Quantitative single-cell analysis of *S. cerevisiae* using a microfluidic live-cell imaging platform

THÈSE N^{O} 6519 (2015)

PRÉSENTÉE LE 13 FÉVRIER 2015

À LA FACULTÉ DES SCIENCES DE LA VIE

UNITÉ DU PROF. NAEF

PROGRAMME DOCTORAL EN BIOTECHNOLOGIE ET GÉNIE BIOLOGIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Johannes BECKER

acceptée sur proposition du jury:

Prof. M. Dal Peraro, président du jury
Prof. F. Naef, Prof. S. Maerkl, directeurs de thèse
Prof. J. McKinney, rapporteur
Prof. R. Schneider, rapporteur
Prof. D. Shore, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2015

Numbers is hard and real and they never have feelings
But you push too hard, even numbers got limits
Why did one straw break the camel's back? Here's the secret
The million other straws underneath it: it's all mathematics
— Yasiin Bey, 1998

Magnets, how do they work? ...
— Shaggy 2 Dope (Insane Clown Posse), 2009

Acknowledgements

They say a midget standing on a giant's shoulders can see
much further than the giant
— Jay Z, 2002

There were a few giants throughout my PhD that helped me to see much further. Prof. Felix Naef gave me the opportunity to work in his lab and always provided the freedom to make my own findings, while giving me the support I needed whenever my studies threatened to go astray. The same goes for Prof. Sebastian Maerkl, who always had a good idea and an even better quip for my support.

During my thesis I had two main collaborators and without their help my studies would not have been the same. Nicolas Dénervaud is a good mentor, great friend and the best at including dirty jokes while explaining how to handle microfluidic devices. Speaking of dirty jokes, I have never been insulted in funnier and more creative ways than by Poonam 'Humpy' Bheda. Her passion for biology (mixed with the right amount of craziness) was good support during the later stages of my PhD. I am very glad that she and Rob Schneider stumbled into my life and I hope that my work will be helpful for them.

When worst comes to worst, my peoples come first
— Havoc (Mobb Deep), 1995

There are a lot of people that make my life far from the worst and I am certain I will forget to mention a few. The Siegerts: Manu, who makes worse jokes than me; Steffi and Emelie, who have to endure them. The Coffee Crew: Julia, friend, little sister and voice of reason all at once; Carrie, who always has my back and most of the time her ish together. The Amanda's: Lund (and her Bouba) aka The Party Amplifier; Verpoorte, always happy and always a scapegoat

when the microscope room was messy. The Flower Family: You guys are like a cross between the A-Team and a wrecking ball. Triple A's, 2moreRaw and extended Moscarojan family: There are too many nice people and too much alcohol involved to mention or remember them all ☺. My two labs, especially the microfluidic lunch (aka the poor people that had to listen to me every day). Kalle Zwo: The bike that carries my butt around for the last eight years. Point velo: The place where Kalle Zwo got repaired at least 100 times. Everyone that ever drank a beer with me at Great Escape/Sat/SV Happy Hour; I'll mention Salem, Jairo, Scott and Ryan Brian as significant representatives of this group. The Amazing PMI Ladies: Glad that I can be your plus one. Lausanne-Ville / Prilly Basket: Often games to forget, always nights to remember. NWK and extended Watzeverdel: I wish I would see you more often (I'll try to change this).

This is family business
and this is for everybody standing with us
—Kanye West, 2004

Finally, I would like to thank my family. Every Christmas time or summer vacation, when we are all stuffed together, I get reminded how lucky I am to have them in my life. So, this is to my parents, you two are the definition of Goethe's "Wurzeln und Flügel" quote. And to my siblings, all three a bit different, all three absolutely great. Also, kudos to extending our family with topnotch spouses and sprouts. And of course this thesis is dedicated to Zeina, who could not be more supportive, positive and simply amazing than she is. Every day with you is a joyful one.

Lausanne, 13 January 2015 J. B.

Abstract

Genome-wide manipulations and measurements have made huge progress over the last decades. In *Saccharomyces cerevisiae*, a well-studied eukaryotic model organism, homologous recombination allows for systematic deletion or alteration of a majority of its genes. Important products of these manipulation techniques are two libraries of modified strains: A deletion library consisting of all viable knockout mutants, and a GFP library in which 4159 proteins are successfully tagged with GFP. In addition, the development of a method that allows for the systematic construction of double mutants led to a virtually infinite number of potential strains of interest.

These advancements in combinatorial biology need to be matched by methods of data measurement and analysis. In order to simultaneously observe the spatio-temporal dynamics of thousands of strains from the GFP library, Dénervaud et al. developed a microfluidic platform that allows for parallel imaging of 1152 strains in a single experiment. On this platform, strains can be grown and monitored in a controllable environment for several days, which results in the imaging of several millions of cells during one experiment.

To objectively and quantitatively analyze this immense amount of information, we implemented an image analysis pipeline, which can extract experiment-wide information on single-cell protein abundance and subcellular localization. The construction of a supervised classifier to quantify localization information on a single cell level is a new approach and was invaluable to detect dynamic localization changes within the proteome.

Using five different stress conditions, we gained insight into temporal changes of abundance and localization of multiple proteins. For example, we found that while localization changes can often be fast and transient, long-term response of a cell is usually enabled by changes in abundance. This shows a well-orchestrated response of a cell to external stimuli.

To extend knowledge about cellular mechanisms, we used our microfluidic platform for two

separate screens, combining GFP-reporter with additional deletion mutants. The advantage of our platform in comparison to more common approaches lies in its simultaneous measurement of fluorescence and phenotypic information on cell size and growth. For each deletion, we can quantify not only its influence onto the respective GFP-reporter under changing conditions, but also its effect on cell growth and size. We showed that it is advantageous to combine this information, as it allows pointing out possible underlying mechanisms of gene network regulations.

In a first screen we investigated the behavior of several gene networks upon UV irradiation damage. We were able to show that four gene deletions influenced the localization of ribonucleotide-diphosphate reductase (Rnr4p).

A second screen was designed to find genes that influence the induction of the galactose network. This screen uses more than 500 deletions of genes mostly related to chromatin in combination with two different reporter strains. A main focus of this study was the inheritance of memory during galactose reinduction. We found several previously unknown genes that potentially influence either induction or reinduction and were picked as candidates for further inheritance studies.

Our microfluidic platform allows for unprecedented studies of proteomes in flux. This thesis shows the potential of the platform and highlights the quantitative analysis, which needs to be able to cope with the amount and complexity of data in high throughput live cell imaging.

Keywords: Microfluidics, Live-cell arrays, Time-lapse microscopy, Single-cell image analysis, *Saccharomyces cerevisiae*, Yeast GFP collection, Protein abundance, Protein localization, DNA damage response, Environmental stress response (ESR), Ultra-Violet irradiation (UV), Synthetic genetic array (SGA), Galactose network (GAL).

Zusammenfassung

Genomweite Manipulationen und Messungen haben in den letzten Jahrzehnten große Fortschritte gemacht. In *Saccharomyces cerevisiae*, einem gut untersuchten eukaryotischen Modellorganismus, ermöglicht homologe Rekombination systematische Löschung oder Änderung der Mehrheit seiner Gene. Wichtige Produkte dieser Manipulationstechniken sind zwei Kollektionen modifizierter Stämme: Eine Löschungs-Kollektion, bestehend aus allen lebensfähigen Knockout-Mutanten, und eine GFP-Kollektion, in der 4159 Proteine erfolgreich mit GFP markiert wurden. Darüber hinaus führt die Entwicklung eines Verfahrens, das die systematische Konstruktion von Doppelmutanten erlaubt, zu einer nahezu unendlichen Anzahl potentiell interessanter Stämme.

Diese Fortschritte in der kombinatorischen Biologie benötigen angepasste Methoden der Datenerfassung und -analyse. Um gleichzeitig die raum-zeitliche Dynamik tausender von Stämmen aus der GFP-Kollektion ermöglichen zu können, entwickelten Dénervaud *et al.* eine mikrofluidische Plattform mit einer parallelen Bildgebung von 1152 Stämmen in einem einzigen Experiment. Auf dieser Plattform können Stämme gezüchtet und in einer kontrollierbaren Umgebung für mehrere Tage überwacht werden, was zu der Aufnahme von mehreren Millionen Zellen während eines Experiments führt.

Um diese immense Menge an Informationen objektiv und quantitativ zu analysieren, implementierten wir eine Bildanalyse-Pipeline, die experimentweit Informationen über Protein-Menge und subzelluläre Lokalisierung extrahieren kann. Der Bau eines überwachten Klassifikators zur quantitativen subzellulären Proteinlokalisierung auf Einzelzellebene ist ein neuer Ansatz und ist von unschätzbarem Wert, um dynamische Veränderungen innerhalb des Proteoms zu erfassen.

Mit der Beobachtung von fünf verschiedenen Stressbedingungen gewannen wir einen Einblick in die zeitlichen Änderungen der Menge und subzellulären Lokalisierung von mehreren Prote-

inen. Zum Beispiel haben wir festgestellt, dass, während die Lokalisierungsveränderungen oft schnell und vorübergehend sein können, langzeitige Reaktionen einer Zelle in der Regel durch Veränderungen in Proteinkonzentrationen gesteuert sind. Dies zeigt eine gut organisierte Reaktion einer Zelle auf äußere Reize.

Um das Wissen über zelluläre Mechanismen zu erweitern, haben wir unsere mikrofluidische Plattform für zwei separate Screens benutzt, in denen GFP-Reporter mit zusätzlichen Deletionsmutanten kombiniert wurden. Der Vorteil unserer Plattform im Vergleich zu gewöhnlicheren Ansätzen liegt in der gleichzeitigen Messung von Fluoreszenz und phänotypischen Informationen. Für jede Löschung können wir nicht nur ihren Einfluss auf das jeweilige GFP-Reportergen unter wechselnden Bedingungen messen, sondern auch ihre Wirkung auf das Zellwachstum und die Größe. Wir haben gezeigt dass es vorteilhaft ist diese Informationen zu kombinieren, denn sie ermöglichen den Hinweis auf potentiell zugrundeliegende Mechanismen der Gen-Netzwerk-Regulierung.

In einem ersten Screen untersuchten wir das Verhalten mehrerer Gen-Netzwerke bei UV-Bestrahlungsschäden. Wir konnten zeigen, dass vier Gen-Knockouts die Lokalisierung von Ribonukleotid-Diphosphat-Reduktase (Rnr4p) beeinflussen.

Der zweite Screen wurde entwickelt, um Gene zu finden, die die Induktion des Galaktose-Netzwerkes beeinflussen. Dieser Screen untersucht mehr als 500 Löschungen von Genen, die weitestgehend mit dem Chromatin zusammenhängen in Kombination mit zwei unterschiedlichen Reporterstämmen. Ein Schwerpunkt der Studie war das Vererben von Informationen während Galactose-Reinduktion. Wir fanden mehrere bisher unbekannte Gene, die möglicherweise Einfluss auf entweder Induktion oder Reinduktion haben. Sie wurden als Kandidaten für weitere Vererbungsstudien aufgenommen.

Unsere mikrofluidische Plattform ermöglicht beispiellose dynamische Studien des Proteoms. Diese Arbeit zeigt das Potenzial der Plattform und unterstreicht die quantitative Analyse, die in der Lage sein muss, die Menge und Komplexität der Lebendzellmikroskopiedaten zu bewältigen.

Stichwörter: Mikrofluidik, Lebendzellmikroskopie, Einzelzell Bildanalyse, *Saccharomyces cerevisiae*, Hefe GFP Kollektion, Proteinvorkommen, Proteinlokalisierung, DNA-Schädigungsreaktion, Umweltbeeinflusste Schadensreaktion, UV-Strahlung, Galaktose Netzwerk

Contents

List of Figures

List of Tables

1 Introduction

Systems biology has emerged as the main approach to study complex interactions, sparked by the ongoing development of high-throughput technologies. Using a holistic viewpoint, it generally tries to identify all parts of a complex network instead of focusing on few interactions. For example, a main part of systems biology is the understanding of gene regulation and protein activity as complete networks or even on a genome or proteome-wide level, respectively [1]. It is a common approach to combine high-throughput techniques, to obtain system-wide information, which is then integrated into a model, aiming to understand the functioning of a cell or organism as a whole.

Two important aspects of systems biology are genomics and proteomics. Genomics on the one hand focuses its studies on the gene content of DNA and its transcription. Since first sequencing techniques have been published 40 years ago, thousands of complete genome sequences became available and gene analysis techniques have evolved from southern blotting over DNA microarrays to genome-wide sequencing methods like Chromatin immunoprecipitation followed by deep sequencing (ChIP-Seq). The advancements of these technologies helped to identify the role of genes in a genome-wide manner, as they allow for large studies that could for example highlight gene abundance variation over time [2] or under different conditions [3].

For the proteome, methods like the yeast two-hybrid system (Y2H) or affinity-purification coupled with mass spectrometry (AP-MS) are well suited to give us a good overview about protein interaction [4]. A major impact on protein analysis was the systematic approach to tag

each open reading frame (ORF) of *Saccharomyces cerevisiae* with green fluorescent protein (GFP) undertaken by Huh *et al.* [5]. This led to a collection of 4159 clones (67% of all ORFs), which helped in proteome-wide understanding of protein localization and later on abundance and noise under static conditions [5, 6, 7].

Fluorescent protein markers combined with time-lapse microscopy enabled dynamic observations like the shuttling of proteins between different subcellular compartments [8] or molecular processes during the cell cycle [9]. Recently, several large-scale screens used the GFP library on standard microtiter plates to study static differences. For example, these screens were able to detect changes in abundance and localization following treatment with methyl methanesulfonate (MMS) and hydroxyurea (HU) [10] and in response to DTT, H2O2, and nitrogen starvation [11].

But the underlying dynamic of responses still remained hidden, as none of these approaches allow for control of culture conditions, as a continuous supply of medium cannot be provided. Microfluidic devices are capable of overcoming this problem, as continuous flow allows for static conditions, while being able to control culture size. Furthermore, they allow for dynamic changes of conditions. For example, first devices with living cell cultures were used on a low number of strains to measure the gene expression of single strains upon changes in the environment [12] or to analyze complete pathways [13].

However, these devices were still limited in the amount of strains that could be used, mostly by relying on manual loading of the device. To overcome this drawback, we developed a microfluidic device that uses a DNA spotter to spot living cells, instead of loading the device via the more common approach of flow-in [14]. This device allows for the parallel study of up to 1152 different strains over several days, with a time resolution of 20 minutes, resulting in close to 100,000 images for one single experiment.

The importance of the parallel study of thousands of genetically modified strains increased over the last decade, as genome wide modifications became possible in both bacteria [15] and simple eukaryotes [16, 17]. For several reasons, we chose *S. cerevisiae* as the primary model of our live-cell imaging platform. A technical reason was the robustness of yeast, which was found to survive the stressful process of cell spotting. The biological reasons are the same that make budding yeast one of the best-studied model organisms. Being a eukaryote, yeast possesses a lot of processes that are comparable in more complex organisms. At the same time,

yeast is similar to a lot of bacteria in that it provides an efficient background for large-scale genetic manipulations [18].

We first used the aforementioned GFP library to study the changes in protein abundance and localization under dynamic conditions. Therefore, a robust automatic image analysis pipeline was indispensable for data quantification. Combining bright field and fluorescence images, this pipeline allowed us to extract valuable single-cell information about the proteome in flux.

Independent of the GFP library, a second library in *S. cerevisiae* was obtained through the development of a deletion mutant array (DMA) [17], containing every viable knockout ORF of the yeast genome, resulting in more than 5000 viable strains. This DMA can be used for Synthetic Genetic Array analysis (SGA), a high-throughput technique that allows to systematically construct collections of double mutants. A typical use of this method is the large-scale analysis of double deletion mutants to detect synthetic lethal and synthetic sick genetic interactions (SSL) [18]. Another example is the combination of a GFP-tagged reporter gene with the deletion set [10]. This can be advantageous over the typical approach of using two deletions, as it expands the detection of network relations that go beyond lethality.

We took advantage of the SGA method and produced two different reporter deletion sets to investigate the dynamics of two different important networks. First, we were interested in different mechanisms of DNA damage repair. Therefore, we crossed a diverse set of reporter genes that we found to respond strongly to UV irradiation with a set of deletion mutants known to be affected in DNA damage response and RNA degradation [14].

A second screen used the well-studied galactose (GAL) network to address the mechanics of transcriptional activation during nutrient induced changes. Previous studies found reinduction of GAL network genes to occur faster than during the initial induction, even after an interim glucose repression of several hours [19, 20]. This is a primary example of epigenetics, as the effect can even be seen in daughter cells born after the first induction. Yet the underlying mechanisms remain largely unclear. Therefore, we combined a Gal1p-GFP reporter with an extensive set of chromatin related deletions, to highlight genes that influence either GAL network induction in general, or specifically a reinduction due to influences on epigenetic mechanisms.

For all aforementioned screens, our microfluidic device allowed for an unprecedented holistic approach. We achieved high temporal resolution on a genome wide level, linking pheno-

typic information about growth and cell size with cell network information about protein abundance and localization. As a result, it became essential to develop new methods to quantitatively analyze and visualize the large spectrum of information.

Chapter 2 will first describe general principles of microfluidics and the development of our device, mainly carried out by Nicolas Dénervaud. Chapter 3 describes the computational methods that are fundamental for single cell image analysis, namely image processing, single cell segmentation and information extraction. For our device, Ricard Delgado-Gonzalo and Nicolas Dénervaud have primarily conducted the first two of these steps. Nicolas Dénervaud and Johannes Becker worked together on the information extraction, with Nicolas Dénervaud focusing more on the question of protein concentration and Johannes Becker investigating the analysis of protein localization and on-chip cell growth.

The possibilities of genome-wide manipulations in a eukaryote make *S. cerevisiae* an excellent model organism for systems biology. Paired with robust growth on chip, it allows for high-throughput studies of complete gene or protein networks. Chapter 4 describes two large-scale screens of GFP reporter deletion double mutants. One screen focusing on the mechanisms of DNA damaging network and the other on transcriptional mechanisms during galactose induction, the chapter highlights ways of how the dynamic information of our microfluidic platform can integrate diverse quantitative information in a holistic manner.

This overlying thematic of quantitative analysis of high throughput microscopy experiments will be summarized in chapter 5. The chapter provides an overview of the advancements and possibilities of our platform. Chapter 5 also points out limitations, possible improvements and future applications for system-wide live cell imaging studies on our microfluidic platform.

