03E-02
	Oxidative phosphorylation	1	9.33E-01	NA	2.02E-01
	Glyoxylate shunt	1	8.76E-01	8.72E-01	NA
	Glycolysis (fermentation)	1	1	2.62E-01	NA
	Glycolysis (respiration)	1	1	NA	9.03E-02
	Gluconeogenesis	1	1	8.72E-01	NA

Of the 85 metabolites included in first order prediction set, only 14 could be identified in the empirical data using peak identification software, and of those only 9 passed the quality control during the pre-processing step, see supplementary table X. This was perhaps not surprising as many of the predicted metabolites are short chain carboxylic acids, which are known to be difficult to detect in conventional LCMS methods due to bad reverse phase retention and inefficient electrospray ionisation[16]. The abundance of phosphoenol pyruvate (p-value = 0.077, log2-fold change = -0.0805, linear model) and glutamic acid (p-value =0.078, log2-fold change = 0.87, linear model) were significantly different in the mutant compared to the reference. During the ethanol phase, increased accumulation of nicotinamide adenosine dinucleotides could be observed in the deletion mutant compared to the reference strain: NAD+ (p-value = 0.024, log2-fold change = 2.82, linear model), NADH (p-value = 1.9 x 10^-4, log2-fold change = 2.82,

linear model), and NADP+ (p-value = 0.039, log2-fold change = 3.96, linear model). While the NADH/NAD+ ratio was unchanged (1.00), an increased NADP+/NAD+ ratio could be observed (1.40). Furthermore, increased accumulation of glutamic acid, which is essential for the anabolism, could be observed (p-value = 0.008, log2-fold change = 1.73, linear model) in the ethanol phase.

Results of the FBA simulation method

Using a FBA model with growth as the objective function, we simulated metabolite presence and gene expression for all metabolites and genes in the Yeast9 model, using a pathway perturbation method to instantiate our hypothesis on *YGR067C* function. As FBA does not simulate metabolite accumulation, metabolite presence means that the compound is predicted to be involved in one or more active reactions. Because of the difference in simulation methods, predictions were made for roughly twice as many metabolites as in the LGEM$^+$ method, and for roughly three times as many genes, meaning the pathway coverage is greater than either the pathway set prediction method or the LGEM$^+$ method, see Table 2.

During the glucose phase we predicted that overall 564 genes would be differentially expressed to some degree in the mutant strain; in the ethanol phase this figure was 554. The pathways predicted to be down-regulated the most during the glucose phase in the mutant were glucose fermentation, pyruvate fermentation, glycolysis, and very long chain fatty acid biosynthesis. In the ethanol phase the pathways predicted by the FBA to be most down-regulated in the mutant were various amino acid and nucleotide biosynthesis pathways, and chorismate metabolism; the pathway predicted most up-regulated was the glyoxylate cycle. In contrast to LGEM$^+$, the FBA model predicted that 32 and 28 of the detectable metabolites in the glucose and ethanol phases respectively were differentially expressed in the mutant; the model predicted the direction of these differences with an accuracy of 22%.

Full tables for the predictions with evaluation against transcriptomics and metabolomics data are provided in the supplementary information, see Supplementary Table 4.2.

Results of the LGEM$^+$ simulation method

LGEM$^+$ expresses the graph structure of metabolic networks in mathematical logic, then uses an automated theorem prover to simulate (through logical deduction) activated reactions, metabolites, and genes[8]. Compounds and genes that appear in the LGEM$^+$ simulation are those predicted to be present. The simulations are not quantitative, so presence is binary, and metabolite presence in the case of LGEM$^+$ is defined in the same

way as for FBA simulations. We simulated metabolite presence and gene expression for all metabolites and genes in the Yeast9 model, using the same pathway perturbation as the FBA simulations, see Methods section for details on the LGEM$^+$ simulations. Most metabolites and genes were predicted not to be present, or expressed, in either the wild-type strain or the *YGR067C* deletant, see Table 2. Therefore, the predictions were largely that there was no difference between the strains.

The LGEM$^+$ simulation extends beyond the localised scope of the pathway set prediction, yielding us a prediction of the second-order effect of the hypothesised consequences of *YGR067C* deletion. During the glucose phase this model predicted 37 genes to be differentially expressed in the *YGR067C* deletant compared to the wild type. Genes involved in glycolysis and pyruvate decarboxylation to acetyl CoA were predicted differentially expressed. Of these, 8 were significantly differentially expressed in the empirical transcriptomic data ($p < 0.05$).

Full tables for the predictions with evaluation against transcriptomics and metabolomics data are provided in the supplementary information, see Supplementary Table 4.2.

Table 2. ***Results of simulation methods*** *(† - no support for LGEM$^+$ metabolomics simulations as there was not overlap between detected metabolites and predicted)*

Simulation method	*LGEM$^+$*		*FBA*	
Phase	Glucose	Ethanol	Glucose	Ethanol
# Pred. diff expr. genes	37	39	564	554
Direction prediction acc. (genes)	2%	2%	23%	19%
# Pred. diff expr. metabolites	11	18	444	423
Direction prediction acc. (metabolites)†	–	–	22%	24%
Predicted affected pathways (in ΔYGR067C)	↑ – ↓ –	↑ pentose phosphate pathway ↓ pyruvate decarboxylation to acetyl CoA	↑ – ↓ glucose fermentation, pyruvate fermentation, glycolysis, very long chain fatty acid biosynthesis	↑ glyoxylate cycle ↓ amino acid and nucleotide biosynthesis