2 Background: A microfluidic live-cell imaging platform

2.1 Introduction to microfluidics

The advantages of microfluidics are manifold. Just like its analogy in microelectronics, its reduced size allows for massive parallelization [21], while reducing costs due to reduced need of reagents. Furthermore, the use of small amounts of samples allows for controlled observations on single molecule level [22].

Several technologies play an important part in the development of a microfluidic device. Photolitography allows the miniaturization of components such as transistors and has been used in microelectronics since the early 1960s. First microfluidic devices were made of silicon or glass, copying their microelectronic counterparts. A further advancement was the use of elastomers like polydimethylsiloxane (PDMS). PDMS is transparent and gas permeable, making it well suited for experiments with optical read-outs and the culturing of cells. PDMS devices are developed by a technique called soft lithography [23].

First, lithography is used to fabricate a mold containing all the microstructures. Second, PDMS is casted upon this mold. This technique allows manufacturing numerous devices using the same mold. Multiple layers of PDMS can be used to add layers of control using microfluidic valves [24]. Using pressure, it becomes possible to deform the elastomeric membrane. This allows for controlled operations, like opening and closing of flow channels. Therefore, complex changes to attributes of the flow become possible, like the change of media sources or rapid mixing of multiple samples [22].

2.1.1 Microfluidic applications in single-cell imaging

It became obvious over the last decades that batch measurements are not sufficient to understand cellular behaviors. Looking at cells on a population level, hides the fact that there can be an extensive cell-to cell-variability on gene expression levels [25, 7]. Another example is the transcription of genes with bursting kinetics [26]. Time-lapse microscopy became the leading method to study single-cells *in vivo*.
To allow for single cell studies, it is important to keep the cells in monolayers and with the advancement of cloning techniques and the accompanying increase in strains of interest, it became mandatory to find possibilities of parallelization. Recently, the use of micro-well plates allowed for the parallelization of imaging, making it possible to image the whole yeast GFP library under different steady state conditions [10, 11]. This method has the drawback of only allowing for static images, as continuous perfusion cannot be achieved in a well. Microfluidic devices on the other hand have shown to be well adapted for the precise control of conditions [27, 28].

2.1.2 High-throughput imaging devices and live-cell arrays for device loading

The traditional technique for loading cells into a microfluidic device is flow-in, where strains are flown into the device and then trapped [29]. While these experiments can allow loading several strains and measuring several conditions simultaneously [13], the architecture of the approach is still limited to a low number of strains. A DNA spotter can precisely deposit small sample sizes, individually selecting them from strains that are stored in micro-well plates.
Previous approaches used this parallelization technique to analyze DNA samples on a microfluidic chip [30] or to spot cells on a coverslip for direct assessment upon nutrient starvation [31]. Of course, spotting live-cells onto a coverslip and then subsequently integrating this into a microfluidic chip adds additional challenges, as it requires the cells to regrow in their new environment. Recently Dénervaud *et al.* succeeded this challenge and engineered a device that allows for parallel continuous growth and observation of more than thousand different microbial strains under changing conditions [14].

2.2 The technical platform fabricated by Dénervaud *et al.*

As described in details here [14, 32], Dénervaud *et al.* set their objective to measure more than 1000 strains in a way that '(i) each micro-culture must grow continuously in a defined area and in a controlled environment, (ii) cells have to be constrained in a monolayer to enable single-cell imaging, (iii) the entire chip should be interrogated under a microscope with adequate spatial and temporal resolution, to enable the analysis of protein dynamics'.
The novelty of this approach lies in two different areas. First, it was necessary to design a microfluidic chip whose dimensions and geometry allow to grow cells continuously in monolayer, while at the same time minimizing effects like cross-contamination. Second, the spotting process of cells in a live cell array needed to be unobtrusive enough to allow exponential growth of the cells on chip upon perfusion after spotting and aligning to the microfluidic device
The optimization of chip design and cell spotting and the consequential alignment of the spotted cells to the PDMS device were iterative. In the following, we will shortly describe the final setting and highlight those parameters that are essential for the success of experiments. We will not describe the steps necessary for mold and microfluidic chip fabrication, two methods that have been described in detail previously [33, 34].

2.2.1 Chip design

A summary of the used chip design can be seen in Figure 2.1. The chip consists of two separate layers, the upper thick layer containing the control valves and the lower thin layer containing the medium flow channels, chambers and sieve channels. Flow channels, chambers and sieve channels have different heights according to their objectives (Figure 2.1d). Chambers are perfused from both sides. The use of two flow lines for perfusion prevents nutrient limitations inside the chambers. Sieve channels on one side prevent the cells from being washed out of the chambers.
Chambers were arranged in pairs, enabling the imaging of two different strains at the same time (Figure 2.1b). The flow channels are separated into 3 groups of 8 rows, a setup that was necessary to reduce pressure, whilst keeping a sufficiently high flow rate. Each group of flow

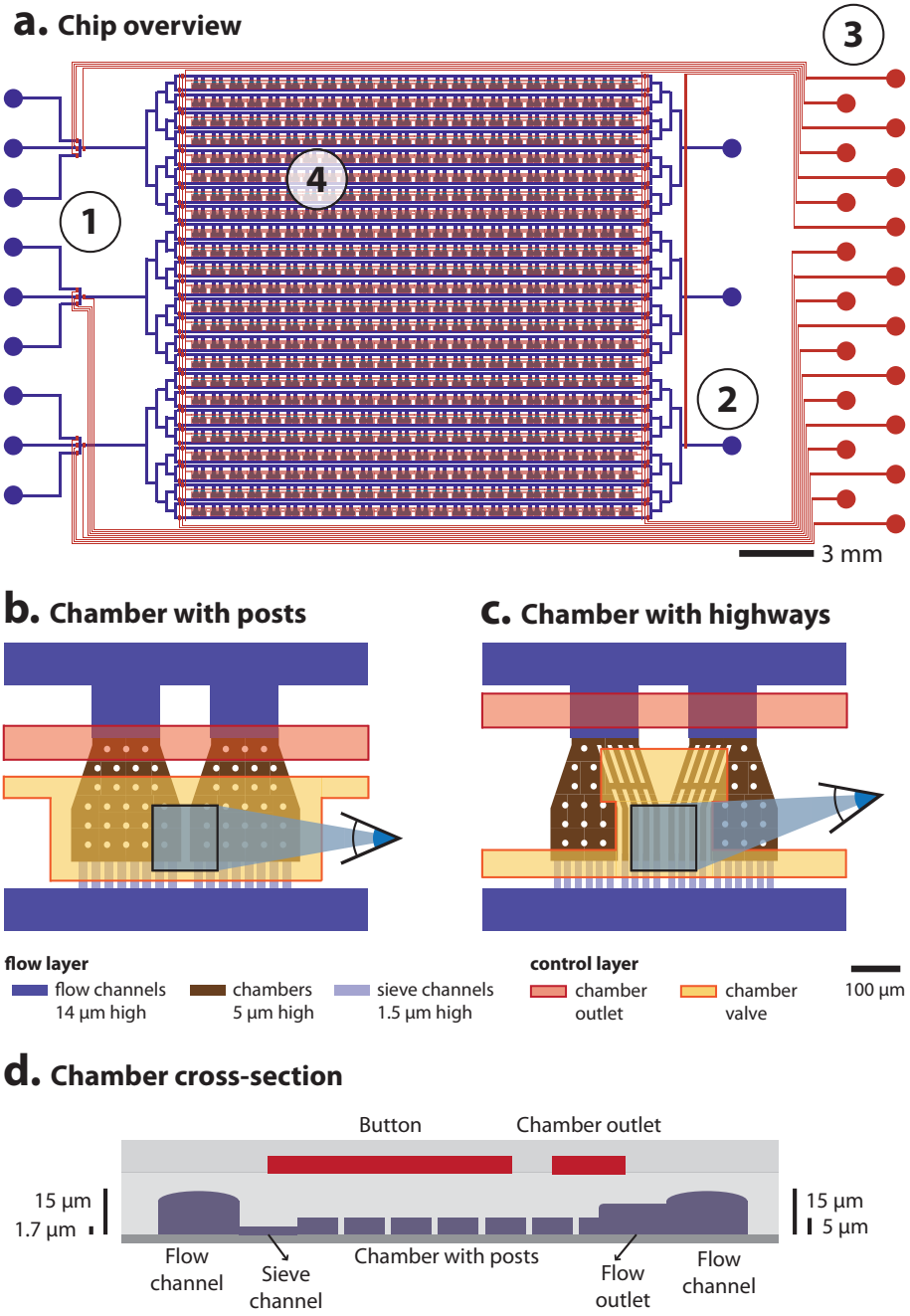

Figure 2.1: **Description of the perfused chamber array. a.** Schematic of the chemostat array with flow and control layers in blue and red, respectively. Components of the device are indicated: (1) three separated medium inlets and their control valves, (2) medium outlets and control valves, (3) connection of the control lines, and (4) chamber array. **b.** Scaled drawing of a unit cell pair, for the principal perfusion design. **c.** Scaled drawing of a unit cell pair, for the alternative perfusion design. The eye shows the imaging area. **d.** Schematic of a chamber cross-section.

channels has 3 inlets, allowing for the use of two different media in combination with one purge. Posts inside the chamber assure a uniform chamber height and prevent its collapse. Control layers consists of valves that can either close flow layers, push a button on top of the chamber or close the chamber outlet. Closing of flow layers is either necessary to control the medium source on the inlet or to facilitate the device perfusion. The button on top of the chamber is necessary to gently push the cells into a monolayer. Closing the chamber outlet button can reduce cross-contamination during initial growth.
A second design included channels inside of the chamber, called "highways". This device could be used for cell-tracking and allowed us to estimate the growth rate on chip (see section 3.6.1).

2.2.2 Live cell arrays

To successfully dispense an array of yeast strains using a DNA spotter, it is necessary to provide optimal conditions for the cell from spotting until final chip perfusion. We found it favorable to spot cells that are grown to stationary phase, as they were found to have a higher chance of survival. To prevent cell colonies from drying out, humidity levels inside the spotter were kept high (73%). It was also found that larger colonies increased the likelihood of cell survival, an effect that could be due to increased probability of a surviving clone. In addition, bigger spots remain moist for a longer period, something that could keep cells within the spot from drying out. We selected a pin with a delivery volume of 0.9 nL, a size that led to high cell density without overloading the chambers.
Essential for cell survival during the spotting process was to keep the time between spotting and perfusion as short as possible. We used 4 pins for spotting, spaced in a 2x2 square. This highly reduced the time needed for spotting. To assure cell survival, subsequent alignment of the PDMS device to the coverslip, followed by priming and installation of the device on the microscope stage should be conducted in approximately 30 minutes. Figure 2.2 gives an overview over the different aspects of spotting and alignment.

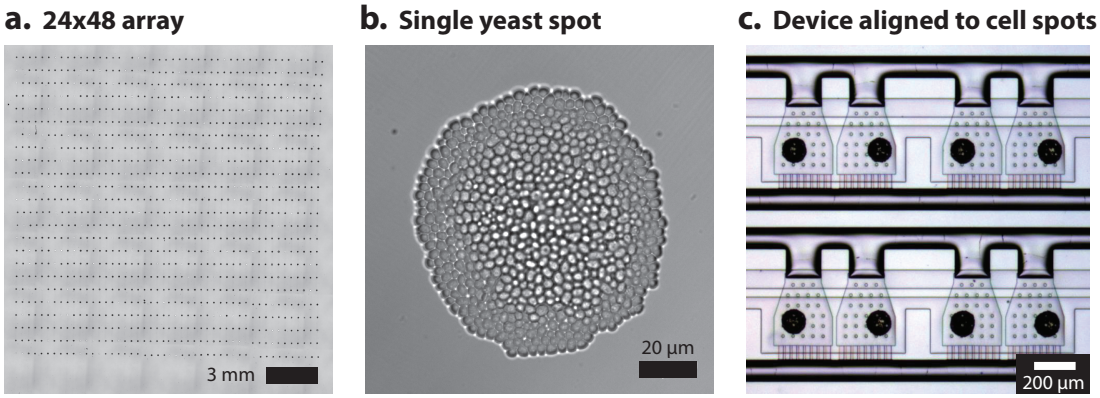

Figure 2.2: **Cell arraying results. a.** Assembly of 7 by 7 images covering the full 24x48 spots of a yeast cell array. **b.** High resolution micrograph of a single cell spot containing hundreds of cells. **c.** Brightfield micrograph of 8 chambers taken after cell spotting and chip alignment.

2.2.3 Live cell arrays

Several steps are necessary for effective and successful priming. First, it was necessary to perfuse the chip at a very low pressure level (1.3 psi). When medium reached the chip outlet, the outlet valve was closed and the remaining air in the device was eliminated by out-gas priming through the PDMS. To minimize the duration of out-gas priming, a step that can be very time consuming under low pressure, flow control valves needed to be opened and closed in a certain order, described in figure 2.3.

After the chip is completely primed, strains were grown for 16 to 20 hours to fully populate the chamber. During that time the entire device was imaged in 30-minute intervals, using 4x magnification and the NIS-element software (Nikon Instruments Inc.).

2.2.4 Overview of automated microscopy

To allow for the dynamic observation of subcellular events in more than 1000 strains simultaneously, it was necessary to optimize both spatial and temporal resolution. In the following, we will briefly describe the used microscope technology and setup.

An epi-fluorescence microscope with hardware autofocus system was used to quickly travel through each imaging position of the device. The chip was immobilized on the stage. Images were obtained with a 60x oil immersion objective. Besides those experiments that focused on

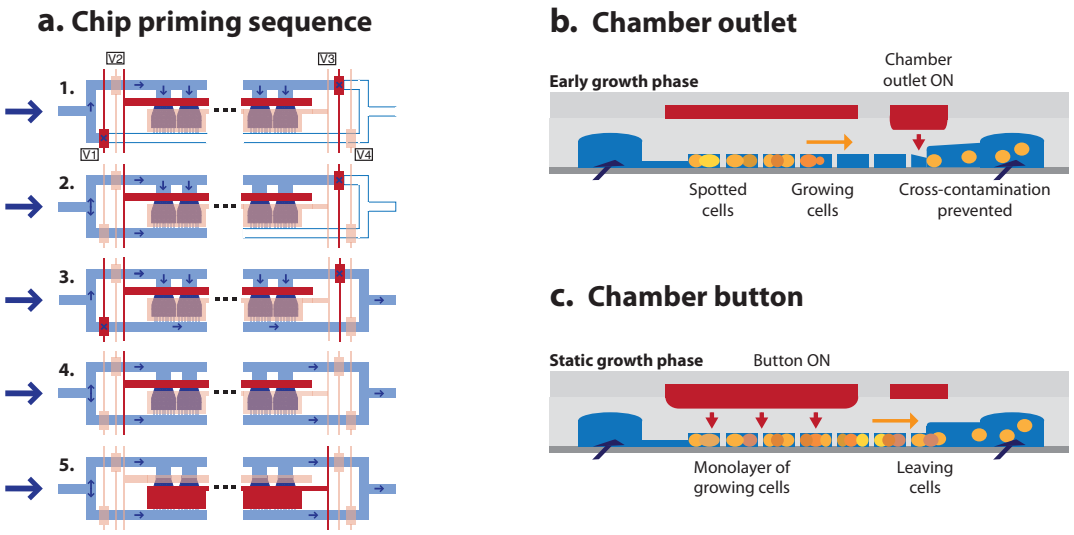

Figure 2.3: **Chip priming. a.** Schematic showing the sequence of valve operations needed to prime the chip. **b.** Schematic of the chamber cross-section, showing the action of the chamber outlet valve on the PDMS membrane. The valve partially closes the chamber outlet to prevent cross-contamination between chambers. **c.** Schematic of the chamber cross-section showing the impact of the button on the topology of the chambers. The button counteracts the pressure exerted by the growing cells to constrain them to grow in a monolayer.

protein abundance, an intermediate 1.5x lens was used to obtain a final magnification of 90x. LEDs provided a stable excitation source, which is crucial for imaging over a long period of time. An Electron Multiplying Charge Coupled Device (EMCCD) camera can obtain maximal sensitivity with minimal exposure time, therefore reducing the necessary amount of time while also preventing bleaching and photo toxicity.

A Visual Basic program controlled the microscope and its peripherals, including the control valves. Using eight reference points that were supplied manually, the position of each double chamber could be calculated with simple trigonometry. In the end, the software defines an optimal path that serpentines through each position, acquiring time-lapse movies on two or more different light channels (e.g. phase contrast and fluorescence).

3 Image analysis

3.1 Background: Image analysis

In most general terms, the object of image analysis is to automatically extract information from images. There are several advantages for automation [35]. It can allow for a fast and complete evaluation of a data set and can quantitatively detect changes that are subtle or not perceptive for a human observer. Furthermore, it can overcome observer bias, which is essential for an objective data evaluation.

One application for automated image analysis in biology is single-cell analysis. To extract information on single-cell level, images need to be first processed and afterwards segmented into single cells. In information extraction for fluorescent-tagged proteins for example, it is of main interest to measure the amount and localization of these proteins. There are several existing open-source image analysis pipelines, for example cell-ID [36] and Cell profiler [37]. Even though these pipelines are versatile and can be adapted to different problems, they were not suited for our microfluidic device. One reason is the crowded cell population structure, which is a hindrance for segmentation. Another reason is the high amount of images and cells. The microfluidic platform is capable of imaging more than 1.000 chambers 72 times per day, each chamber containing hundreds of cells. Therefore, a fast algorithm was mandatory.

To overcome these difficulties, we build a fast and stable fully integrated pipeline [32]. In the following, we will describe the different steps in general and how they have been implemented in our work. Further information can be found in the publication of Denervaud *et al.* [14].

General assembly of different parts of the pipeline, its implementation on a computer cluster, and the analysis of abundance data was performed by Nicolas Dénervaud. Cell segmentation has been the main work of Ricard Delgado Gonzalo. The automatic classification of subcellular localizations was the task of Johannes Becker.

One of the remaining shortcomings of automated image analysis is qualitative assessment, something a human observer is well capable of [35]. For our microfluidics platform, this becomes the most noticeable for the automatic classification of subcellular localizations, especially as an observation that changes over time. Therefore, a main focus during the following sections will be on protein localization and localization change.

The use of GFP as a reporter is generally found to be not influential on the phenotype [5]. This is not the case for gene deletions, where a change in size or growth can be a known consequence [9]. Therefore, we showed a great interest in size and growth of our reporter-deletion strains. While the size of a cell is a directly deductible result of cell segmentation, the estimation of growth in a densely populated environment is a non-trivial task and is described in section 3.6.

3.2 The automated image analysis pipeline

3.2.1 Image processing

The task of image processing is not to interpret the image content, but to transform it to emphasize particular aspects [35]. Processing steps can be contrast or color enhancements, or noise reducing steps like filtering. As our microfluidic device images two chambers at the same time, it was necessary to separate these two chambers into independent image sequences. In addition, small positional or rotational drifts needed to be corrected. As a result, we received cropped images that separate our chambers and are rotated for positional adjustment. This would also facilitate cell tracking, as it assures a stable relative position. The sequences were saved with a 14bit depth. Further step-specific image processing was done for each part of the image analysis separately.

3.2.2 Single-cell segmentation

The general objective of cell segmentation is to find all cells in an image and return their outlines. Even though this can be a simple (yet daunting) task for a human observer, its computational implementation is often not straightforward [38, 39]. The reasons for occurring problems are manifold and often implementation specific.

For example, first segmentation techniques used intensity thresholds, assuming that cells have a detectable different intensity than the background. However, this assumption often does not hold true, as both cell and background intensity can change gradually. Another technique, watershed based segmentation, imagines an image as a contour and fills darker parts of the image (basins) with water, until neighboring basins start to touch. While watershed based segmentations can have problems that are as well threshold related, their shortcomings are usually slightly different, as they are not subject to absolute thresholds. An additional problem is that cell contours of neighboring cells may not be well defined, something of importance for our densely packed microfluidic device.

To overcome these limitations, we combined watershed-based segmentation with an approach of deformable model fitting, summarized in figure 3.1. Deformable model fitting uses a parametric contour to minimize an energy function. In our case, it was implemented using an ImageJ plugin [40]. First, pre-processing was performed, using morphological dilation and a smoothing filter to reduce noise. Then, a watershed algorithm split images into small areas, each area expected to contain a single cell. Finally, these areas were used as initiation point, using parametric active contours or snakes. These snakes maximize the intensity difference between the dark inside of a cell and the usually brighter halo or gray background surrounding a cell. A first minimalistic snake named the Ovuscule [41] was used for a first estimate. This snake is parameterized by three control points and has the shape of an ellipse. In a second step, a snake with a variable number of control points, named the E-snake, refined its result. While the Ovoscule and the E-snake themselves are robust segmentation tools, they both converge to a local minima. Therefore, the validity of their results heavily depends on the quality of the image and subsequently the watershed segmentation. To assure that only well segmented cells are further analyzed, we used single-cell filtering based on different criteria. We removed all cells that were close to the boundary, as those cells were generally not well

segmented. To remove wrongly segmented objects like chamber posts or small artifacts, a cell size threshold was implemented.

Cell segmentation quality was classified using a manually annotated set of good and bad quality cells for training with a support vector machine (SVM). Comparing automated segmentation with manual annotation, we correctly identified 83.7% of the cells, with a specificity of 92.3%. As a final step, cells that were strong outlier in abundance were discarded as well. Outliers were defined as cells that were beyond the outer fences (using interquartile outlier detection criteria), or more than 3 standard deviations away from the mean.

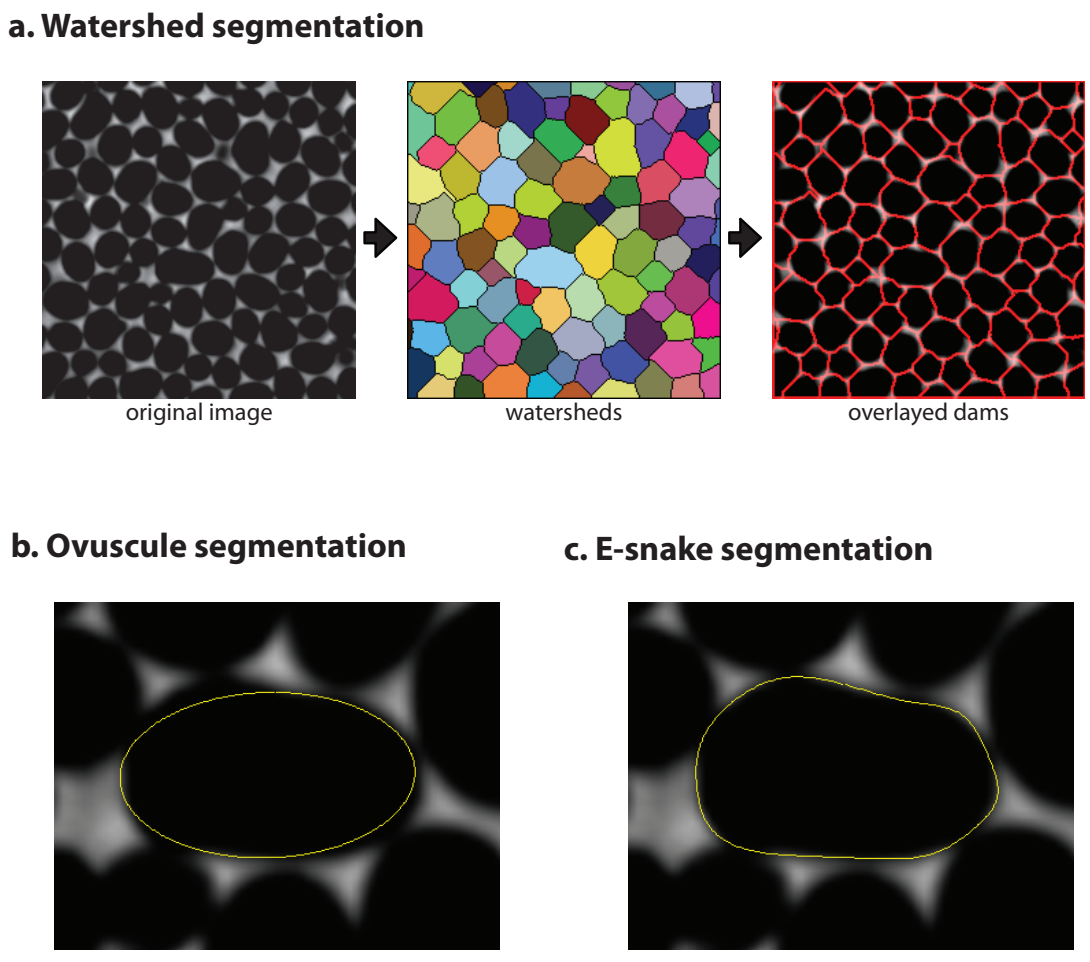

Figure 3.1: **Cell segmentation - Watershed, Ovuscule, E-snake. a.** Process flow of watershed segmentation using phase contrast images. **b.** Example of a cell contour determined by an ovuscule (a snake with three nodes) and **c.** by an E-snake (with unlimited number of nodes).

3.2.3 Abundance extraction and estimation of protein copy numbers

The estimation of protein copy numbers is of central interest in fluorescence microscopy. As the abundance of neighboring cells influences the general intensity, it is necessary to estimate the local background intensity. To estimate this intensity, we used the minimal pixel value of the watershed region. Even in the absence of GFP, cells have fluorescence values above background due to auto-fluorescence. To measure the distribution of a cell populations auto-fluorescence, we imaged the GFP library parental strain (BY4741) under the same conditions as the GFP strains.

This distribution can be used for deconvolution, a probabilistic approach that assumes that the abundance of a cell is the sum of the cell auto-fluorescence and the contribution of GFP. Assuming a log normal distribution for the auto-florescence and a gamma distribution for GFP in cells, we obtain the following formula

$$P(s,k,\theta) = \int_0^s P_{gamma}(x,k,\theta) \ln \mathcal{N}(s-x,\mu,\sigma)) dx \tag{3.1}$$

The log-likelihood of this probability for the set of measured cells $\{s_i\}$ is:

$$LogL(k,\theta) = \sum_i \ln P(s_i,k,\theta) \tag{3.2}$$

We maximized the log likelihood using a Markov chain Monte Carlo (MCMC) method. The thereby obtained estimations of actual GFP distributions were then compared with other techniques. We compared the data from 18 proteins, measured by both TAP-western and mass spectrometry [6, 7], and found a correspondence with our measurements, which allowed us to estimate absolute protein counts (Figure 3.2).

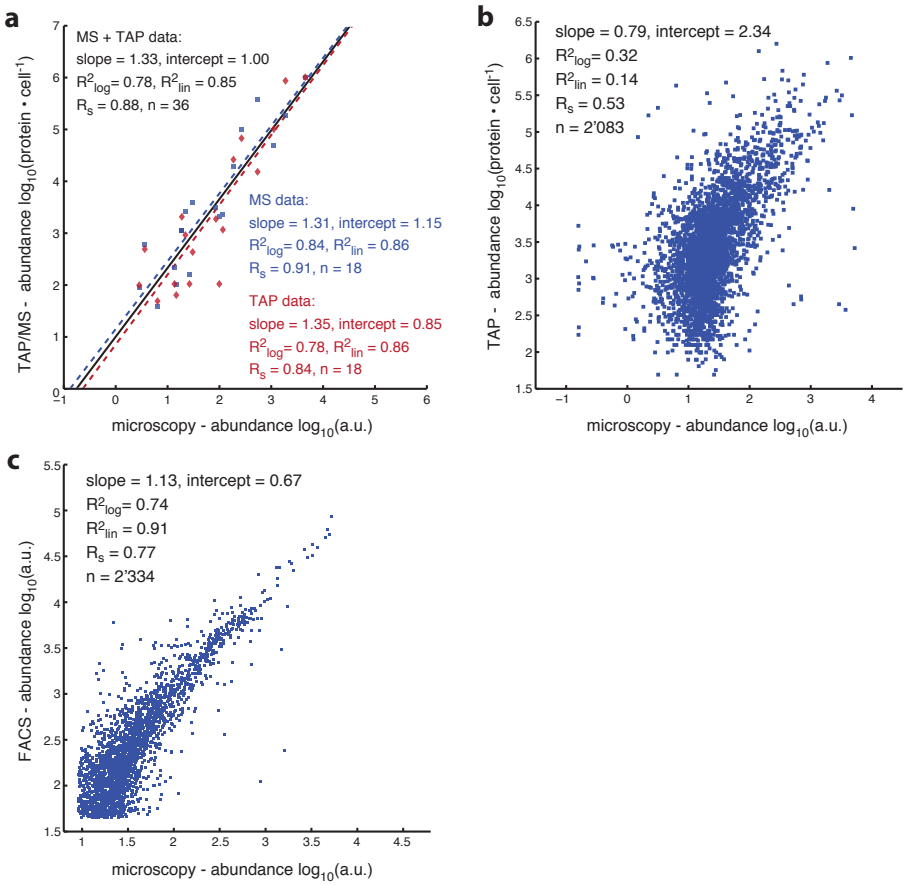

Figure 3.2: **Comparison of protein abundance measurements in our studies with existing datasets. a.** Correlation of our measurements with mass-spectrometry (blue) [42] and TAP-western data (red) [6]. The black line shows the correlation of our measurements with the mass-spectrometry and TAP-western datasets. This relationship was used to infer absolute protein abundance from our data. **b.** Correlation of abundance values from our data with TAP-western data [6]. **c.** Correlation of abundance values from our data with flow cytometry data [7].

3.3 Classification of subcellular localization

3.3.1 Background: Automated classification of subcellular localization

The principles of automatic localization of subcellular protein patterns remain mostly the same since their first applications at the end of the Nineties [43]. In general, after a first imaging process to normalize the images and remove background noise, image analysis methods are used to calculate numerical features, which contain valuable information such as the morphology of a cell [44]. A classifier to categorize cells into different categories then evaluates this information.

There are several automated classifiers with a good predictive value recognizing the patterns of subcellular structures in HeLa cells [44, 45] and on the yeast ORF-GFP fusion library [46, 47]. But in contrast to the image data given by the micro-fluidics platform, these studies had circumstances that facilitated the classification. In the case of HeLa cells, it is simply the fact that HeLa cells with a diameter of 25μm are more than five times larger then the diameter of a typical haploid yeast cell (~4μm). In the case of the ORF-GFP fusion library, a DAPI image to determine the position of the nucleus and mitochondria and a differential image contrast (DIC) image to determine localization at the vacuole and vacuolar membrane where available [5]. Furthermore, strains were grown in low density, which facilitates a precise segmentation. Over the last decade, several studies used the original yeast GFP library data to enhance the detection quality and computational efficiency of subcellular localization [46, 47, 48]. For example, latest efforts in the lab of Robert Murphy classified 2,655 images from the UCSF collection with an overall accuracy of 87.8% in less than one hour [47]. But these studies are accompanied by several caveats that make a direct adaptation to our task impossible.

The main setback is that all studies focus on the classification of populations that were attributed to a unique subcellular localization. As we are mostly interested into strains that exhibit localization changes, our main interest is on those strains that are steadily or transiently found in two or more subcellular compartments. The automatic analysis of heterogeneous localization distributions is still scarcely discussed. Studies that were interested in the dynamic changes of localization were generally focused on one protein and a well-defined change [25], which considerably reduces the complexity.

Recent studies were interested in analyzing these mixed distributions and even extended the analysis to cases were fluorescence in one cell can be localized to several compartments [49]. But these studies focused on only two subcellular compartments, lysosome and mytochondria, using osteocarsoma cell lines, which have an approximately 10-fold bigger radius than yeast.
A second difference between previous approaches and our requirements is of a more technical nature. While accuracies of around 90% can be reached for classification [47], it is at least partially the result of a bias due to the chosen classifier. Support vector machines (SVM), the predominantly chosen type of classifier, are constructed in a way that favors groups with more data points. This helps to achieve a better overall accuracy, but leads to a unidirectional misclassification between different classes.
A slightly different approach to describe a large amount of different strains is the use of unsupervised clustering. While this approach starts as well with the measurement of features, it does not need any further information about different classes. Instead, it defines a method for linking the data and calculates a distance metric. It subsequently sorts the data set into clusters in regards to the used method and metric. A recent study used this approach for *S. cerevisiae* to describe general protein distributions during different parts of the cell cycle [48]. While this method is well suited to give a general overview over the subcellular localization distribution of the proteome, its drawback is that there is no certainty information for one specific reporter strain.
Overall, the unique questions that our dynamical data set poses have not been answered satisfactorily. The most general question that subcellular classification tries to solve is the localization of a protein inside a single cell at any given time point. As we laid out previously, this holistic approach is hardly achievable. We therefore posed different questions that we tried to answer separately.
Our main objective was to quantify subcellular protein localization and especially localization changes. Thus, our goal was to build a robust classifier for single cell localization to answer questions about the general localization of a protein. To achieve this robustness, we addressed the underlying issue that subcellular localizations are even for a trained observer often impossible to distinguish. To overcome this obstacle, we focused on geometrical pattern instead of subcellular localizations. We used the terms Periphery, Structure, Punctate, Disk, Corona, and Homogeneous to describe pattern that were computationally as well as manually

separable. We were also able to show that geometrical pattern information can be traced back to subcellular compartment information.

In the following, we will first describe our localization pattern classifier, which was used in recent studies [14, 50]. Nicolas Dénervaud and Johannes Becker carried out feature extraction and Johannes Becker developed and evaluated the classifier. Even though our classifier was found to be robust, it was not suited to work without human supervision. We will explain the reasoning of this approach in section 3.3.6. Published material will demonstrate typical results extracted from our classifier in section 3.4, while section 3.5 will show extended possibilities of our classifier.

3.3.2 Feature extraction for protein localization

The object of feature extraction in image analysis is to condense the complex information contained in pixel intensities into a set of information that is comparable between different images. Feature extraction can be done either for single cells [46] or using complete images [47]. While the latter is less computationally expensive, it is not suited for single cell analysis. The selection of meaningful features is arguably the most important step of classification, as subsequent steps can only be successful if the extracted information is sufficient.

For each cell, we extracted a small rectangular image, surrounding the cell contour. Contrast was increased by stretching the image in the full 8-bits range (0-255), between a minimal value, defined as the 5th percentile of the cell pixels, and the maximal pixel value. Starting with a large set of more than 200 features, we used Analysis of Variance (ANOVA) to test their efficiency. If features were found to be redundant, we removed those that were computationally more expensive.

In the end, we calculated a set of 97 features for each cell. This set consists of 17 histogram-based, 3 geometrical and 7 morphological features [44, 51], 10 granulometry measures [52] and 60 threshold adjacency statistics [53]. The complete list of features can be found in Appendix A.

3.3.3 Supervised classification into six spatial patterns

Our experimental design implies a certain compromise between high-throughput temporal imaging and the level of details of subcellular localization analysis. Thus, we had to distinguish between very fine localization patterns as usually defined in cell biology and more objective geometrical shapes. Exploratory analysis indicated that we can robustly distinguish six spatial patterns, shown in figure 3.3. To train our classifier, we built a training set by manually annotating cells extracted from 104 images as representatives of one of the following patterns:

- **Periphery:** Represents a fine outline of the cell contour, generally well distinguishable. Representatives include cellular membrane proteins, or in some cases protein located in bright dots distributed on the cell contour.
- **Structure:** Includes filaments, circles and shape-forming dots that are often a direct indication for proteins localized in the endoplasmic reticulum (ER), the Golgi apparatus, or the mitochondrion.
- **Punctate:** Detects one or more distinct dots, smaller than 1 μm in diameter (< 20% of the size of a cell). Typical representatives are nulear foci, cytoplasmic aggregates, actin, lipid particles, endosomes and peroxisomes.
- **Disk:** Highlights one dominant area of GFP signal contained in the interior of the cell. The diameter of these objects is at least around 25% of the diameter of a cell. Typical representatives of this group are proteins localized in the nucleus and nucleolus, but also proteins in the vacuole or the vacuolar membrane.
- **Corona:** Includes pattern showing a broad ring around the center that can also be more sickle-shaped. Typical subcellular compartments that have a corona-like appearance are the cytoplasm and in some cases the ER.
- **Homogeneous:** Represents cells where the fluorescence is uniformly distributed. In many cases homogeneous cells are of low intensity, reflecting background levels.

As the boundaries between these shapes can be fuzzy, we chose to assign to each cell a probability vector reflecting the likelihood to belong to each of the six patterns. We first used

Reduced-Rank Linear Discriminant analysis (RRLDA) [54] to compress the dimensionality of the feature space. This method is similar to Principal Component Analysis (PCA), but considers the separation into classes as additional information. Instead of considering the directions in the parameter space with the largest variance in the data, RRLDA maximizes the between-class variance relative to the within-class variance. We used RRLDA together with our training set and obtained a matrix that reduces the dimensions of our feature set from 97 dimensions to five.