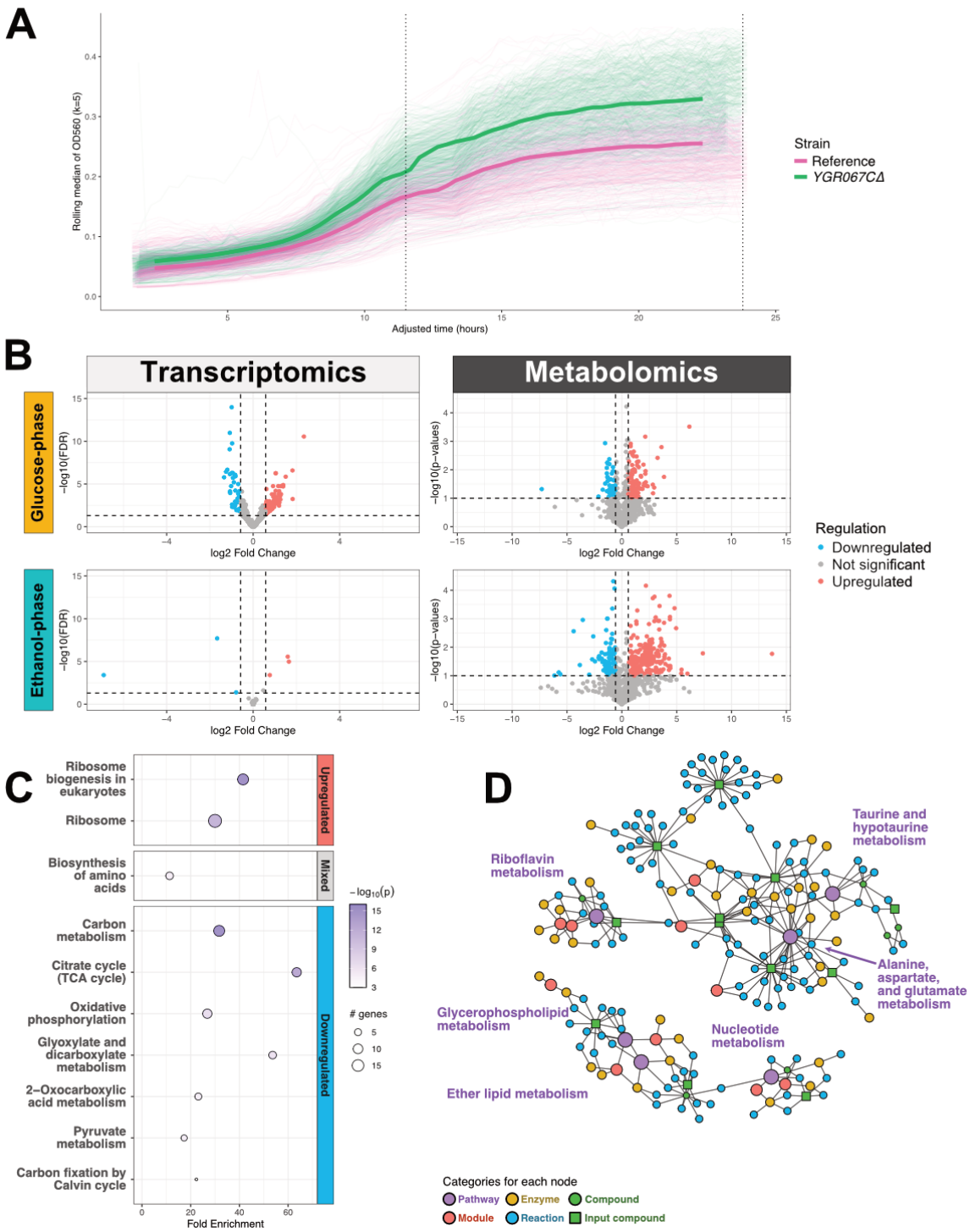

Figure 2. ***Results from exploratory analyses****. A. Averaged growth curves of the ygr067cΔ strain (pink) and the reference strain (green) with the 11.5-hour and 24-hour sampling points as a dotted vertical lines. OD_{560} in the y-axis and time in hours in the x-axis. The bold line represent a rolling median applied over the measured OD_{560}. The thinner lines represent the actual OD_{560} measurements from each replicate. B. Volcano plots representing differential expression of transcriptomics (left column) and metabolomics (right column) during the glucose phase (top row) and the ethanol phase (bottom row). C. Dotplot of pathway enrichment analysis using sub-active networks of the glucose-phase. The top section represents upregulated pathways, middle section mixed regulation, and bottom downregulated pathways. D. Diffusion based topological enrichment using FELLA with significantly enriched pathways in red (FELLA p-score < 0.05).*

Results outside the initial hypothesis

Our initial hypothesis was only concerned with a relatively small subset of the recorded data, namely genes and metabolites related to respiration. However, RNAseq and untargeted LCMS attempts to capture the entire transcriptome and metabolome, respectively. Here we present observations regarding the transcriptome and metabolome which was outside of our initial hypothesis. Additionally, since no physiological prediction was made prior to the experiments, this section also covers growth parameters.

Growth curves were obtained by measuring OD_{560} every 20 minutes with OMEGA Polarstar. The growth curves were then used to estimate physiological parameters (maximum biomass specific growth rate during exponential growth, μ, and carrying capacity, $OD_{560, max}$). A permutation test, n = 10,000, was then performed to compare the growth parameters, see Methods section for detailed explanation. The *ygr067cΔ* strain grew faster on average and reached a higher OD_{560} at 24 hours compared to the BY4741 reference strain, see Figure 2A. For the reference strain, μ was 0.402 h^{-1} and 0.450 h^{-1} for the *ygr067cΔ* strain. The Δμ was therefore 0.048 h-1 and the permutation test yielded a p-value = 0.0001. The OD_{560} at 24 hours post-inoculation was 0.265 for the reference strain and 0.338 for the *ygr067cΔ* strain. The mean difference in OD_{560} at 24 hours was thus 0.073 and the permutation test yielded a p-value of 0.0001. The expression of genes responsible for mitochondrial catabolism in the TCA cycle were downregulated in the *ygr067cΔ* mutant, see the subsection on pathway set prediction. The increased biomass yield and biomass specific growth rate in the *ygr067cΔ* mutant could then perhaps be explained by more energy being funnelled into biomass/population growth instead of mitochondrial proteome investment[17]. Notably however, the cultures in this study were sampled before ethanol could be depleted. It is therefore unknown if the biomass yield in the *ygr067cΔ* mutant would remain higher than the reference strain post-ethanol depletion.

The overall trend of gene expression and metabolite accumulation during each growth phase was investigated. Interestingly, the gene deletion appears to have affected gene expression during the glucose phase, while the gene expression is relatively unchanged during ethanol phase, see Figure 2B. On the other hand, the gene deletion seems to have affected metabolite expression during both phases, see Figure 2B. Furthermore, in the reference strain, we found that *YGR067C* was significantly under expressed during the ethanol phase compared to the glucose phase (FDR = 4.71x10^-33, log2-fold change = 2.05, Wald test with Benjamini-Hochberg (BH) correction), see Supplementary Table 2.1. This suggests that the activity of *YGR067C* — which is assumed to be a transcription factor — is more pronounced during the fermentative glucose phase, while the metabolic consequences of *YGR067C* mediated regulation can be observed in both glucose- and ethanol phase.

There were some notable differences in the intracellular metabolism that occurred outside the scope of the original hypothesis, see Table 4. Increased accumulation of riboflavin, glutamic acid and asparagine was observed during both phases in the deletion mutant. Amino acid expression was mixed during the ethanol phase — accumulation of glutamic acid, asparagine, and tryptophan were increased in the mutant while aspartate and phenylalanine abundances had decreased.

Table 4. **Statistically different metabolite abundances between the deletion strain and reference strain in each phase (p < 0.1, linear model) .** The metabolites were filtered by the list of compounds found in the yeast9 model.

	Query	KEGG	m/z similarity	l2fc	p
Glucose phase	(-)-Riboflavin	C00255	0.99	0.60	0.021
	3',5'-Cyclic AMP	C00575	1	2.94	0.042
	Asparagine	C00152	0.98	1.17	0.083
	Glutamic acid	C00025	0.97	0.87	0.020
	L-Carnosine	C00386	0.85	0.43	0.096
	Phosphoenolpyruvic acid	C00074	0.93	-0.08	0.077
Ethanol phase	(-)-Riboflavin	C00255	0.99	0.41	0.007
	Asparagine	C00152	0.98	0.77	0.085
	Aspartic acid	C00049	1	-0.74	0.090
	Cytidine 5'-diphosphocholine	C00307	1	1.54	0.025
	Glutamic acid	C00025	0.97	1.74	0.008
	Guanine	C00242	0.94	-0.43	0.089
	NAD+	C00003	0.81	2.82	0.024
	NADH	C00004	0.89	2.82	0.000
	NADP+	C00006	1	3.96	0.039
	O-Phosphoethanolamine	C00346	1	-3.56	0.001
	Phenylalanine	C00079	0.86	-1.06	0.057
	Taurocholate	C05122	0.99	1.36	0.078
	Tryptophan	C00078	1	1.48	0.056

To investigate whether pathways outside of the main hypothesis were statistically enriched, topological pathway enrichment was performed on the significantly differentially expressed genes (FDR < 0.05, Wald test with BH correction) and metabolites (p-value < 0.1, linear model). Topological pathway enrichment of the gene expression was performed using pathfindR and only on the DEGs found during the glucose phase as the number of DEGs during the ethanol phase was too small. The TCA cycle, glyoxylate pathway, and oxidative phosphorylation, were all downregulated, see Figure 2D. This was consistent with the hypothesis driven analysis. Furthermore, other pathways related to carbon utilisation which were not part of the hypothesis were downregulated, e.g. the carbon metabolism [sce:01200], 2-oxocarboxylic acid metabolism [sce:01210], and pyruvate metabolism [sce:00620], see Figure 2C.

Ribosome related pathways were significantly upregulated in the *ygr067cΔ* strain, Ribosome [sce:03010] and Ribosome biogenesis in eukaryotes [sce:03008]. Finally, the pathway Biosynthesis of amino acids [sce:01230] was significantly enriched but did not have a distinct direction of regulation.

The topological pathway enrichment of the metabolome was performed using FELLA[18] and both glucose- and ethanol phase datasets. During the glucose phase, several signalling pathways were significantly enriched due to the upregulation of 3'-5'-Cyclic AMP [C00575] in the *ygr067cΔ* mutant, see Supplementary Table 5.1. Alanine, aspartate, and glutamate metabolism [sce:00250] and Riboflavin metabolism [sce:00740] were enriched in both phases, most likely due to the increased accumulation of glutamate and asparagine for the prior pathway and increased accumulation of riboflavin for the latter. During the ethanol phase, apart from the previously mentioned pathways, the following pathways were enriched: Glycerophospholipid metabolism [sce:00564], Taurine and hypotaurine metabolism [sce:00430], Ether lipid metabolism [sce00564], and Nucleotide metabolism [sce:01232] see Figure 2D.