We verified that increasing the number of features does not improve performance. To assign probabilities to each cell, we used the MATLAB function 'classify', specifying a 'quadratic' discriminant function, which fits multivariate normal densities with a separate covariance estimate for each shape. In addition to the probability vector, the function 'classify' gives for each cell an estimate of the probability density of the feature set of this cell. This is useful for discarding cells that are atypical (e.g. dead cells) or have ambiguous fluorescence patterns.

For each image, we obtained a distribution of 6-dimensional vectors representing the cell population. To compare distributions, we modeled each population as a Dirichlet distribution, which is well-suited to model a population of probability vectors [55]. We used the Bhattacharyya distance to estimate the similarity of two Dirichlet distributions [56, 57]. Using the Dirichlet probability density function

$$p(\mathbf{p}) \sim \mathcal{D}(\alpha_1, \ldots, \alpha_d) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k p_k^{\alpha_k - 1}, \quad p_k > 0, \quad \sum_k p_k = 1 \tag{3.3}$$

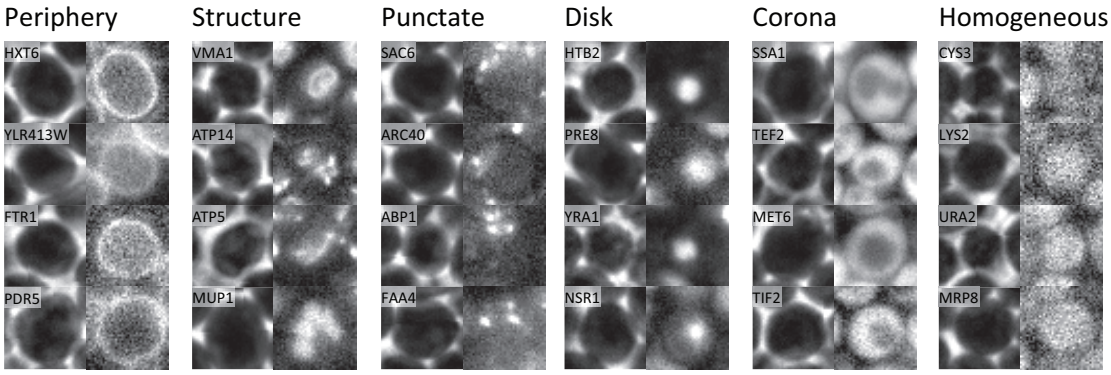

Figure 3.3: **Single cell representations of the six localization pattern.** Single-cell micrographs illustrating the six localization patterns. For each cell, the phase contrast channel is shown on the left, and the GFP channel is given on the right. Each image has a width of 5.8 μm.

we obtain for the probabilistic distance function between two parameter sets $\alpha_{\mathbf{a}}, \alpha_{\mathbf{b}}$ the Bhattacharyya distance J_B

$$\begin{aligned} J_B(\alpha_{\mathbf{a}}, \alpha_{\mathbf{b}}) = \ln\Gamma\left(\sum_{k=1}^{d}\frac{1}{2}(\alpha_{ak}+\alpha_{bk})\right) + \frac{1}{2}\left(\sum_{k=1}^{d}\ln\Gamma(\alpha_{ak}) + \sum_{k=1}^{d}\ln\Gamma(\alpha_{bk})\right) \\ - \sum_{k=1}^{d}\ln\Gamma(\frac{1}{2}(\alpha_{ak}+\alpha_{bk}) - \frac{1}{2}(\ln\Gamma(|\alpha_{\mathbf{a}}|) + \ln\Gamma(|\alpha_{\mathbf{b}}|)) \end{aligned} \tag{3.4}$$

3.3.4 Validation of the classifier

We first assessed the performance of the classifier using 10-fold cross-validation of our training set. Each cell was assigned to its most probable pattern. The confusion matrix indicated that, expectedly, most misclassifications happened between groups with the most fuzzy boundaries (Figure 3.4.a). For example, a low intensity cytoplasmic signal could be misclassified as being homogeneous. Or, mitochondrial proteins were classified either as punctate or structure, depending on the density of the signal. Note that assigning hard classes removed information and performance was thus expected to be lower in this assessment.

We also compared manual against automatic annotation. For this, we randomly picked 200 images for which more than 60% of the cells were classified to belong to the same group. Those images were independently and blindly annotated by ND and JB. For the 182 images where the two manual annotations agreed, there were only two cases of disagreement with the automatic annotation (Figure 3.4.b).

To assess the repeatability of the classification, we picked 2741 strains, for which we had duplicate recordings. We selected only those 1034 strains with high intensity to avoid the problem that two randomly selected strains that contain only background noise are as well very similar. Those selected strains have an average Bhattacharyya distance of 0.11 (0 means identical). In comparison, when selecting two of these strains at random, the distances were significantly higher (Figure 3.4.c), showing that our method is reproducible.

Finally, we wanted to validate the Bhattacharyya distance as a measure of localization change. We randomly picked a sample of 110 image sequences, from an experiment where a chemical stimulus was added about half-way through. To quantify potential localization changes due to the stimulus, we calculated the Bhattacharyya distance between the end and the beginning of

the experiment. Meanwhile, ND and JB manually annotated if they could perceive a change in the spatial distribution of fluorescence. For a Bhattacharyya distance of 0.4 or bigger, 75.6% of the sample was annotated as changing. On the contrary, for a Bhattacharyya distance smaller than 0.4, only 25% of the sample was annotated as changing (Figure 3.4.d).

Together with the finding that more than 90% of replicate experiments have a Bhattacharyya distance smaller than 0.2, we can say with high certainty that our method is reliable and that it can identify changes in the spatial distribution of the signal. However, whether those changes

a

Actual \ Predicted	Periphery	Structure	Punctate	Disk	Corona	Homogeneous
Periphery	92.6%	0.7%	0.7%	0.0%	3.0%	3.0%
Structure	1.2%	65.0%	10.9%	11.9%	7.1%	4.0%
Punctate	2.4%	12.7%	60.5%	7.5%	1.3%	15.7%
Disk	0.0%	25.7%	2.5%	67.4%	2.3%	2.2%
Corona	0.9%	11.4%	2.5%	0.7%	48.5%	36.1%
Homogeneous	0.0%	0.8%	1.7%	0.2%	9.0%	88.3%

b

Manual \ Computer	Periphery	Structure	Punctate	Disk	Corona	Homogeneous
Periphery	13	0	0	0	0	0
Structure	0	22	1	0	0	0
Punctate	0	0	4	0	0	1
Disk	0	0	0	50	0	0
Corona	0	0	0	0	69	0
Homogeneous	0	0	0	0	0	22

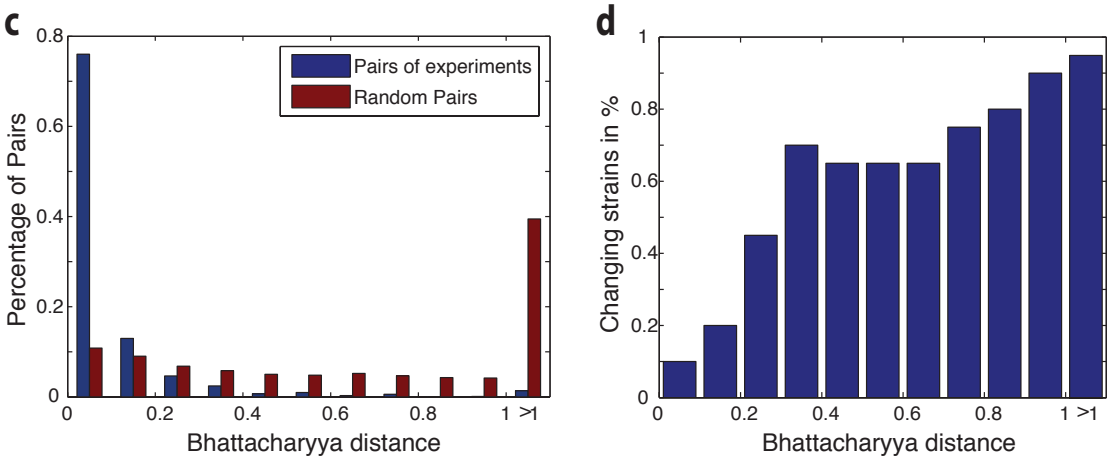

Figure 3.4: **Validation of the classifier. a.** Confusion matrix to compare the manual annotation of single cells with the predicted geometrical shape using 10-fold cross validation. **b.** Comparison of automatic and manual annotation: 200 images for which the classifier found one characteristic group, where manually and independently predicted by ND and JB. **c.** Agreement of replicated recordings: Histogram of the Bhattacharyya distance of high intensity strains with duplicate recordings, in comparison to the distance of randomly selected strains. **d.** Relation between Bhattacharyya distance and the visual perception of a change between the images.

are biologically relevant remains to be assessed by a critical viewer. For example, a change identified for a mitochondrial protein might result from a change of the total number and spatial distribution of mitochondria within the cell. But, the protein might remain in the mitochondria, and it is therefore not a localization change.

3.3.5 Comparison with the original yeast GFP library annotations

The yeast GFP collection was initially observed in static conditions and protein localization was manually assessed and determined using 22 biologically relevant annotations (UCSF annotations [5]). We wanted to find the correlation between our spatial patterns, determined by a continuous 6-dimensional space, and theses annotations. To simplify the comparison, we chose the strains with only one clear annotation in the UCSF dataset. We also concentrated only on the strains that showed a sufficient intensity in our recordings (above background). We grouped the strains based on their manual annotations and compared the average cell population probabilities (Figure 3.5.a) and the correlation of the Bhattacharyya distances within those groups (Figure 3.5.b).
We found that specific subcellular localizations have a distinct probability profile. For example, mitochondrial localization is defined by a mix of structure and punctate and nucleolus is a mix of disk and punctate. Nuclear periphery is well characterized by the structure pattern and vacuolar membrane is a mix of disk and structure. This shows that our continuous 6-dimensional space contains more information than six binary classes. A precise comparison of the population distribution allows to find subtle differences.

3.3.6 Supervised quantification of localization change

There are two main reasons why an automatic classification of localization changes is not advised, both related to the small number of strains with localization changes that we detected manually. We manually annotated ~120 of the over 4000 proteins (~3%) as localization changes. Even if we were able to achieve a very low false positive rate (the sum of all false positives divided by the sum of all negatives), we would still obtain a high false discovery rate. In addition, the low number of strains that were found to change their localization would make it even more important to keep the sensitivity of our automatic detection high.

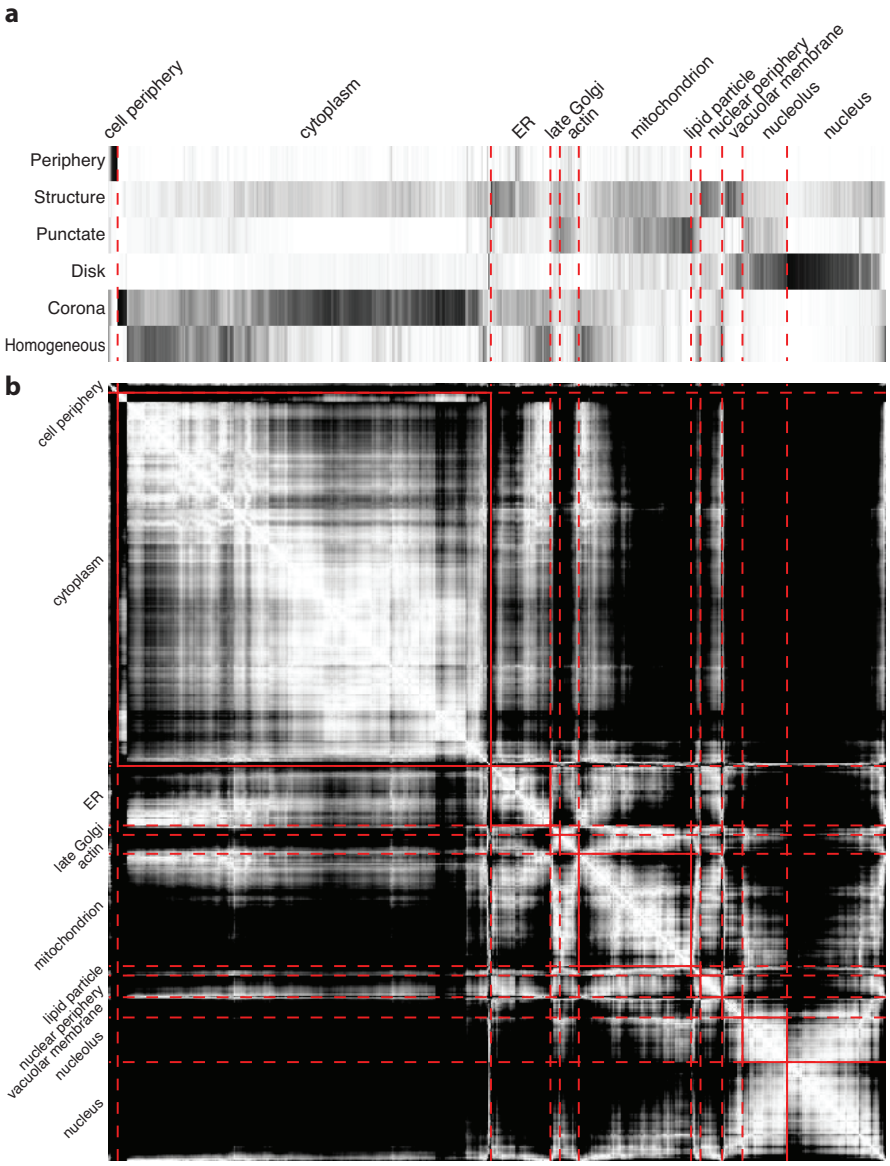

Figure 3.5: **Comparison of our six spatial patterns with the UCSF annotations.** Evaluation of strains with high intensity and well defined localization patterns. The strains are clustered within their groups, as defined by Huh *et al.*[5] . **a.** Average probability of the six geometrical classes for each strain. White means 0% probability, black 100%. **b.** Heat map that shows the Bhattacharyya distances between the strains. The distance goes from zero (white) to one or above (black).

Therefore, we decided to use our classifier in two ways. Firstly, we used hierarchical clustering to find strains with changing localization that were overlooked during manual annotation. By focusing on strains with a short distance to manually selected strains, we were able to detect eight additional strains showing similar changes. Three of these strains were found to be

previously missed strains of interest. As these were only 2.6% of all annotated changes, it was valid to assume that this list is exhaustive. Figure 3.6 is exemplary for the practical application of this method.

Secondly, we combined our manual annotation with our pattern classification to achieve a

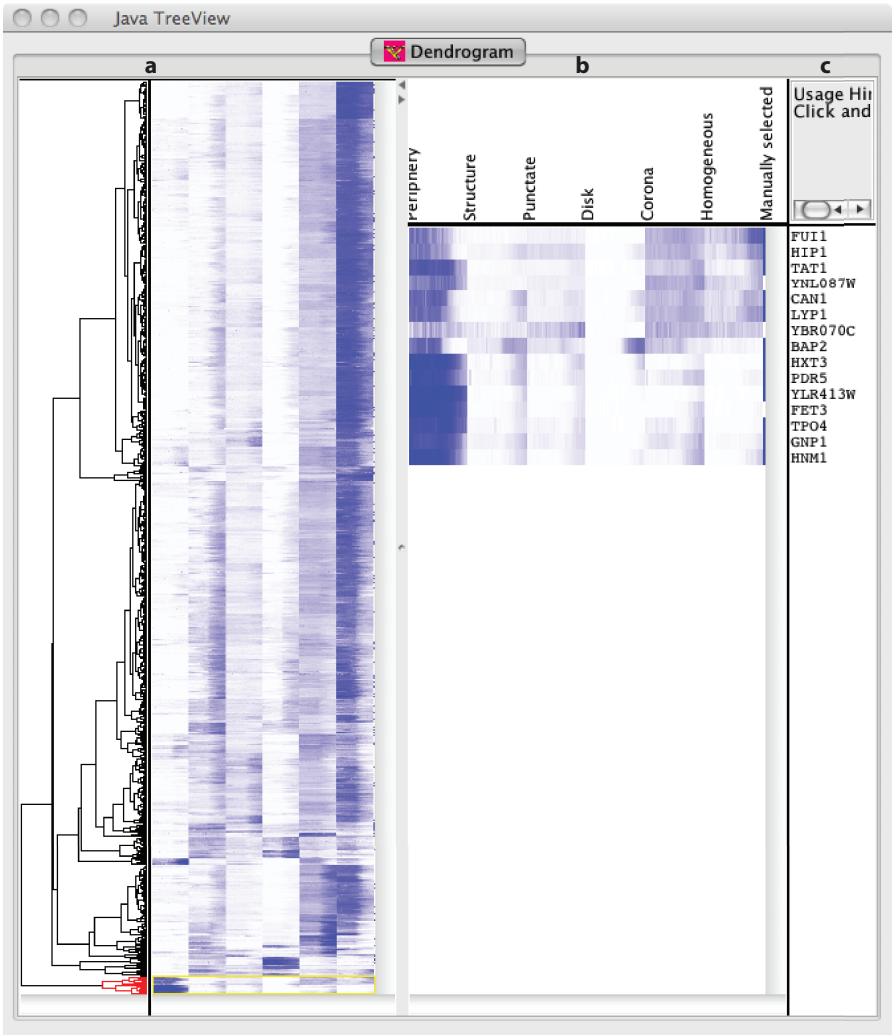

Figure 3.6: **Example of a clustergram of localization over time in Java TreeView.** (a) View of the complete clustergram, which contains only strains with a temporal Bhattacharyya distance > 0.3. (b) Focus on strains for which the classifier detected a localization change of proteins away from the cell periphery. Average probabilities indicate that most of these proteins relocate not directly after MMS treatment, but gradually after around 3 hours. The last column of the clustergram shows if strains where manually annotated as an interesting protein localization change (blue) or not (white). After evaluation of those 3 strains that where not manually annotated as changing, FET3 was added to the list of annotated localization changes. (c) Standard names of the proteins, providing a link to SGD.

supervised quantification of localization changes. To further quantify the localization changes and analyze their rate and timing, we focused on the geometrical pattern, which is the most relevant to a given transition (i.e. showing the clearest change). We found that these are (i) disk for transitions between nucleus and cytoplasm, (ii) punctate for proteins that aggregate and (iii) periphery for everything transiting from or towards the cell membrane.