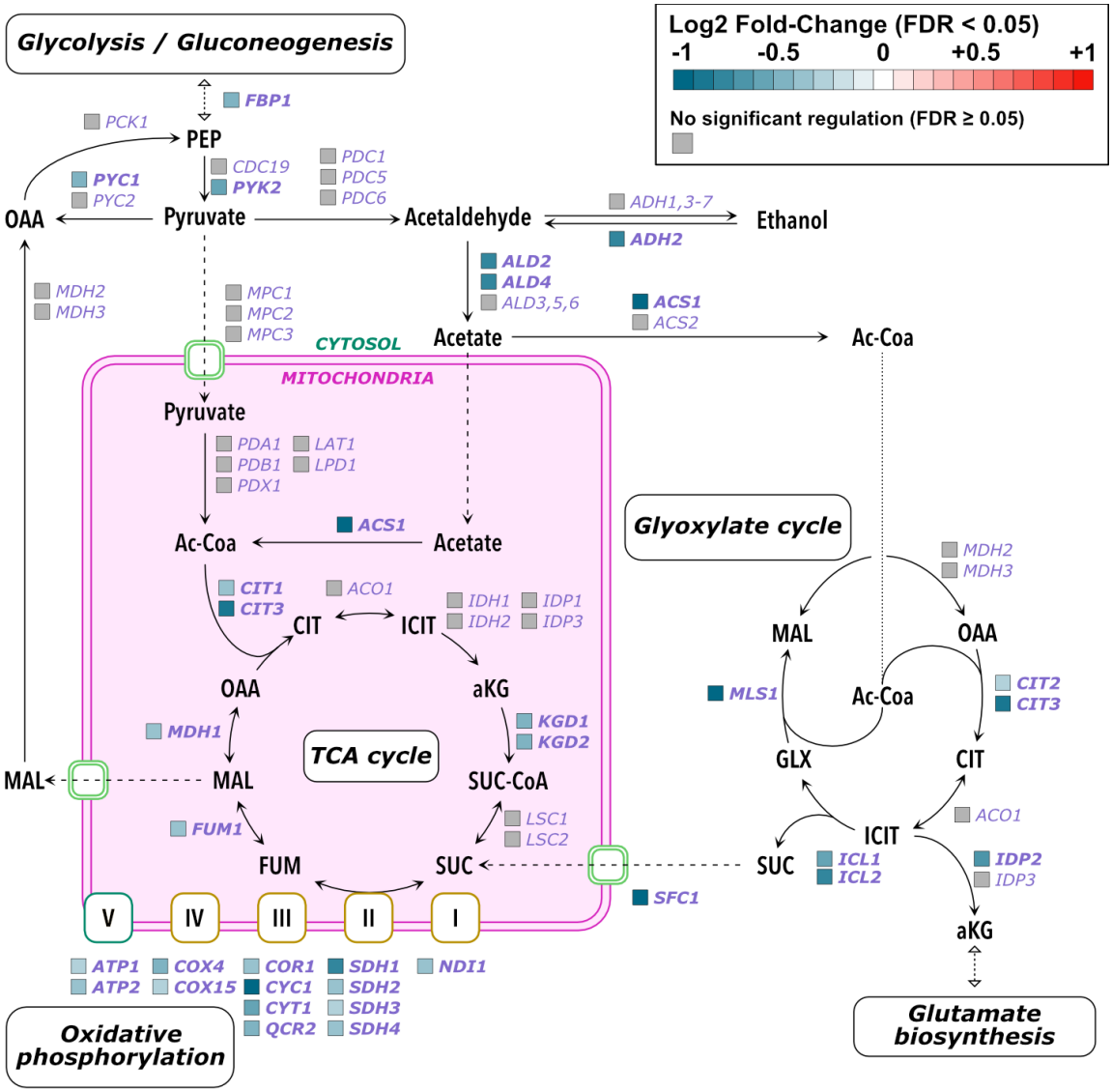

Figure 3. ***Pathway visualization with transcriptomic regulation during the glucose phase.***

Discussion

Automation of science is necessary to address the challenges in the quest to fully elucidating the biological function of every gene in *S.cerevisiae*, and other organisms. When automating the scientific cycle, it is necessary to use a model to translate abstract hypotheses into predictable and measurable outcomes. In systems biology the use of mathematical models is particularly important, as each local change (e.g. a gene knockout) will have not only first-order effects on the systems that gene directly interacts with, but higher order effects on other systems in the organism, and these effects are impossible to calculate without a mathematical model. The process of choosing, refining, and applying a model involves taking many decisions. Each of these modelling choices can have a large effect on the predictions, and thus on the evaluation

of the hypothesis. To fully automate the scientific cycle, the hypothesis instantiation process must be formalised and recorded.

We hypothesised that *YGR067C* induced respiratory pathways during the diauxic shift based on evidence from previous studies. Pathway set prediction (first-order effects) and model-based approaches such as LGEM$^+$ and FBA (higher-order effects) transformed this hypothesis into predictions of differential expression of genes and metabolites. Based on the pathway set prediction it was postulated that a downregulation of respiratory genes and decreased abundances of respiratory metabolites would be observed in the *YGR067C* deletion mutant. The LGEM$^+$ simulations also predicted downregulation in respiratory pathways. The FBA simulations predicted a wider effect on metabolism, including: disruption to fermentation pathways during the glucose phase; and down-regulation of various amino acid and nucleotide biosynthesis pathways, and up-regulation of the glyoxylate cycle during the ethanol phase.

The success of a model-driven approach to hypothesis instantiation is dependent on the quality of the model used. Each of the three simulation methods resulted in different predictions, and neither of the model-based predictions worked exceptionally well. LGEM$^+$ being a discrete model predicted less disruption than the FBA simulations. A significant modelling challenge was how to translate the abstract hypothesis into the simulation. The technique we used in this study to remove subsets of reactions was chosen because these models only contain metabolic genes and reactions. The low predictive accuracy of both models using this random gene removal indicates this technique needs refinement. A hybrid model, that had representations of signalling and gene regulation connected to the metabolic component, could instantiate the hypothesis differently. Since we hypothesise *YGR067C* to be a transcription factor, developing and implementing a hybrid or whole-cell model, such as that proposed in[16], and instantiating the hypothesis in the gene regulation part of a simulation would be closer to our hypothesis.

Another challenge in the modelling is how to predict metabolite accumulation and transcript levels. Both the LGEM$^+$ and FBA models are qualitative in their predictions, and we arrived at a prediction of up- or down-regulation by averaging over repeated simulations. It would be more desirable to have a model that predicted transcript levels and metabolite accumulation quantitatively. This is the subject of active research. One recent approach that might be worth future investigation is to use a deep-learning model to predict accumulation from flux[20].

Pathway set hypothesis testing showed that the TCA cycle, oxidative phosphorylation, glyoxylate cycle were statistically enriched during the glucose phase ($p < 0.05$ FGSEA, Fisher's test, and Boschloo's test) and were furthermore distinctly down regulated ($p < 0.05$ FGSEA), see Table 1. However, transcription of genes responsible for reactions

directing flux from pyruvate to either the TCA cycle or ethanol production were not significantly expressed, see Figure 3. Instead, genes responsible for irreversible steps in the conversion of ethanol to acetyl-Coa, glyoxylate, TCA cycle, and gluconeogenesis were significantly downregulated, see Figure 3. Furthermore, many of the downregulated genes were either glucose repressed (e.g. *PYK2*, *ADH2*, *CIT1*, etc)[5] or induced during consumption of ethanol (*ALD5*, *ALD6*, *PYC1*, etc)[21]. The evidence from this study therefore suggests that it is more likely that *YGR067C* regulates respiratory genes responsible for flux from ethanol rather from glucose, which passes through pyruvate.

There were second-order metabolomic evidence that the glycolytic flux is upregulated in the *ygr067cΔ* strain during the glucose phase. Fermentative glycolysis produces organic acids, such as acetic acid, which acidifies the medium and the cytosol[22]. The pH homeostasis is then maintained through V-ATPase-mediated vacuolar acidification[23]. Two subunits of the V-ATPase complex were upregulated in the deletion mutant, which could be an indication towards increased glycolytic flux. Furthermore, metabolomic analysis showed that 3'-5' cyclic AMP (cAMP) was significantly upregulated in the glucose phase (p-value < 0.1, linear model). cAMP is a crucial signalling metabolite in the Ras/cAMP-pathway that activates protein kinase A, which in turn regulate many processes related to cell growth, such as the diauxic shift[24]. The Ras/cAMP-pathway is activated by intermediate metabolites in the glycolysis, and the increased cAMP levels could thus be due to the increased glycolytic flux in the deletion mutant[23].

While respiratory genes were downregulated in the glucose phase, there was no apparent transcriptomic regulation in the ethanol phase. Interestingly however, a significant increase in accumulation of NAD+, NADH, and NADPH was observed in the ethanol phase. Moreover, tryptophan, the precursor molecule for *de novo* NAD synthesis[25], showed increased accumulation while phenylalanine, which like tryptophan requires chorismite as a precursor[26], saw decreased accumulation. While the relation between the NADH/NAD+ ratio and fermentation, respiration, and aging have been extensively studied[27], there does not appear to be much research on increased accumulation of NAD and its derivatives. It is therefore difficult to explain the mechanism behind the NAD+/NADH/NADP+ accumulation based on knowledge from previous studies. One possible explanation would be if the *ygr067cΔ* strains had a less developed mitochondria prior to the diauxic shift and was rapidly producing mitochondrial proteins as glucose was depleted. The increased NAD levels could then be explained by the requirement of NADP(H) during amino acid synthesis, and it would also explain the increased levels of glutamate in the deletion mutant.

We found that our hypothesis regarding the role of the uncharacterised ORF *YGR067C* was accurate at a high level, but not specific enough about the predicted effects. The

sub-hypothesis, "*YGR067C* induces *ethanol consuming* respiratory pathways *prior to* the diauxic shift" was consistent with the evidence. This lack of specificity in the initial hypothesis has effects on the simulation-based predictions. The accuracy of the simulation-based predictions might have been improved if we had exclusively targeted reactions related to ethanol consumption or had separated the hypothesis into smaller sub-hypotheses, e.g. induction of glucose consuming genes versus induction of ethanol consuming genes.

To conclude, we demonstrate several methods to instantiate hypotheses of uncharacterised genes starting from limited knowledge. The performance of the model-driven approaches showed that the techniques and models require more refinement, which we believe is a worthwhile investment for the future of the field. Finally, based on the results of this study, we suggest that previously uncharacterised ORF *YGR067C* induces ethanol consuming respiratory pathways prior to the diauxic shift.

Methods and materials

Pathway set prediction

The hypothesis stated that the transcription of genes and the metabolism in pathways associated with respiration would be disrupted in the absence of *YGR067C*. In the pathway set prediction approach, the hypothesis was instantiated by selecting KEGG pathways that were predicted to be differentially expressed between fermentation and respiration: the citric acid cycle (sce00020), oxidative phosphorylation (sce00190), Glyoxylate and dicarboxylate metabolism (sce00630), ethanol synthesis from pyruvate (part of sce00010), genes exclusively expressed during gluconeogenesis (part of sce00010), the pyruvate dehydrogenase complex (part of sce00010), and mitochondrial pyruvate carriers (*MPC1, MPC2,* and *MPC3*). Consensus set enrichment, using Fisher's test, Boschloo's test, and Fast Gene Set Enrichment Analysis was then performed FGSEA[28,29]. Fisher's exact test and Boschloo's test tends to be overly conservative while FGSEA produces excessive false positives at times[30,31]. Thus, the methods were chosen to provide a balanced the biological interpretation. The consensus set enrichment was then performed on complete KEGG pathways (sce00020, sce00190, and sce00630) using the predicted gene- and metabolite sets against the empirical transcriptomics and metabolomics data, respectively. The cut-off for the enrichment methods were set to $\alpha = 0.05$. Individual gene- or metabolite regulation was considered when assessing partial pathway predictions.

Simulation using LGEM⁺

Metabolic networks can be described in a graph structure which can then be expressed in mathematical logic. Using automated theorem provers we can conduct simulations through logical deduction, and theory repair (hypothesis generation) through abduction[8]. We constructed a first-order logic model of yeast metabolism based on the consensus genome-scale metabolic model Yeast9 (yeastGEM v9.0.2). This model takes as input a given set of available compounds (in this case the minimal growth medium used for the empirical study), and a goal in the form of a subset of metabolites (the production of a set of compounds deemed essential for yeast to grow). Predictions are logical proofs which correspond to activated reactions, metabolites, and genes. As genome-scale models do not inherently model concentration, those compounds and genes that are included in the LGEM⁺ simulation are those predicted to be present. The simulations are not quantitative, so presence is binary.

Simulation using flux balance analysis

To conduct flux balance analysis simulations, we used the Python library CobraPy (version 0.26.3) with the same version of yeastGEM we used to build the LGEM⁺ model (yeastGEM v9.0.2). The default configuration is for growth in a glucose-rich medium and we used this configuration for the simulations for the glucose phase. For the ethanol phase, we set the bounds for glucose exchange to zero and set the ethanol exchange to be 1.0. We used the default growth objective defined in yeastGEM. To obtain predictions for compounds and genes that are expressed, we took the metabolites and genes associated with each reaction that had a flux greater than a stated threshold (1×10^{-9} mmol g^{-1} $DW h^{-1}$) in the found solution. Similarly to LGEM⁺, presence for each simulation is therefore binary.

Metabolism disruption simulation

YGR067C is not present in Yeast9 v9.0.2 which means that simulating the effect of its deletion from the genome is not directly possible with any computational model built upon Yeast9. We also want to avoid using the empirical transcriptomic data to constrain the simulations, as this would introduce a bias in the simulation, we then want to compare our predictions with the empirical data. So, we need another method of introducing the effect of the deletion into the simulation.