For each of those cases we fitted a logistic function to approximate the average probability of the respective relevant pattern, P_t, by minimizing the error between L_t and P_t. L_t is given by:

$$L_t = \frac{\alpha}{1 + e^{(-\lambda(t-\delta))}} + \tilde{P}_{Pre} \tag{3.5}$$

where $\tilde{P}_{Pre}$ is the median value of P_t for the last two hours before the stress stimulus. If the localization change was transient, we fitted the logistic function only for the time until the change reached its peak. For pulse experiments, we focused our fit on the first pulse. α, δ and λ are estimations for the rate, the timing and the slope of a localization change. An example is shown in figure 3.8a.

To filter strains for which the change was not quantified robustly, we set a minimal threshold to the score T_P , given by the following equation:

$$T_P = \frac{L_{max} - L_{min}}{\sqrt{P_\sigma^2 + L_{max}}} \tag{3.6}$$

T_P takes into account the variance of the average probability of the pattern before treatment, P_σ, and the maximum and minimum values for the logistic fit, L_{max} and L_{min}. This way, we were able to sort out strains for which the automatic analysis could not detect a change with certainty.

In addition, we required that strains with at least double coverage showed a change in more than one repeat. For the second screen, we discarded strains with single coverage. The threshold of 0.12 was determined empirically. The requirement that all 119 strains in the first experiment show a localization change was used to optimize sensitivity. As a result, 111 of 119 strains (93.3%) passed this filter. Strains that we manually detected as not changing in the second set of experiments were used to maximize specificity. Of all 97 strains that were automatically detected during the four additional experiments of the second screen, we had

to discard only 3 strains (3.1%) as misclassification errors. Thus the choice of our threshold was adequate.

To make different stresses comparable, we normalized the data of our relevant pattern by a modified z-score, Z_t, that is closely related to T_P :

$$Z_t = \text{sgn}_{man} \frac{P_t \tilde{P}_{Pre}}{\sqrt{P_\sigma^2 + max(L_{max}, P_t)}} \tag{3.7}$$

sgn_{man} sets the directionality of the function after our annotated localization change (e.g a protein moving from the nucleus to the cytoplasm has a positive change, although the disk probability decreases). The observed localization changes in MMS_{high} are given in figure 3.8b.

3.4 Results of quantitative localization analysis

3.4.1 Screening of the GFP library in MMS

Studying proteome-wide changes upon MMS treatment, we found 111 proteins that change their localization. Based on the involved localizations, changes could be assigned to one of five transition classes: transitions between cytoplasm and nucleus (28 proteins), transitions from the nucleus to nuclear foci (11 proteins), nuclear periphery aggregations (21 proteins), formation or dissolution of cytoplasmic foci (33 proteins), and transitions from the cell periphery into the cell interior (18 proteins) figure 3.8b.

Timing of localization changes was found to differ strongly in timing and intensity. For example, the formation of protein foci in two heat shock proteins (Hsp104 and Hsp42) occurred in less than an hour. Further rapid transitions were the relocation of Bmh1p from the cytoplasm to the nucleus and Rnr4p from the nucleus to the cytoplasm. Bmh1p and Rnr4p are known to be involved in DNA binding and DNA damage repair respectively. These fast responses can stay in stark contrast to the more gradually changes found in abundance. This discrepancy is exemplified in figure 3.7.

Our results agree with previous studies that showed localization changes to be a fast and efficient way of a cell to respond to a change [8, 58, 59]. The gradual increase in abundance on the other hand, where protein levels can increase by more than 100 fold, allows for a wide

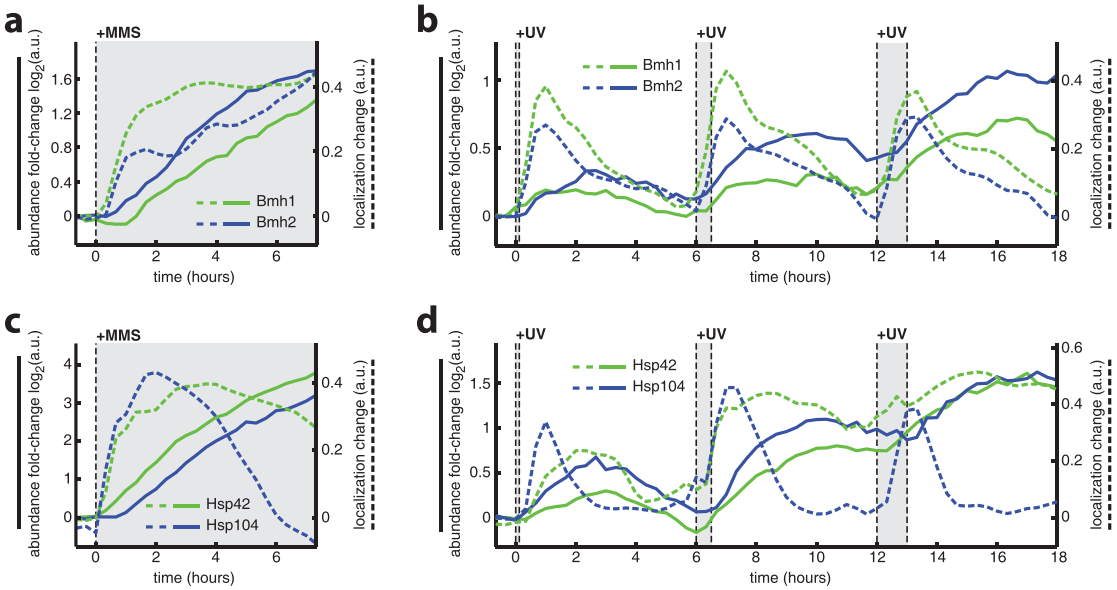

Figure 3.7: **Dynamics of Bmh1p/Bmh2p and Hsp42p/Hsp104p.** Abundance (solid line) and localization (dashed line) changes of Bmh1p and Bmh2p in response to **(a)** MMS treatment and **(b)** UV pulses. Abundance and localization change of Hsp42p and Hsp104p in response to **(c)** MMS treatment and **(d)** UV pulses.

range of well-adapted responses.

3.4.2 Comparison of localization changes for different stress conditions

We extended our observation of those proteins that showed localization changes in MMS to five additional conditions (MMS pulses, lower MMS concentration, MMS pulses, HU, UV pulses and Sorbitol). We found localization changes to be condition and protein function specific (3.8c-f). MMS, HU and UV, three stresses that lead to severe cell damage, were similar in that two Heat Shock Proteins (Hsp104p and Hsp42p), the DNA related Bmh1/2p complex and Rnr4p relocalized during the first hour. The Mcm2-7p proteins changed their localization in a similar fashion in all 3 of those stresses, showing that cell cycle arrest occurs precisely and in a stress unrelated manner. Proteins of the nuclear pore complex were found to form foci after 3 to 5 hours in both HU and MMS. They were not as prevalent during the transient stress of the first UV pulse.

The most striking differences between different conditions can be found for proteins related to mRNA processing and membrane transportation. P-Body related proteins, which are known to

have a function in mRNA degradation [60], were found to change their localization after more than 4 hours in MMS, between one and four hours in HU, in less than two hours in sorbitol and after less than 15 minutes in UV irradiation. A previous study that looked at localization changes after two hours, found P-Body formation in HU, but not in MMS [10]. Another study tested the P-Body protein Dhh1p in *Candida albicans* under different conditions, finding that P-Body formation was the highest for UV irradiation and osmotic stress [61], which can be induced by sorbitol. These findings confirm our results, while showing the increase of information that can be achieved by our temporal resolution.

Localization changes of membrane transporter were found to occur between one and 6 hours in MMS, in less than three hours in HU and practically immediately for sorbitol. As may be expected, they did not occur in UV irradiation. Under all three conditions, they tend to slightly precede the formation of P-Bodies. Proteins localized to the cell membrane were additionally found to localize to the vacuole [5], whose role in autophagy and protein degradation is well known [62]. We therefore hypothesize that P-Body formation under certain conditions may precede a first step of degradation inside the vacuole. This possibility of a relation between P-Bodies and vacuolar degradation is mentioned in the literature [63, 64], but a lack of temporal and quantitative information made causal assumptions previously impossible.

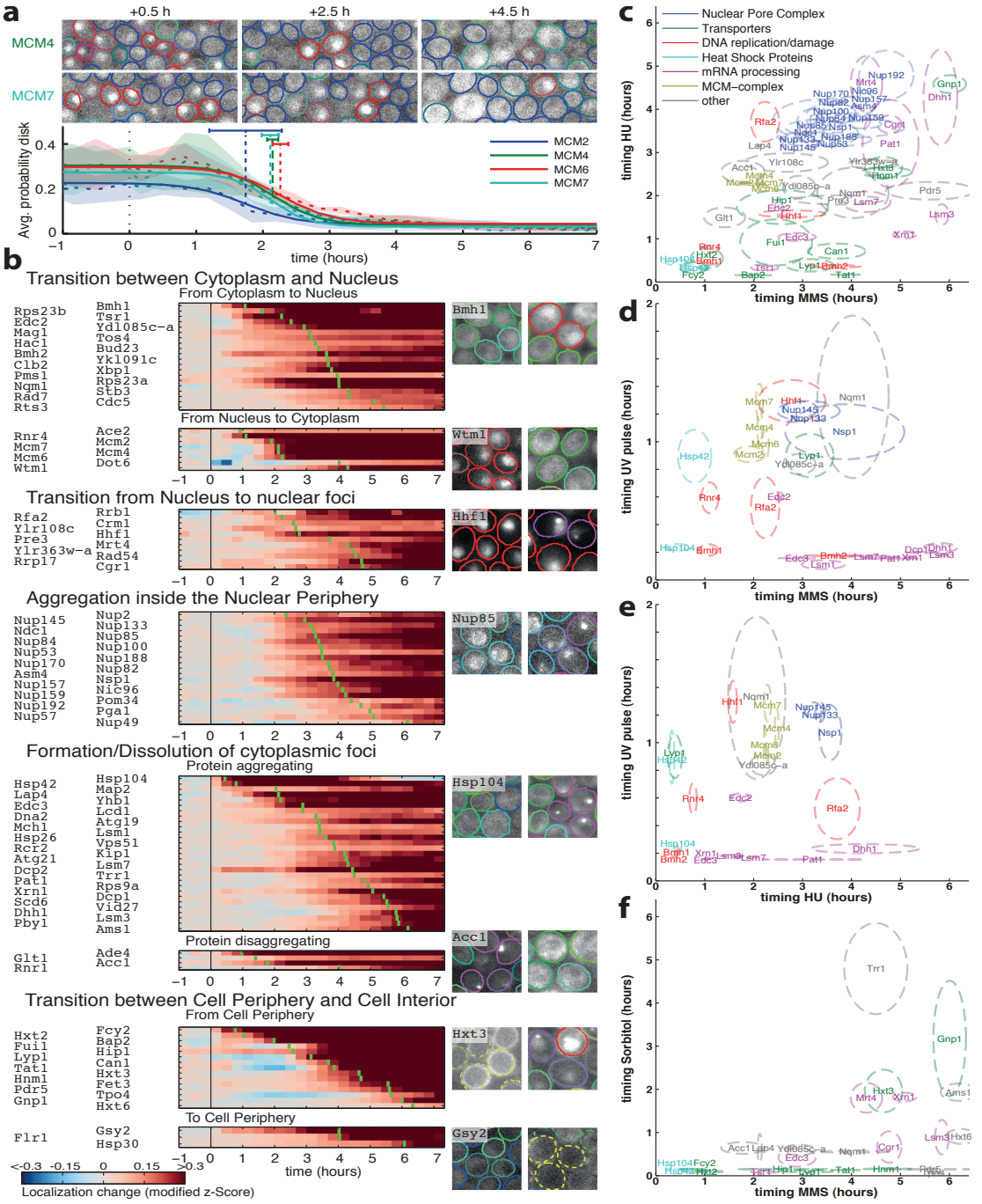

Figure 3.8: **Summary of proteome-wide localization changes.** **(a)** Average disk probability is shown over time for four Mcm proteins that translocate from the nucleus to the cytoplasm. Traces were averaged over multiple repeats. The dashed line represents the average and the transparent area shows the error (±s.d.). Traces were fitted with a sigmoid (solid line). The vertical dashed lines show the transition times and their corresponding error bars (±s.d.). **(b)** Localization change is shown over time for all proteins that relocated after MMS treatment. Proteins were grouped into five transition classes. For each heatmap, proteins were ranked by their timing, as shown by the green bar. Images show examples for each transition class. **(c-f)** Comparison of the set of 111 relocating proteins in MMS with four additional stress conditions.

3.5 Visualization of localization using our six geometrical patterns

Compressing the information of hundreds of time-lapse microscopy movies under different conditions into meaningful figures is a necessary step, both to visualize and to evaluate the data. Even with our geometrical patterns, the data still contains 10 'dimensions' (Conditions x Strains x Time x Abundance x 6 geometrical patterns). In the following, we will describe different techniques to visualize localization in different aspects of our analysis.

Most of the images contain one or two major populations of cells representing heterogeneity in localization. This is expected for several reasons: (i) some shapes such as Punctate may be intrinsically 3D and not be well captured in all cells due to limited focal depth; (ii) cellular responses are stochastic; (iii) cells may be asynchronous in their response to MMS.

Using K-means clustering (looking for two groups) of the probability vectors, we find that different populations usually show at most three non-vanishing (out of six) probabilities with non-zero values. Three probabilities can naturally be represented on a 2D equilateral triangle (or simplex) (figure 3.9a). Cells that are not exclusively part of these three probabilities can be plotted proportionally smaller and darker, so they appear to be in smaller planes in the background of the triangle.

This representation of data on a simplex can be either helpful to understand the differences in distribution of protein localization on single-cell level (figure 3.9b), or allows us to compare the distribution of proteins on a proteome-wide level (figure 3.9c). The latter was a visual confirmation that our geometrical pattern classifier was capable to give extended information about the subcellular localization of a protein, as strains with the same manually annotated subcellular localization were found to form distinct subspaces in our six dimensional space.

Another powerful way of visualization is to link our proteome-wide experiments with previously annotated data of protein complexes. *S. cerevisiae* is a well-studied model organism has a large number of known complexes [65]. Complexes are collected in catalogues [66], which can be either curated manually from low throughput experiments or using high throughput approaches like yeast two-hybrid (Y2H). It is in theory possible to use our information about localization and abundance to find proteins that form complexes. But it is more robust to use the existing information as a prior information, as so-called 'Gold Standards' of protein-protein interaction (PPI) sets are extensively validated.

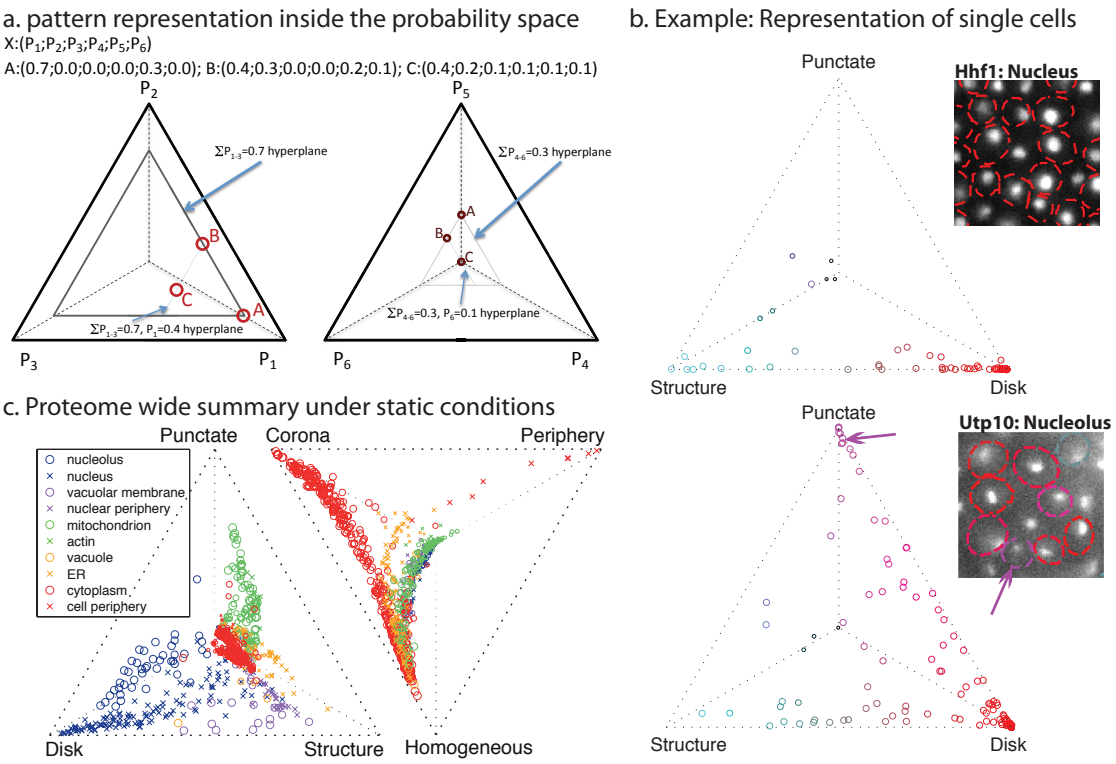

Figure 3.9: **Representation of spatial patterns inside a 6D simplex a. Three examples of single cell representation inside the probability space.** For the left simplex, the sum of the probabilities P_1, P_2 and P_3 decides for the size and brightness of the cell, for the right triangle it is the sum of the probabilities P_4, P_5 and P_6. As both sums are the same for all three cells, they have the same size and lie in the same hyperplane. As P_1 is identical for cell B and Cell C, they lie on the same line parallel to P_2 and P_3. **b. Representation of single cells inside the simplex.** Comparison of Hhf1p (annotated: Nucleus) and Utp10p (annotated: Nucleolus). While Hhf1p is uniformly distributed in disk-like pattern, Utp10p can be found in either Disk or Punctate pattern. **c. Comparison of the 6 geometrical patterns and localization annotations by Huh *et al.*.** Groups with low number of strains and strains with low intensity were excluded for clarity.