In an initial naïve approach, we looked at the compounds and genes in the pathways associated with respiration. For the LGEM⁺ and FBA simulations, we take these same pathways and randomly remove a subset of them before running a growth simulation. This method aims to model the biological effect of a disruption to respiratory pathways,

which we hypothesise would be the impact of *YGR067C* deletion. (Note that this method assumes deletion of *YGR067C* would have a negative impact on the respiratory pathways in yeast during the diauxic shift; simulation of a positive regulation after deletion would require a different approach.)

This random deletion is repeated N_{sim} times, each time removing a subset of reactions of random size between R_{del}^{-} and R_{del}^{+}. Each simulation results in a prediction of the reactions, metabolites, and genes that are activated, see Table 3. We then calculate the difference between the simulation and the non-disrupted pathway.

Table 3: *Parameters used in the metabolism disruption simulations. * - minimum and maximum number of reactions were found by testing the tolerance of the models to random perturbation, so that the perturbation has a measurable effect on the simulation but does not result in non-growth.*

Parameter	Description	Value used	Comment
N_{sim}	Number of simulations conducted	500	
R_{del}^{-}	Minimum number of reactions removed during disruption simulation	5	Found after testing*
R_{del}^{+}	Maximum number of reactions removed during disruption simulation	12	Found after testing*

Each of these simulations results in a slightly different prediction for the metabolic and transcriptomic activity. Our simulation results are stochastic by nature, the randomness introduced in the size and location of the disruption applied to the model.

The empirical data from growth experiments also have stochasticity. In this case, the randomness arises from many different sources but will vary across cells within the culture. When measuring growth, transcriptomics, and metabolomics, we are measuring the sum of effects of *YGR067C* deletion across all individual cells, smoothing out this stochasticity.

We also sum across our simulations to arrive at data that can be compared to the empirical data, see Figure 1.

Strain selection and cultivation conditions

The *S. cerevisiae* wildtype strain BY4741 (Accession number: Y00000) and single-gene deletion strain BY4741 *YGR067C::kanMX4* (Y04697) were taken from the EUROSCARF deletant library[32]. The strains were revived from -80°C glycerol stocks by cultivating them overnight in YPD (2% (w/v) dextrose) media at 30°C, 220 rpm. The strains were then streaked on YPD plates and incubated at 30 ° C for 3 days. Single colonies were then used to inoculate precultures containing YPD (2% (w/v) dextrose) for 15 h at 30°C, 220 rpm. Finally, the main cultivations were performed in Thermo Fisher 384 well MATRIX plates (Thermo 4332), with a working volume of 80 µL YNB medium (10.5 g/L YNB without amino acids, 1.25 g/L glucose, 75 µg/L ampicillin, and 0.625 g/L of L-

methionine, L-leucine, L-histidine, and Uracil respectively (Brunnsåker et al., 2023}. Each culture was inoculated with an initial OD600 of 0.05, and subsequently incubated at 30°C. Every 20 minutes, the well plate was removed from the incubator, agitated using an orbital shaker, aerated by removing the plate lid, and the OD550 was measured using a plate reader (Polarstar). RNAseq and metabolite samples were taken twice, once after 12 hours after inoculation and again after 24 hours.

Multiomic extraction and processing

Current RNA extraction protocols and LCMS protocols require biomass concentrations which are not feasible with 80 µL cultures. 96 wells were therefore pooled into one biological replicate for the RNA and metabolite extraction protocols using the liquid handler Bravo. The pooled cell broth meant for RNA extraction was then centrifuged (5,000 rcf, 5 min) and the RNA was immediately extracted using RNeasy kits (QIAGEN). The extracted RNA was stored in 30 µL RNAse free water at -80°C. Total RNA quantity and quality was measured using BioAnalyzer. The library construction and sequencing were performed by Azenta in Leipzig, Germany. Data are deposited at European Nucleotide Archive (PRJEB60302). The raw .fastq files were processed using the nf-core/rnaseq v3.10.1 pipeline[33], using the *S. cerevisiae* reference genome Ensembl entry R64-1-1, STAR[34] for fragment alignment and Salmon[35] for quantification.

The extraction protocol is described in a previous study[6]. Untargeted metabolomics profiling was performed on a Waters Xevo G2-sX qTOF high-resolution mass spectrometers (HRMS) coupled to a Waters Acquity Classic UPLC instrument. Metabolites were separated on an UPLC HSS T3 (1.8 µm, 2.1 × 100 mm, Waters) column with a water-MeOH gradient solvent system containing 0.04% formic acid. The gradient started at 5% MeOH with formic acid (MPB) and ramped to 100% MPB over 6 min and held for 4.50 min at 100% MPB. Column temperature was set to 45 °C and the flow at 0.4 mL/min. Mass spectra was acquired using an electrospray ionization (ESI) source in either positive or negative ionization mode scanning from 40 to 1200 m/z at 5 spectra/second. The capillary voltage was set at 1500 V (ESI negative) and 2000 V (ESI positive), and cone voltage at 40. The source temperature was set at 120°C, desolvation a gas temperature at 600°C, desolvation and cone gas flow at 700 and 10 L/min, respectively. Data-dependent MS2 data was collected in both positive and negative ionization by using the following parameters: mass range 40-1200 m/z, MS survey switching threshold 5000, MS survey scanning 0.2 sec, maximum number of precursors 6, scan rate for MS/MS 0.1 sec, collision energy ramp LM CE ramp 6-9 to 60-80 over a mass range of 40-1200 m/z. The raw mass spectra were converted into .mzML files using ProteWizard's msConvert[36]. Peak picking and initial processing were performed using MSDIAL (v5.4)[37]. Identification was performed using the Riken library of both positive and negative ion mode[37]. The identified peaks were then processed using the Notame[38] pipeline in R (v. 4.5.0).

Statistical analysis of empirical data

The phenomic analysis was performed by first compiling the recorded measurements from OMEGA into a .csv file. The .csv file was then used to generate input files compatible for AMiGA[39]. AMiGA then calculates ln(OD_{560}) and d/dt ln(OD_{560}) at each timepoint t. The carrying capacity $OD_{560, max}$ is obtained by finding the maximum value of ln(OD_{560}) during the experiment while the maximum biomass specific growth rate, μ, was obtained by finding the maximum of d/dt ln(OD_{560}). Since we did not know the distribution of the biomass specific growth rate and maximum OD_{560}, the statistical difference between the reference strain and mutant strain were assessed by performing a permutation test, which is a non-parametric test. The observed test statistic, T_obs, was calculated by taking the difference in median response value between the strains. The null distribution was then generated by resampling the growth parameter data and recalculating the test statistic n = 10,000 times. The two-sided p-value was then calculated by counting how many times 1 + the absolute value of the sampled permutations exceeded the absolute value of the observed test statistic, divided by n.

The transcriptomics analysis was performed using the DESeq2 software package[40]. Raw expression data, see Supplementary Table 2.1, were normalised, fit to a negative binomial distribution, and the log2-fold change of low expression genes was adjusted using the DESeq2-package in R. Hypothesis testing was performed using the Wald test and were corrected for false positives using FDR/Benjamini-Hochberg method with a cut-off of FDR < 0.05. The following contrasts were used for this study: *ygr067cΔ* versus reference during glucose phase, *ygr067cΔ* versus reference during ethanol phase. The log2-fold changes (Log2FC) were shrunk using DESeq2's lfcShrink function with the 'ashr' setting[41].

The metabolomics analysis was performed using the notame package[38]. Univariate significance testing of the identified peaks was performed using linear modelling. The signal intensity was set as the dependent variable while the group (*ygr067cΔ* mutant versus reference) was the independent variable. The p-value cut-off was set to p-value < 0.1, similar to previous studies[42].

Topological enrichment analysis was performed using active-subnetwork-oriented enrichment analysis through the pathfindR package for the transcriptomic data. For the metabolomics datam, a diffusion based method was performed using FELLA[18]. The protein-protein interaction network used in both topological enrichment analysis were constructed using KEGG graph objects downloaded from KEGG (date). Topological pathway enrichment was performed on the significantly differentially expressed genes (FDR < 0.05, Wald test with BH correction) and metabolites (p-value < 0.1, linear model).

Data availability

Data deposition: RNA-seq data has been submitted in the form of raw reads in the form of .fastq files under the accession number PRJEB60302 at the European Nucleotide Archive (ENA). Metabolomics data has been submitted in the form of derived spectral .mzML files under the accession number MTBLS12663 at the Metabolights.

Code availability

All code required for reproduction of the analysis and figures in the study can be found on GitHub at https://www.github.com/erikbju/YGR067C_Dshift.

References

1. Wood, V. *et al.* Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? *Open Biol.* **9**, 180241 (2019).
2. Böhm, S., Frishman, D. & Mewes, H. W. Variations of the C2H2 zinc finger motif in the yeast genome and classification of yeast zinc finger proteins. *Nucleic Acids Res* **25**, 2464–2469 (1997).
3. Young, E. T., Dombek, K. M., Tachibana, C. & Ideker, T. Multiple Pathways Are Co-regulated by the Protein Kinase Snf1 and the Transcription Factors Adr1 and Cat8. *Journal of Biological Chemistry* **278**, 26146–26158 (2003).
4. De Deken, R. H. The Crabtree Effect: A Regulatory System in Yeast. *Journal of General Microbiology* **44**, 149–156 (1966).
5. Kayikci, Ö. & Nielsen, J. Glucose repression in *Saccharomyces cerevisiae*. *FEMS Yeast Research* **15**, fov068 (2015).
6. Brunnsåker, D. *et al.* High-throughput metabolomics for the design and validation of a diauxic shift model. *npj Syst Biol Appl* **9**, 11 (2023).

7. Sordo Vieira, L. & Laubenbacher, R. C. Computational models in systems biology: standards, dissemination, and best practices. *Current Opinion in Biotechnology* **75**, 102702 (2022).
8. Gower, A. H., Korovin, K., Brunnsåker, D., Tiukova, I. A. & King, R. D. LGEM+: A First-Order Logic Framework for Automated Improvement of Metabolic Network Models Through Abduction. in *Discovery Science* (eds Bifet, A., Lorena, A. C., Ribeiro, R. P., Gama, J. & Abreu, P. H.) vol. 14276 628–643 (Springer Nature Switzerland, Cham, 2023).
9. Murphy, J. P., Stepanova, E., Everley, R. A., Paulo, J. A. & Gygi, S. P. Comprehensive Temporal Protein Dynamics during the Diauxic Shift in Saccharomyces cerevisiae. *Molecular & Cellular Proteomics* **14**, 2454–2465 (2015).
10. Schlossarek, D. *et al.* Rewiring of the protein–protein–metabolite interactome during the diauxic shift in yeast. *Cell. Mol. Life Sci.* **79**, 550 (2022).
11. Brauer, M. J., Saldanha, A. J., Dolinski, K. & Botstein, D. Homeostatic adjustment and metabolic remodeling in glucose-limited yeast cultures. *Mol Biol Cell* **16**, 2503–2517 (2005).
12. King, R. D. *et al.* The Automation of Science. *Science* **324**, 85–89 (2009).
13. Roper, K. *et al.* Testing the reproducibility and robustness of the cancer biology literature by robot. *J. R. Soc. Interface.* **19**, 20210821 (2022).
14. Sigurdardóttir, S. *et al.* An automated positive selection screen in yeast provides support for boron-containing compounds as inhibitors of SARS-CoV-2 main protease. *Microbiol Spectr* **12**, e01249-24 (2024).
15. Williams, K. *et al.* Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *J. R. Soc. Interface.* **12**, 20141289 (2015).