Using a comprehensive catalogue of 408 manually curated heteromeric protein complexes denoted as CYC2008 [66], we can use the extracted information from our microfluidic platform to visualize the behavior of different protein complexes in different conditions (figure 3.10). This visualization has several advantages. First, it allows direct identification of strains that behave not as expected. For example, Rps22bp and Rps23ap, two proteins of the cytoplasmic ribosomal small subunit, were found to be partially or completely localized inside the nucleus. This is known for Rps22ap and its homolog Rps22bp, but not for Rps23ap [67]. Closer evaluation of Rps23ap raised the assumption that this strain might be contaminated.

Second, this visualization is an effective way to quickly understand the functioning of cells under different conditions. Combining our proteome-wide data with curated knowledge, we can readily see how cell cycle related complexes like the Mcm2-7p or the Exocyst complex arrest during DNA damaging conditions, or that specifically complexes involved in phosphate synthesis show an increase in abundance during a nutrient change to sorbitol.

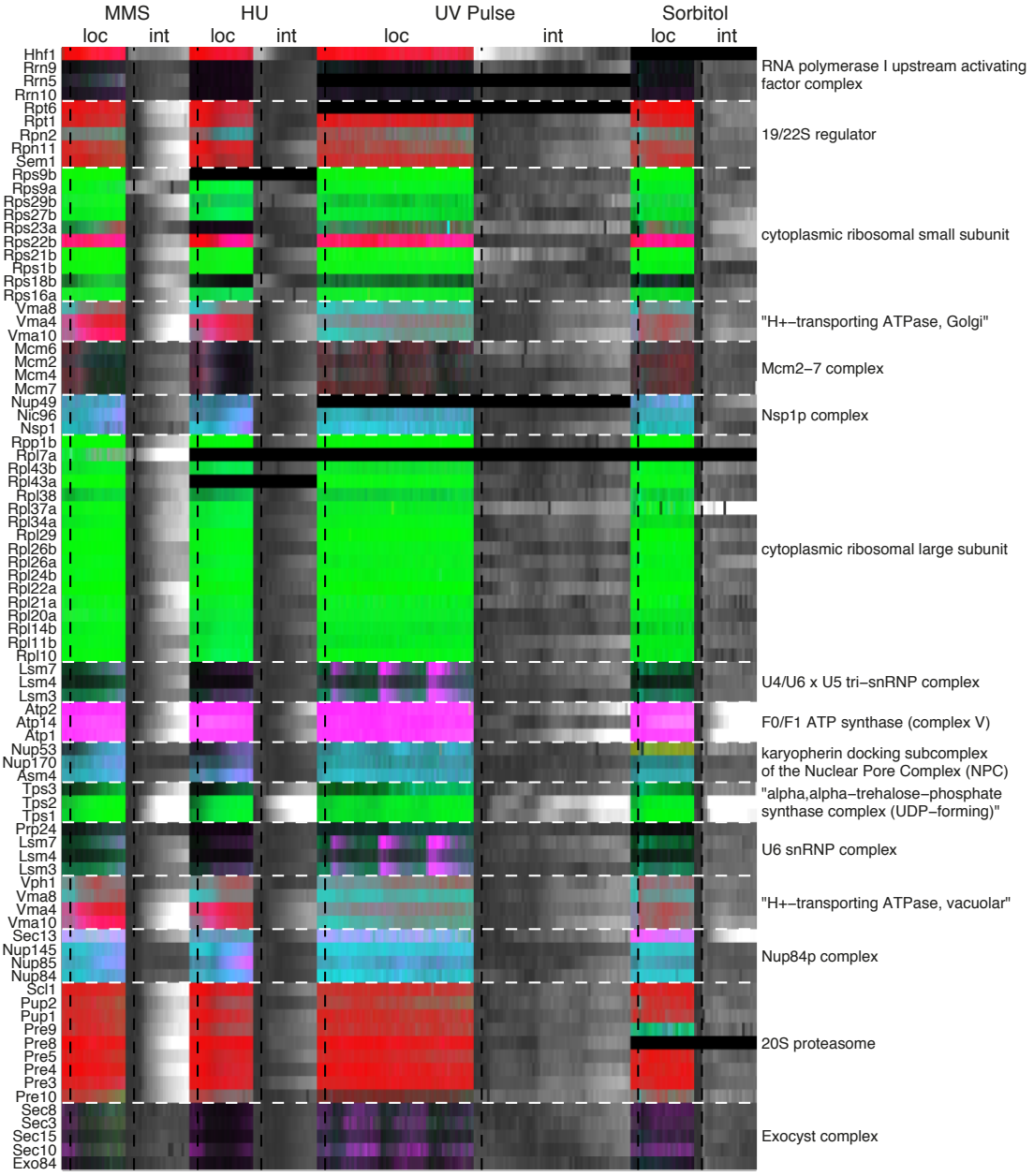

Figure 3.10: **Visualization of a manually curated complex catalogue.** Colors for localization (loc) are merged between Disk (red), Corona (green), Punctate (purple), Structure (magenta) and Periphery (yellow). Abundance (int) is shown as fold-induction, normalized using the preinduction abundance of each strain individually.

3.6 Measurement of cell growth

In addition to the information extracted from GFP as a reporter, a mutation can also influence the general phenotype of a cell. This is not an eminent concern for strains that are solely tagged with GFP, as GFP is generally found to not drastically influence the phenotype in most conditions [5]. Deleterious mutations on the other hand are expected to influence cellular networks – and therefore the phenotype of a cell.

There are two phenotypical changes that can be detected by brightfield time-lapse microscopy: cell morphology and cell growth. While there can be drastic changes in cell morphology in budding yeast, e.g. in regard to α-factor Pheromone [68], we only observed changes in size, an information that can be directly extracted from cell segmentation. During different experiments under stress conditions, we observed partial or complete cell cycle arrests. As dilution is an important factor in protein abundance [14], cell growth needs to be controlled for in a screen on gene deletions.

The optimal method for estimation of cell growth and doubling time depends highly on the data available [69]. If the total number of cells is known at any given time, a growth curve can be directly deducted. When it is possible to robustly track cells and their division, measuring doubling time is similarly straightforward [70]. Another possibility that works for growth at steady state, is the measurement of cell size distribution, using the observation that cell division in yeast is coupled with cell size [71].

All the aforementioned methods are not suitable for our high-throughput microfluidic device. Therefore, we decided to control cell growth in two separate steps, first assuring that on-chip cell growth is comparable to batch measurements and then finding a way to estimate the relative impact of changing conditions.

3.6.1 On-chip cell growth under stable conditions

To estimate the on-chip growth rate, we used a chip design with slightly modified chambers (see figure 2.1c). These chambers have highways on the part that is imaged, guiding cells along one axis (in our case the y-axis). Using a higher time resolution with time steps of 30 seconds allowed us to track cells. We decided against the direct measurement of cell division, as the

high cell density makes this prone to error.
Furthermore, most cells got pushed out of our imaging field in a short time, strongly reducing our sample size. Instead, we decided to use the cells vertical displacement. As cell movement is restricted by the chamber walls (in our case 'upwards') and depends solely on the doubling time of the cells below itself, we can estimate the cell displacement. The position y of a cell at time point $t_0 + \Delta t$ can be estimated by the cells position y_0 at time t_0 and the average doubling time T_d:

$$y(t_0 + \Delta t) = y_0 \cdot 2^{\frac{\Delta t}{T_d}} \tag{3.8}$$

Thus, the displacement $\Delta y = y(t_0 + \Delta t) - y_0$ that occurs to a cell during time step Δt is a linear function of y_0:

$$\Delta y = y_0 \cdot (2^{\frac{\Delta t}{T_d}} - 1) \tag{3.9}$$

We estimated the slope a of this function by averaging single-cell displacements. Knowing a, the doubling time T_d was approximated with the following relationship:

$$T_d = \frac{\Delta t}{log_2(a+1)} \tag{3.10}$$

We estimated T_d independently for 12 image sequences. The average doubling time was found to be 129 min (Figure 3.11), with a standard deviation of 17 min. Batch measurements gave an average of 120 min with a standard deviation of 12 min.
This is comparable to observations made in *S. pombe* [70]. The reason for a slightly decreased growth rate could be marginally unfavorable conditions for on-chip cell growth.

3.6.2 Growth estimation using local image correlation

The approach described above is obviously not suitable to measure growth during high-throughput experiments with imaging intervals of several minutes. If reporter proteins that include cell cycle markers are at hand (e.g. Mcmp2-7), changes of their population-wide location distribution can be an indication for an arrest inside the cell cycle. In our previous

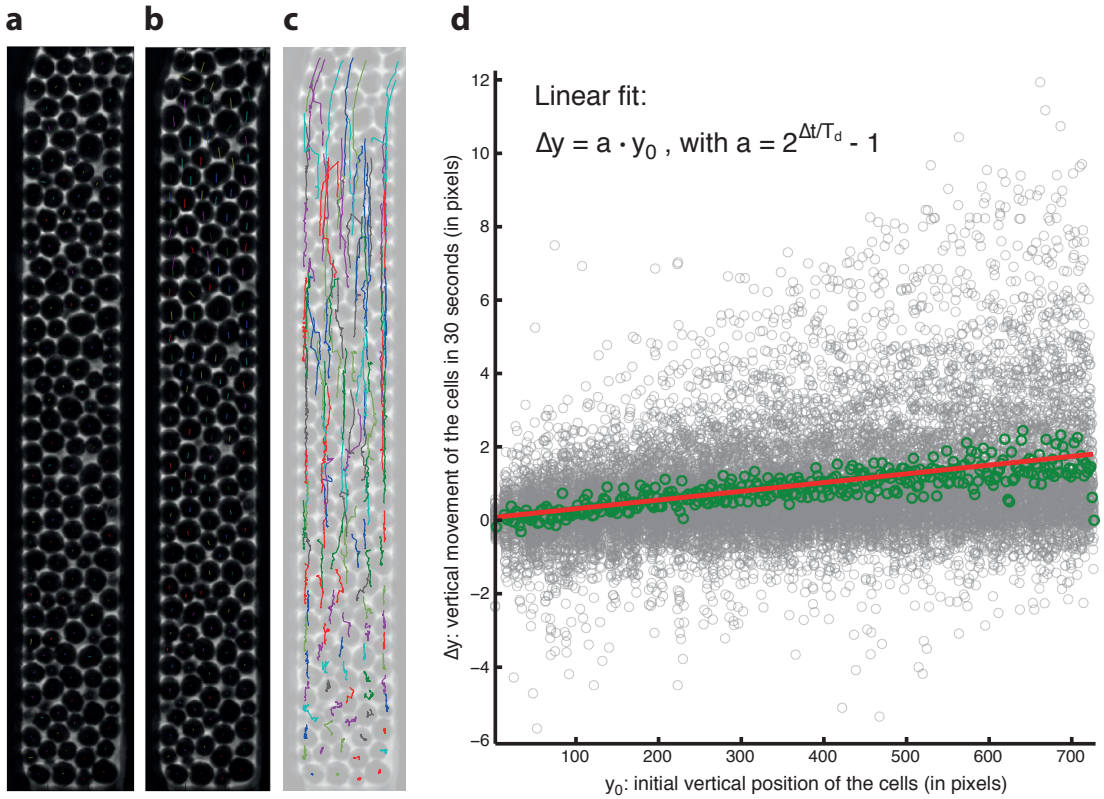

Figure 3.11: **Estimation of doubling time on chip. a&b.** Phase contrast images showing cells in a single-highway at two different time points. The colored lines represent the movement of the cells Δy, between the previous and the current frames, separated by a time-step Δt. **c.** Single-cell trajectories are given for the first 30 min of acquisition. **d.** Graph showing the correlation between the movement of cells between two frames Δy and the initial position of the cells y. Gray points represent all the cells at every time points. Green points show a moving average of the gray points, with a moving window of 2 pixels. The red line is a linear fit of the data: $\Delta y = a \cdot y_0$. The slope a enables the estimation of the doubling time T_d.

work this approach proved successful, using a bud neck marker to show the arrest of the cell cycle after induction with MMS [14]. However, this method is not practical in experiments with deletion genes where deletions of interest are expected to cause differences in growth. To overcome these restrictions, we used an inherent property of on-chip cell growth for an estimation of growth over time for each individual chamber. As cells are densely packed inside a chamber, they often get displaced in groups. This leads to stable cell formations, something that is detectable by eye even for a 20 minutes frame rate. We used a local 2-D correlation coefficient between two frames to estimate the amount of displacement between two images. The image at the first time point was 'cut' into a number of adjacent squares (Figure 3.12d).

For each square we looked for the highest normalized 2D cross-correlation value, using an

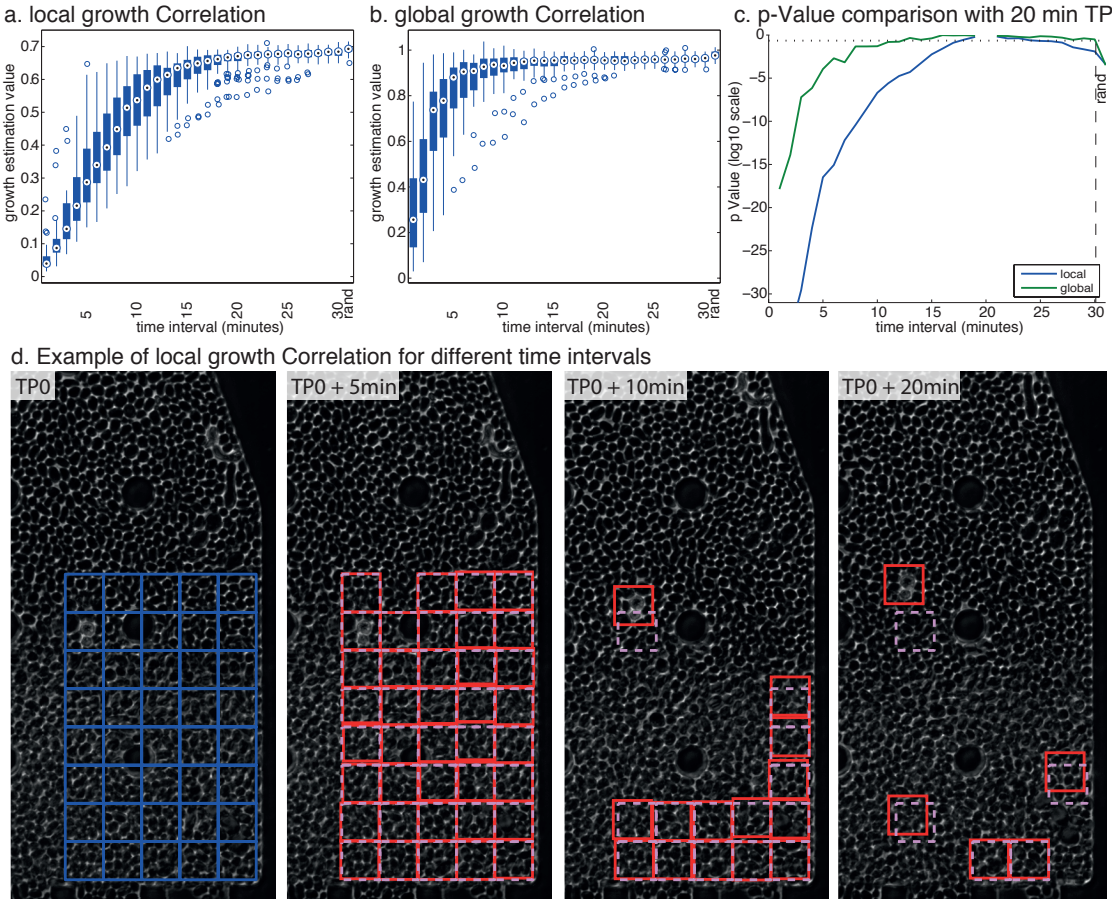

Figure 3.12: **Local correlation of images for growth rate estimation.** Measurement of **a.** local and **b.** global correlation for 44 chambers and increasing time intervals. **c.** Comparison of local and global correlation values with the 20 minutes time interval (p-Value). **d.** Example of the local growth correlation estimation method for 5, 10, and 20 minutes time interval. Red squares mark those local parts with high correlation to the respective blue squares (position indicated by light red squares in the respective frame).

overlaid but slightly increased rectangle in the second time point to find the square that is most related to our initial time point. Even though the cell formation itself remains stable, cell divisions at other parts will lead to a displacement. To keep different chambers as comparable as possible, we used a 5 (width) x 8 (height) squares grid, starting from inside the visible corner of the chamber. Square size was 60x60 pixels. We considered that up to 8 chambers could be influenced by the chamber posts. Furthermore, we never found more than 50% of squares to be considerably correlated under normal growth. We therefore estimated our growth value

g_{chamber} as

$$g_{\text{chamber}} = 1 - \text{percentile}(\vec{g}_{\text{local}}, 0.75) \text{ ,with} \tag{3.11}$$

$$g_{\text{local}} = \max(\text{normxcorr2}(\text{TP1}_{\text{inner}}, \text{TP2}_{\text{outer}})) \tag{3.12}$$

using the Matlab function 'normxcorr2' for normalized 2-D cross-correlation.
Imaging 46 chambers in one-minute steps over a time interval of 30 minutes gave us a good idea about the sensitivity of our measurements (Figure 3.12a). For example, there is a significant difference between time intervals of 10 and 20 minutes. The average value at 10 minutes gives us a lower threshold for a 50% decrease in doubling time. It can be expected to be a lower threshold, as we would expect the same change in biomass for either a 10 minutes time frame or a 20 minutes time frame at half the doubling time, under the assumption that cell size distribution remains similar.
Given that the cells would still have twice as much time to be displaced, the 20 minutes time frame is still expected to have a lower growth correlation value and therefore a higher growth estimation value. Computing the growth estimation for the images of two random chambers showed an only slightly, but significantly greater value(Figure 3.12a). This gave us an upper limit for the growth estimation value and would allow us to even detect a theoretical increase in growth. Comparing our local correlation approach with the global correlation of two images, we found the local approach to be considerably more sensitive (Figure 3.12a-c). Furthermore, if the local growth correlation estimation is taken on a chip wide level, even changes of doubling time in the range of minutes can be found with significant precision.
It has to be noted that local growth correlation is computationally expensive compared to global correlation. Requiring around one second of CPU time for the correlation of two images, it is more than 50 fold slower than a global correlation approach. However, the gain in the precision of our estimation neglects these expenses. Furthermore, we found the local approach to be robust and intuitively understandable. This facilitates adaptation in case of parameter changes (e.g. magnification or imaging interval).