16. Rathod, R., Gajera, B., Nazir, K., Wallenius, J. & Velagapudi, V. Simultaneous Measurement of Tricarboxylic Acid Cycle Intermediates in Different Biological Matrices Using Liquid Chromatography–Tandem Mass Spectrometry; Quantitation and Comparison of TCA Cycle Intermediates in Human Serum, Plasma, Kasumi-1 Cell and Murine Liver Tissue. *Metabolites* **10**, 103 (2020).
17. Di Bartolomeo, F. *et al.* Absolute yeast mitochondrial proteome quantification reveals trade-off between biosynthesis and energy generation during diauxic shift. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 7524–7535 (2020).
18. Picart-Armada, S., Fernández-Albert, F., Vinaixa, M., Yanes, O. & Perera-Lluna, A. FELLA: an R package to enrich metabolomics data. *BMC Bioinformatics* **19**, 538 (2018).
19. Lu, H., Kerkhoven, E. J. & Nielsen, J. Multiscale models quantifying yeast physiology: towards a whole-cell model. *Trends in Biotechnology* **40**, 291–305 (2022).
20. Morrissey, J., Barberi, G., Strain, B., Facco, P. & Kontoravdi, C. NEXT-FBA: A hybrid stoichiometric/data-driven approach to improve intracellular flux predictions. *Metabolic Engineering* **91**, 130–144 (2025).
21. Brewster, N. K., Val, D. L., Walker, M. E. & Wallace, J. C. Regulation of pyruvate carboxylase isozyme (PYC1, PYC2) gene expression in Saccharomyces cerevisiae during fermentative and nonfermentative growth. *Arch Biochem Biophys* **311**, 62–71 (1994).
22. Casal, M., Cardoso, H. & Leao, C. Mechanisms regulating the transport of acetic acid in Saccharomyces cerevisiae. *Microbiology* **142**, 1385–1390 (1996).

642 23. Deprez, M.-A., Eskes, E., Wilms, T., Ludovico, P. & Winderickx, J. pH homeostasis
643 links the nutrient sensing PKA/TORC1/Sch9 ménage-à-trois to stress tolerance and
644 longevity. *Microb Cell* **5**, 119–136 (2018).
645 24. Thevelein, J. M. & De Winde, J. H. Novel sensing mechanisms and targets for the
646 cAMP–protein kinase A pathway in the yeast *Saccharomyces cerevisiae*. *Molecular*
647 *Microbiology* **33**, 904–918 (1999).
648 25. Savitz, J. The kynurenine pathway: a finger in every pie. *Mol Psychiatry* **25**, 131–147
649 (2020).
650 26. Braus, G. H. Aromatic amino acid biosynthesis in the yeast Saccharomyces
651 cerevisiae: a model system for the regulation of a eukaryotic biosynthetic pathway.
652 *Microbiol Rev* **55**, 349–370 (1991).
653 27. Odoh, C. K., Guo, X., Arnone, J. T., Wang, X. & Zhao, Z. K. The role of NAD and NAD
654 precursors on longevity and lifespan modulation in the budding yeast,
655 Saccharomyces cerevisiae. *Biogerontology* **23**, 169–199 (2022).
656 28. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach
657 for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**,
658 15545–15550 (2005).
659 29. Boschloo, R. D. Raised conditional level of significance for the 2 × 2-table when
660 testing the equality of two probabilities. *Statistica Neerlandica* **24**, 1–9 (1970).
661 30. Abatangelo, L. *et al.* Comparative study of gene set enrichment methods. *BMC*
662 *Bioinformatics* **10**, 275 (2009).
663 31. Dinu, I. *et al.* Improving gene set analysis of microarray data by SAM-GS. *BMC*
664 *Bioinformatics* **8**, 242 (2007).

32. Giaever, G. & Nislow, C. The Yeast Deletion Collection: A Decade of Functional Genomics. *Genetics* **197**, 451–465 (2014).
33. Harshil Patel *et al.* nf-core/rnaseq: nf-core/rnaseq v3.10.1 - Plastered Rhodium Rudolph. Zenodo https://doi.org/10.5281/ZENODO.7505987 (2023).
34. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
35. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417–419 (2017).
36. Adusumilli, R. & Mallick, P. Data Conversion with ProteoWizard msConvert. in *Proteomics* (eds Comai, L., Katz, J. E. & Mallick, P.) vol. 1550 339–368 (Springer New York, New York, NY, 2017).
37. Tsugawa, H. *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods* **12**, 523–526 (2015).
38. Klåvus, A. *et al.* "Notame": Workflow for Non-Targeted LC–MS Metabolic Profiling. (2020).
39. Midani, F. S., Collins, J. & Britton, R. A. AMiGA: Software for Automated Analysis of Microbial Growth Assays. *mSystems* **6**, 10.1128/msystems.00508-21 (2021).
40. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
41. Stephens, M. False discovery rates: a new deal. *Biostat* kxw041 (2016) doi:10.1093/biostatistics/kxw041.

42. Raguz Nakic, Z., Seisenbacher, G., Posas, F. & Sauer, U. Untargeted metabolomics unravels functionalities of phosphorylation sites in Saccharomyces cerevisiae. *BMC Syst Biol* **10**, 104 (2016).

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by Knut and Alice Wallenberg foundation and the Swedish Research Council Formas (grant agreement no. 2020-01690). Funding was also provided by the Chalmers AI Research Centre.

This work has been supported by the UK Engineering and Physical Sciences Research Council (EPSRC) [EP/R022925/2, EP/W004801/1 and EP/X032418/1],

The computations/data handling were enabled by resources in project NAISS 2023/22-185 (computation) and NAISS 2023/23-80 (data storage) provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at UPPMAX, funded by the Swedish Research Council (grant agreement no. 2022-06725).

Author Contributions

Conceptualisation and design of study: E.Y.B., I.A.T, and R.D.K. Hypothesis instantiation and simulation: A.H.G. Cultivation experiment: E.Y.B., A.H.G., and P.L. Multiomics extraction: P.L. and E.Y.B. LC/MS processing: O.I.S. Data analysis: E.Y.B. and A.H.G. Writing: E.Y.B., A.H.G., O.I.S, I.A.T., and R.D.K. All authors read and approved the final manuscript.

Funding

Open access funding provided by Chalmers University of Technology.

Competing Interests

The authors declare no competing interests.

Ethics Approval and Consent to Participate

The authors declare no competing interests.

Supplementary material

S. Table 1 – Log-files for growth curves obtained using omega polarstar

S. Table 2.1 – Normalized transcript counts

S. Table 2.2 – Differential expression tables of RNAseq data

S. Table 3.1 – Aligned metabolite data

S. Table 3.2 – Differential expression tables of Metabolite data

S. Table 4.1 – Pathway set prediction

S. Table 4.2 – Simulated prediction

S. Table 5.1 – Topological enrichment, metabolomics