4 Quantitative analysis of reporter-deletion systems in yeast

4.1 Background: Recombinant genetic techniques in yeast

The yeast GFP-library was groundbreaking to describe protein amount and localization on a proteome-wide scale. Another approach of genome-wide modification of ORFs in *S. cerevisiae* was the systematic deletion of all ~6200 known or suspected genes [17]. This Yeast Knockout collection (YKO) showed that while ~1100 genes are essential, 5100 deletions still result in viable haploid gene-deletion mutants. Viability of these deletions was tested in 1144 different chemical conditions [72], revealing that 97% of genes influence growth in at least one condition.

A second method that extensively made use of the YKO was the development of a method for systematic construction of double mutants, termed systematic genetic array (SGA) analysis [18]. This technique uses a series of steps that combine the recombinant properties of budding yeast with a robotic system for manipulation of high-density yeast arrays. Briefly, a query mutation of interest is crossed to a yeast deletion collection in which each strain contains a single gene knockout with a kanMX cassette for antibiotic selection. The query mutation itself is linked to a natMX cassette. The subsequent use of antibiotic resistance markers allows for automated selection of double mutant haploid strains.

SGA has been predominantly used for genome-wide study of double deletions [73, 74], where it allows for the detection of genetic interactions due to synthetic lethality. There are two drawbacks with viability studies. First, they only give information about drastic changes that

are strongly influential on cell growth. Second, their insight into the underlying mechanisms of genetic interactions is limited, as their phenotypic information is mostly related to growth. One system of double mutations that can overcome these limitations is the use of a reporter-deletion double mutant. Combining the deletion of one gene with another gene whose transcription or translation leads to a readout can give additional information about systematical changes. For example, studies using double deletion mutants commonly use a low number of reporter-deletion mutants to confirm deletion-related functional changes. In a recent study [10], SGA was used to combine a small number of GFP reporter genes with the genome wide deletion set, identifying previously unknown response pathways of DNA damage response.
A further advantage of reporter-deletion systems in comparison to deletion double mutants is their capability to detect changes that are transient or gradual, an information for which our microfluidic platform is well adapted. The combination of information about phenotypic response to a deletion on the one hand and information about changes to the behavior of the reporter gene on the other can be of considerable value. It allows for conclusions about causality and actual influence of a gene in regard to environmental changes that otherwise cannot be easily extracted. It has to be mentioned that there are caveats to the use of the yeast deletion collection, which we will address in section 4.2.
Thereafter, we will describe and investigate two different networks in budding yeast that are well conserved in higher eukaryotes. The first part focuses on the response of yeast to UV irradiation. Of special interest is the formation of processing bodies (P-Bodies), distinct foci within the cytoplasm of eukaryotic cells. The second part focuses on the galactose gene network (GAL), a model network in the field of transcriptional regulation.

4.2 Limitations of the yeast deletion collection

Both the yeast deletion collection and its use in SGA have been powerful tools for genome-wide studies in *S. cerevisiae* and are the foundation of a multitude of discoveries. As expected in such a large collection, a few cautions and caveats need to be considered [75]. The most important limitation is that strains in the YKO can undergo unexpected transformations. One example is a study that found aneuploidy in 22 of 290 (7.6%) deletions of genes [76]. The study screened for up- or down regulation of transcription that was deletion and chromosome

specific, which is indicative of chromosomes that are tri- or monosomic.

Using PCR to control for correct gene insertions for the reporter-deletion mutants used in our UV irradiation study [14], we found several strains carrying both a deletion or GFP tag and the respective wild type ORF simultaneously, a sign of aneuploidy. An example are all deletions of XRN1 and several deletions of PAT1. These cases of aneuploidy were often related to significant changes in cell size. In addition, we found several other strains to be increased in size, while passing our PCR. For these strains, (e.g. *mms1*Δ, *rai*Δ and *nam7*Δ), we cannot exclude the possibility of aneuploidy, as PCR could only detect changes to the respective chromosomes that include the ORF of reporter or deletion gene. Previous genome-wide deletion strain studies also found an increased size of *xrn1*Δ and *pat1*Δ [71, 77], without considering these strains to be aneuploid.

For strains generated for the GAL network screen, strain integrity has not been controlled yet. However, several deletions show an increased cell size, including those of *xrn1*Δ and *pat1*Δ. As this screen was based on more than 500 strains, it is possible to make estimations based on strain cell size distribution. Average cell size for the wild type strain and strains that are not affected would be expected to show similarity to a lognormal or gamma distribution [78], which would not account for strains with a size defect or aneuploidy. Figure cellSizeAneua shows that the distribution of our strain size cannot be fitted by a unimodal distribution. In comparison, we found a bimodal gamma distribution

$$f(x|a_n, b_n, a_\Delta, b_\Delta, p) = (1-p)\frac{1}{b_n^{a_n}\Gamma(a_n)}x^{a_n-1}e^{\frac{-x}{b_n}} + p\frac{1}{b_\Delta^{a_\Delta}\Gamma(a_\Delta)}x^{a_\Delta-1}e^{\frac{-x}{b_\Delta}} \tag{4.1}$$

to fit the distributions of strain size exceptionally well (Figure 4.1a). In this distribution, the parameters a_n, b_n are the shape and scale parameters of unaffected strains, while a_Δ, b_Δ are the respective parameters for a deletion-dependent distribution.

The value p, indicating the percentage of cells in the affected cell size distribution, is 12.2% and 7.9% of strains are more likely to be part of the distribution with increased size. Even though this number agrees well with 7.6% found in previous screens of of aneuploid strains [76], it has to be noted that our estimations are vague and presumptuous, as it cannot separate between deletions that cause actual increases in cell size and deletions for which the change in size is a consequence of aneuploidy.

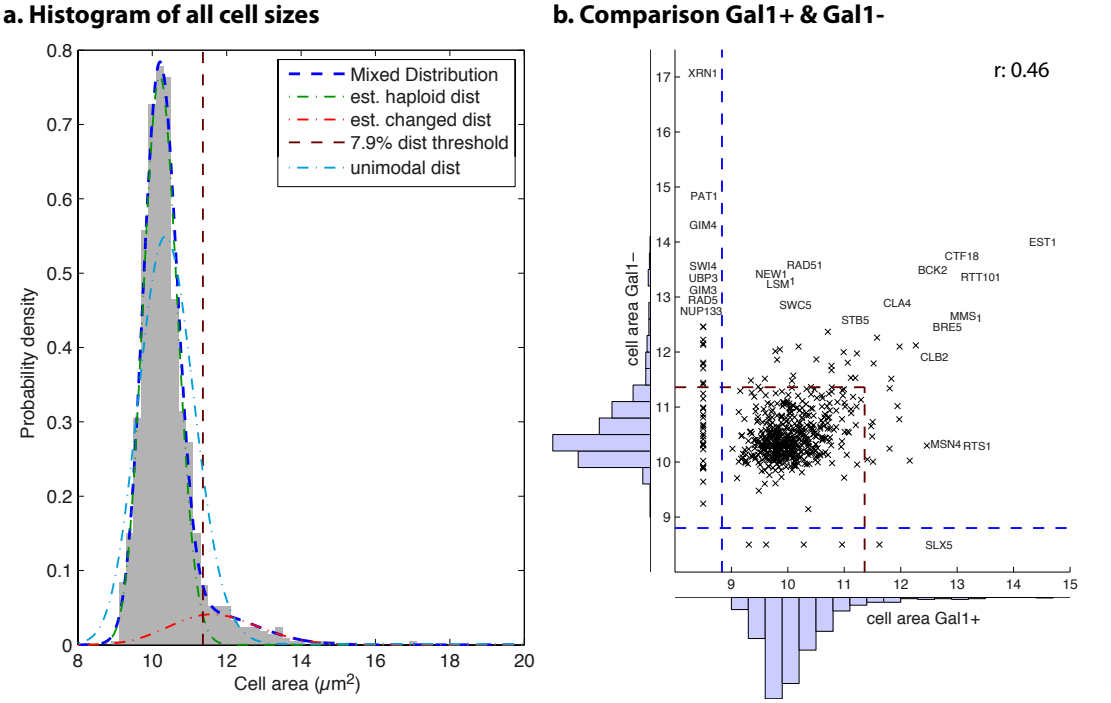

Figure 4.1: **Analysis of strain size distribution highlights the influence of deletions on cell size. a. Distribution of strain size shows bimodality.** Fit of a bimodal gamma distribution to our population sizes (blue line). Light blue line shows unimodal fit for comparison. **b. Strain sizes are correlated for strains with same deletions and different reporter proteins.** Strains outside of the blue line had good repeats for only one of the two reporter proteins. Red dashed line shows 7.9% threshold. Gene names are shown for strains with cell sizes greater than 12.5μm^2.

Nonetheless, we decided to use the threshold of 7.9% (a cell area of 11.36 μm^2) as an indicator of conspicuous size in section 4.4. 70 strains were found to have a cell population larger than the threshold in one of the two reporters and for 13 deletion strains cells were larger than the threshold for both reporters (Figure 4.1b). Of the 70 strains, 28.6% were deletions of genes that are annotated as 'cell size increased' and 32.9% as 'haploidinsufficient' in the Saccharomyces Genome Database. The latter is a moderate number, as 2134 genes in the database have the annotation 'haploinsufficient'. In comparison, of the strains without perceptible increase in growth in our experiment 6.8% are annotated as 'cell size increased' and 35.0% as 'haploinsufficient'.

Under this light, it is interesting to note that the GO annotation of genome wide screens can become misleading in a self-preserving way. For example, we found 20 of the 70 deletion strains to be annotated as 'regulation of cell cycle' genes, including XRN1. We suspect that at

least some of these annotations are the result of high-throughput annotations, which linked an increased cell size with regulatory effects.

To conclude, while YKO and SGA are invaluable tools in genome-wide screens, their results always need low throughput confirmation. Our microfluidic platform is in this regard advantageous to classic spot arrays, as it returns information on cell sizes and growth, which can be a good indication for potential pitfalls.

4.3 Gene network regulation upon UV irradiation

4.3.1 Background: Cell damage and its pathways

The response of regulatory genes or gene networks to damage is a complex question posed in all organisms. Cells need to rapidly assure that they are not irrevocably damaged and that their DNA remains intact to avoid error propagation. The elements that are involved in these processes are manifold. A typical mechanism is damage checkpoint response, leading to cell cycle arrest and damage repair [79]. Many of the genes involved in DNA damage response are expected to interact. We constructed a set of reporter-deletion strains for further investigation. Comparing our GFP library studies in different conditions, UV irradiation was the obvious candidate, as it triggered by far the fastest response in previous experiments. UVA is linked to single stranded RNA breaks by oxidatively generated damage [80]. UV induces pre-mutagenic lesions and is the most important cause for the development of skin cancer [81].

We chose 14 different reporter genes that we previously found to react to different DNA damging conditions [14]. The largest group consisted of seven proteins that are involved in P-Body formation (Dcp1p, Edc3p, Lsm1p, Pat1p, Pby1p, Scd6p, Xrn1p). P-Bodies are accumulations of proteins inside the cytoplasm [82, 83]. Figure 4.2, adapted from Parker and Sheth [84], summarizes the known relationships between different P-Body proteins. Two important aspects of mRNA decay are undertaken by proteins known to be components of P-Bodies. Dcp1p/2p (mRNA DeCaPping) is a decapping enzyme complex that removes the 5' cap structure from mRNAs prior to their degradation and Xrn1p (eXoRiboNuclease) degrades RNA by 5' to 3' exonuclease activity. Therefore, the formation of P-Bodies is assumed to play a role in mRNA degradation. Unsurprisingly, P-bodies predominantly occur during conditions that are stressful for the cell. Nonetheless, their exact role remains unclear, as previous studies showed that their formation is not a cause, but merely a consequence for RNA-mediated gene silencing [85].

Other proteins that were used as reporter genes are known to be involved in DNA damage conditions (Bmh1p, Hsp104p, Lcd1p, Rnr4p) or central cellular functions (Hhf1p, Mcm7p, Nsp1p). Bmh1p (Brain Modulosignaling Homologue) is a protein of the 14-3-3 family, which are acidic dimeric molecules that likely play a role in signal transduction [86]. It binds other

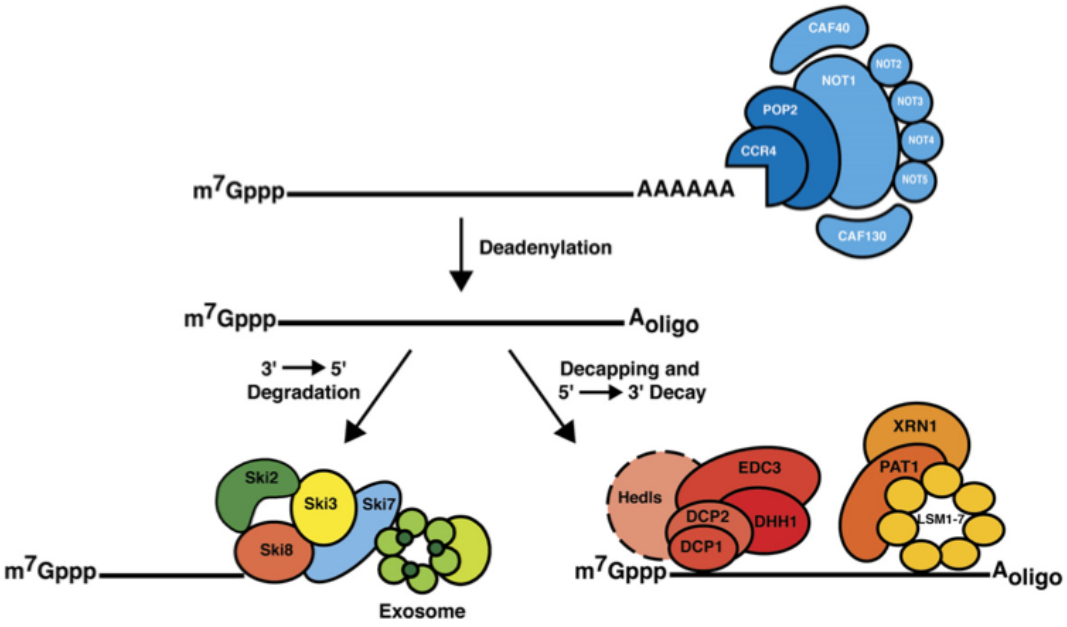

Figure 4.2: **The General Pathways and Nuclease Complexes for Degradation of Eukaryotic mRNAs involved.** Adapted from [84].

proteins and DNA and we found that its abundance increases and its localization shifts toward the nucleus upon DNA stress [14]. Hsp104p (Heat Shock Protein 104) is part of a mechanism to refold and reactivate previously denatured, aggregated proteins [87]. It was found to be produced quickly, thereby forming cytoplasmic foci [14]. Lcd1p (Lethal, Checkpoint-defective, DNA damage sensitive) is an essential protein required for DNA integrity checkpoint pathways [79]. It interacts physically with Mec1p, which itself plays a main role in regulation of P-Body formation [10].

Rnr4p (RiboNucleotide Reductase) is part of the RNR complex, which catalyzes the conversion of nucleotides to deoxynucleotides. This is the rate-limiting step in *de novo* deoxyribonucleotide biosynthesis, and therefore plays an essential role in DNA replication and repair [88, 89]. The function of the complex is controlled in two different ways [88]. Under standard conditions, Rnr2p and Rnr4p are localized to the nucleus during most of the cell cycle, while Rnr1p and Rnr3p are predominantly found in the cytoplasm. Rnr2p:Rnr4p heterodimer are found to transfer to the cytoplasm to form an active complex during S-phase or cell damage. In the case of cell damage, transcription of several RNR proteins is additionally increased, with an increase of abundance by several orders of magnitude for Rnr3p and Rnr4p.

Hhf1p (Histone H Four) is part of a core histone protein required for chromatin assembly and

chromosome function [90]. It forms foci inside the nucleus during stress and is one of the few proteins that were found to reduce its abundance during some stress conditions (UV and HU) [14]. Mcm7p (MiniChromosome Maintenance) is a component of the Mcm2-7p hexameric helicase complex. The Mcm2-7p complex localizes to the nucleus during G1, where it is a part of the prereplicative complex [91]. Nsp1p (NucleoSkeletal-like Protein) is a component of the nuclear core complex (NPC). The NPC is important for the transport of macromolecules between nucleus and cytoplasm and studies suggest that Nsp1p plays a key role in this function [92]. Nsp1p and other proteins of the NPC form foci inside the nuclear periphery under ongoing stress conditions [14, 10].

We crossed our 14 reporter genes with 40 deletion genes from a diverse range of networks. The deletion set includes the deletion of four strains that are directly part of the P-Body network, as well as 7 strains known to be part in general mRNA and rRNA processing. Other deletion strains are for example part of DNA damage networks.

The response of a deletion-reporter system to damaging stress conditions has different aspects of interest. Observing the behavior of the reporter itself (as described in previous chapters) yields information about the conditional response and function of a gene, but is not necessarily an indication for its importance for viability. We can obtain this information from gene deletions, where we can measure which of them cause significant phenotypic changes.

Finally, reporter-deletion systems allow us to combine the information about which genes influence viability with the knowledge of which reporter networks get influenced. In the end, it still remains a challenging task to deduct causality. In the section 4.3.3, we will show in which regards it is possible to answer these questions for our system and where a conclusive answer cannot be found. Most of these findings are published [14]. Reporter-deletion mutants were generated by Pascal Damay. Johannes Becker performed and analyzed UV irradiation experiments.

4.3.2 Materials and methods

Strains containing C-terminal GFP fusions at 14 genes of interest (genotype: MAT**a** his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 goiX-GFP::HIS3MX; obtained from ATTC and confirmed by PCR of the ORF-GFP junctions) were first crossed to Y9230 (MATα can1Δ::STE2pr-URA3 lyp1Δ

ura3Δ0 leu2Δ0 his3,1 met15Δ0). The resulting diploids were then sporulated and haploid segregants of the following genotype were identified: MATα his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 lyp1Δ yfg-GFP::His3MX can1Δ::STE2pr-URA3. These 14 GFP fusion strains were then crossed to a set of 40 different single-gene deletion strains (ORF replacements by kanMX4; generated by the Saccharomyces Genome Deletion Project) of the following genotype: MAT**a** his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 yfg::KanMX using a liquid-handling robot. The resulting diploids were sporulated and haploids of the following genotype MAT**a** his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 yfg::KanMX lyp1Δ yfg-GFP::His3MX can1Δ::STE2pr-URA3 were selected. Robotic mating, sporulation and haploid selection were done according to Tong *et al.* [93]. The gene disruptions in strains giving rise to phenotypes in our screen were confirmed by 5 PCR reactions designed to detect both junctions of the kanMX deletion/insertion, the absence of the corresponding junction fragments of the wild-type allele, and a full-length cassette insertion (lack of the wild-type allele).

All deletion-GFP strains were imaged in quadruplicate in response to UV-pulses. The experimental setup was identical to previous experiments with UV [14]. The samples were exposed to three successive UVC doses of increasing duration (10 min - 6.6 J/m^2, 30 min - 19.8 J/m^2, 1 hour - 39.6 J/m^2), separated by 6-hour recovery intervals. For all strains, we extracted temporal information about cell size, strain growth, abundance and localization. For localization changes we focused on foci formation and the transition between nucleus and cytoplasm, as these were the changes we previously found in our 14 reporter genes.

4.3.3 Results

Cell size and growth

We first evaluated cell size and cell growth, to look for deletion-specific phenotypes (Figure 4.3a). Most deletion strains show similar changes both in size and growth. During the first two hours after UV irradiation, cell size increases, while cell growth decreases. As two hours are close to the cell doubling time, we can assume that UV radiation leads to a general cell cycle arrest in G1. Therefore there is a time window during which cell division is not possible, but ongoing metabolism is leading to increased cell sizes.

We identified several deletions that influenced size and growth. While these two observations are clearly related, partial separation between cause and effect can be made. For example, even before UV irradiation, three deletions (*bmh1*Δ, *rai1*Δ and *mrt4*Δ) were significantly smaller and five (*tsa1*Δ, *nam7*Δ, *mms1*Δ, *rtt101*Δ and *rad55*Δ) significantly larger than wild-type strains. But for all three strains that were smaller and *tsa1*Δ and *nam7*Δ, cell growth estimation was only slightly different from normal, which has to be expected for cells that already show differences in growth in the first place.

The other three strains (*mms1*Δ, *rtt101*Δ and *rad55*Δ) were more heavily influenced in their growth, while their size remained remarkably bigger than for average cells. This indicates that, while the cells were still capable of metabolism, repair of DNA damage was more hindered than in average strains.

Two strains (*rad9*Δ and *pat1*Δ) that were not conspicuous in size and growth under standard conditions were strongly influenced by irradiation. As observed in previous studies [94, 95], *rad9*Δ and *rad55*Δ exhibited a significant growth defect upon low-dose UV irradiation. *Rad9*Δ expectedly failed to arrest after the first UV pulse, as it is known to be a vital part of checkpoint control [96]. By the second UV pulse, *rad9*Δ strains accumulated too much damage, most likely to errors in DNA copies, leading to cell death.

*Pat1*Δ, a gene mostly associated with P-Body formation [97], showed a similar behavior. This finding leads to the suggestion that Pat1p may also play a role in checkpoint control. *Rad55*Δ on the other hand arrested after the first UV pulse but also failed to repair accumulated damage, leading to cell death after the second pulse.

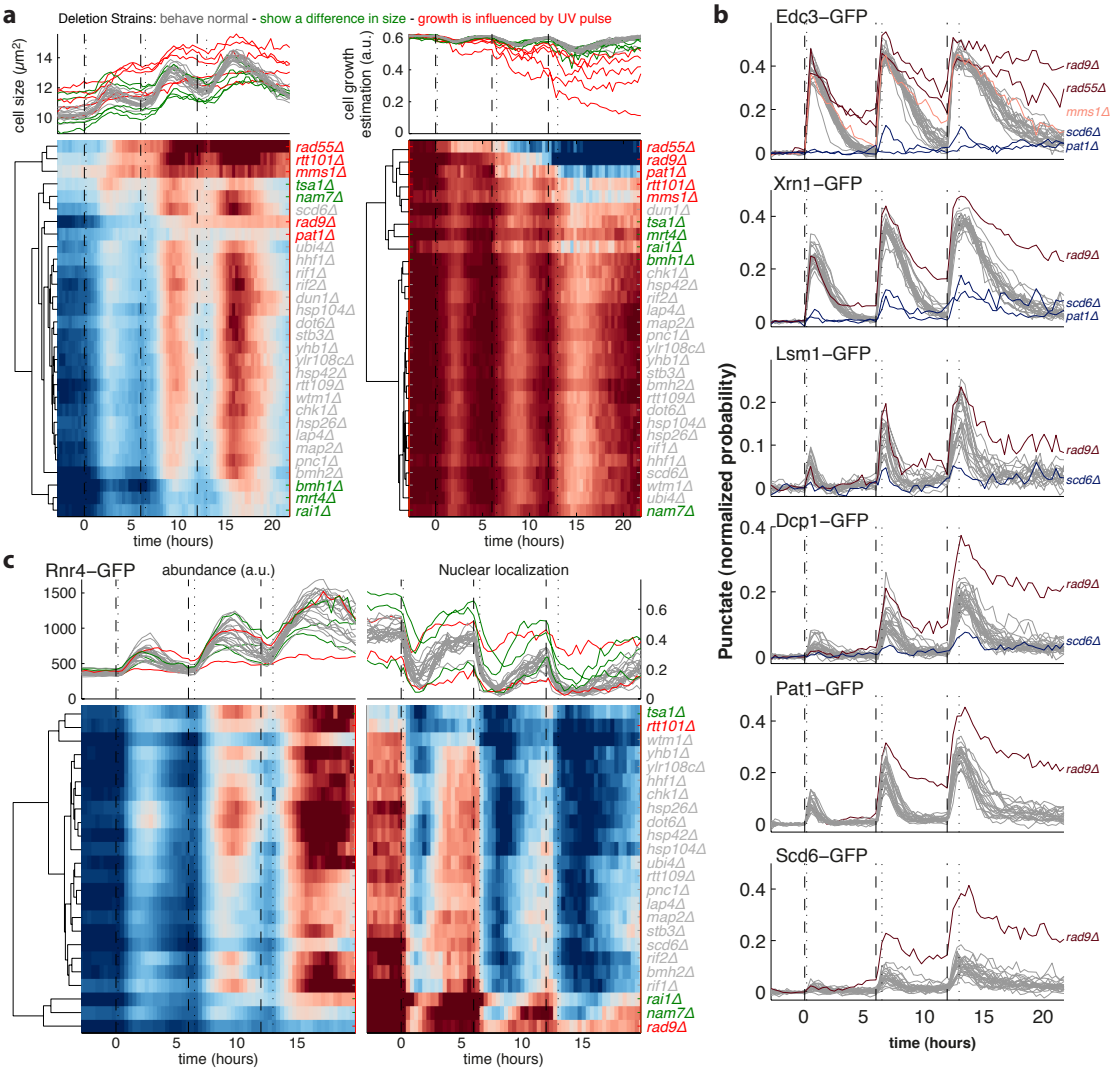

Figure 4.3: **Summary of reporter-deletion UV irradiation screen. a.** Median of strain size and cell growth of deletion-GFP strains. **b.** Punctate formation for six P-body strains in various deletion backgrounds. **c.** Changes in abundance and nuclear localization of Rnr4 as a result of gene deletions.

P-Body regulation

Most gene deletions had no significant effect on P-Body formation (Figure 4.3b). Therefore, those deletions that had an effect were clearly distinguishable. P-Body amount was increased in strains with cells that do not recover from UV radiation, most notably *rad9*Δ and *rad55*Δ. The raised amount of P-Bodies is presumably due to an increased amount of non-functional mRNA.

We found *scd6*Δ and *pat1*Δ to inhibit the formation of foci for several P-Body proteins. This was previously shown for *pat1*Δ in different conditions [83], but not for *scd6*Δ. Interestingly, deletion of SCD6 did not lead to significant changes in growth, which solidifies the hypothesis that their formation is not a cause, but merely a consequence for RNA-mediated gene silencing [85]. Even though different proteins involved in P-Body formation (e.g. Xrn1p, Dcp1/2p, Pat1p) play important roles for cell survival, their accumulation into foci is seemingly not decisive. We also compared the averages of foci formation dynamics of our 6 P-Body related GFP reporter genes 4.4. We found striking similarities and differences between different proteins. Some of these similarities can be explained functionally. For example, Pat1p and Lsm1-7p are known to form a complex [98]. A similar relation could be the case for Edc3p and Xrn1p, respectively for Scd6p and Dcp1.

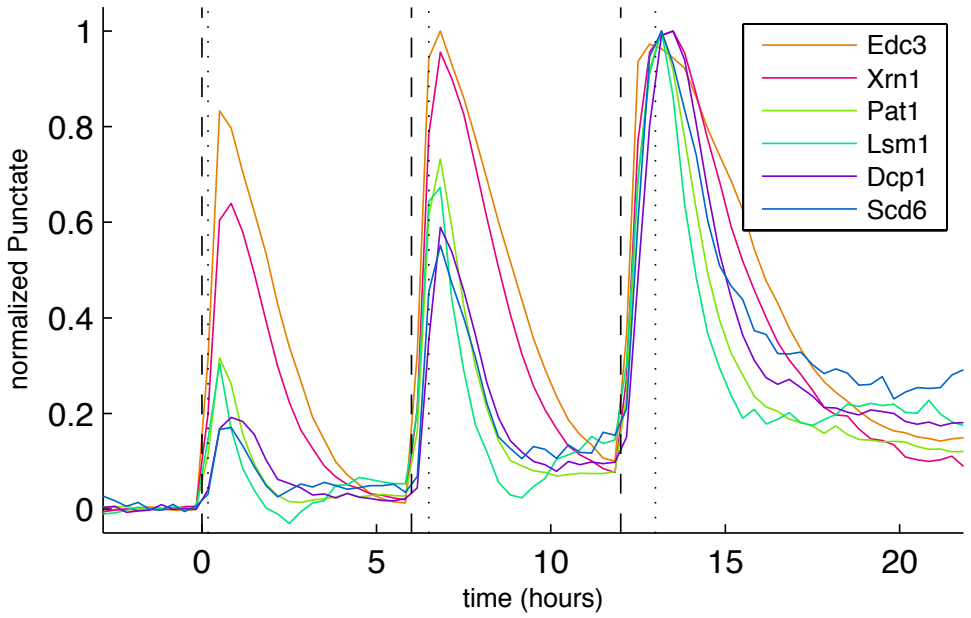

Figure 4.4: **Comparison of foci formation in different P-Body proteins under UV irradiation.** All probabilities for the localization Punctate are normalized to the same maximum value.

Rnr4p regulation

Rnr4p was found to be one of the fastest and strongest responders in all our previous experiments focusing on cell damage. Upon DNA damage, it relocates to the cytoplasm and

its abundance increases, leading to an overall increase in dNTP levels [99]. We measured the influence of deletions on both abundance and localization of Rnr4p. One deletion that influenced localization was *wtm1*Δ, for which we found no effect in growth or size. In *wtm1*Δ mutants, Rnr4p would not relocate to the nucleus during recovery phases, something that is well in accordance with the previously identified function of Wtm1p as a nuclear anchor of Rnr4p [100].

Two other strains where Rnr4p showed increased localization to the cytoplasm, *rtt101*Δ and *tsa1*Δ, showed this behavior even before the first induction and throughout the experiment. *Tsa1*Δ mutation was previously shown to lead to induction of RNR1 and RNR3, accompanied by increased dNTP levels and genomic instability [101]. Rtt101p was shown to form a complex with Mms1p and Crt10p. Crt10p is a transcriptional regulator of RNR2 and RNR3 [102]. Unlike *wtm1*Δ, both *rtt101*Δ and *tsa1*Δ showed either an influence in size or in growth. This indicates that while delayed relocation after an initial response to DNA damage might not be disadvantageous, it is of high importance for cells to regulate their dNTP levels during standard conditions.

Three deletions were found to have increased nuclear localization. *Rai1*Δ and *nam7*Δ, two genes involved in mRNA processing, showed increased size as well as nuclear localization in standard medium. *Rad9*Δ was found to have a decreased localization change up UV irradiation. This result once again shows the role of RAD9 as part of DNA damage checkpoint control. Along with the information that deletion of RAD9 does not decrease the formation of P-Bodies, it is a further indication that P-Body formation is not part of the DNA damage pathway, but instead triggered differently. A similar argument can be made for network signaling of the Bmh1p and Hsp104 related pathways.

Regulation of other reporter proteins

We detected no further indications for direct relationships between our deletion and reporter genes. The deletion of RAD9 led to no relocation of Mcm4p, a cell cycle related protein. This is a further indication for the role of RAD9 in the control of cell cycle arrest during cell damage. It is interesting to note that *rad9*Δ knockout had also no effect on abundance and foci formation in Hsp104p and abundance and relocation to the nucleus of Bmh1p. Both proteins were

shown to change their subcellular localization upon stress earlier than Rnr4p for all DNA damaging conditions [14]. This is so far not surprising, as Hsp104p and Bmh1p are known to be regulated by different pathways[103, 104], but it additionally indicates that there is no cross-talk between the respective pathways.

4.4 The Galactose network

4.4.1 Background: Galactose and transcriptional memory

The use of glucose as an energy source is highly conserved throughout evolution, starting with glucose being the main product of photosynthesis [105]. A possible reason is that it is more stable than other sugars, which is an advantage for cellular storage . In addition, it is less likely than other hexose sugars to react non-specifically with the amino groups of proteins. Yet many organisms, *S. cerevisiae* as well as mammals included, have conserved several gene networks that allow for the use of alternative energy sources. One example is the metabolism of galactose, a monosaccharide sugar found in dairy products, natural gums, and mucilages. The Leloir pathway converts galactose to glucose and is the main pathway for metabolism of galactose in humans and other species [106].

The central genes involved in this galactose network (GAL) are well studied in yeast. Figure 4.5, adapted from Stockwell *et al.* [107], summarizes these general components of the GAL network. To describe the GAL network shortly, in the absence of galactose, Gal80p inhibits Gal4p, which is a transcription factor for several GAL network proteins. The presence of glucose on the other hand leads to several mechanisms that repress GAL genes. For example, there is a glucose-dependent decrease in Gal4p levels due to transcriptional repression and active degradation of Gal4p [108]. Another example would be sequences upstream of GAL genes, where glucose-dependent proteins can bind and subsequently inhibit Gal4p binding. An example protein would be Mig1p, which binds upstream from GAL1. Overall, the amount of GAL genes is tightly controlled by a number of feedback loops, leading to a 1000-fold increase in mRNA copy numbers when galactose is present and glucose absent. This control highly increases the efficiency of the cell, as it allows the cell to only invest energy into the production of specific proteins if those are actually needed.

Another biological process has to occur after Gal4p binds to the upstream-activating sequence (UAS) and before GAL network genes can be transcribed. As DNA is tightly packed by histones, a high percentage of ORFs are not accessible. Thus, in the event of a change of nutrient sources, chromatin remodeling usually needs to precede transcription [109]. The most fundamental level of chromatin organization is the nucleosome, where DNA is wrapped

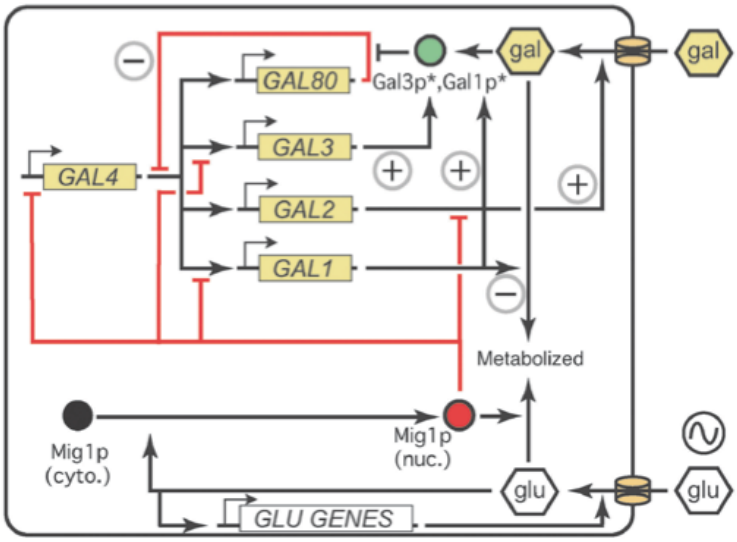

Figure 4.5: **Summary of the known GAL network.** The GAL network is controlled by interlocking positive and negative feedback loops. Asterisks indicate activation by intracellular galactose. Red indicates repressive effects; green represents inducers. Positive and negative feedback loops are marked with circled + and - signs, respectively. Adapted from [107].

around a histone. Histones are subject to post-translational modifications, which are linked to events in chromatin synthesis and assembly. For example, post-transcriptional modifications like histone acetylation can increase the accessibility to transcription factors [110].

The concentration of GAL network proteins and the reorganization of chromatin are both examples of possible epigenetic influences inside a cell, changes that may be heritable, but are not caused by changes in the DNA sequence. Main questions of epigenetics are the conservation of epigenetic information and its inheritance after cell division. There are two general ways to conserve information epigenetically.

One way is related to mRNA and protein concentrations. In the case of the GAL network, the principles of this state conservation are easily fathomable. Studies in yeast have identified many proteins with half-lives that are longer than the cell cycle [111], making dilution due to cell division the main source for decreasing protein concentrations. Most cytoplasmic proteins are expected to be evenly divided between mother and daughter cells.

Therefore, proteins involved in feedback loops can remain significantly increased over several generations. In the case of the GAL network for example, Zacharioudakis *et al.* showed that reinduction with galactose after 12 hours of glucose (6-7 generations) resulted in a rapid and uniform increase in galactose genes [19]. In the case of the same experimental setup and a

*gal1*Δ mutation, reinduction kinetics were found to be the same as for cells without galactose memory, giving rise to the claim that Gal1p concentration is an important part of epigenetic inheritance. In the following, we will call this phenomenon "cytoplasmic memory".
The other possibility for epigenetic inheritance is the propagation of mechanisms like DNA methylation and histone modification from mother to daughter cell. For this kind of "nuclear memory" explanations are not as easily at hand. Previous studies indicate that such chromatin-related mechanisms could occur in budding yeast. For example, a decrease in reinduction-rates was found in a *swi2*Δ strain [20], which is a part of the SWI/SNF chromatin remodeling complex. Another study done in a bistable state of the GAL network shows that cells that are more closely related to each other transit state formations in a related manner [112].
But both studies have limitations to their explanatory power. For example, the first induction in a *swi2*Δ deletion strain shows already strong delay in Gal1p abundance. And as most fluorescence studies can only show the protein amount of one or two selected proteins, it is in many experimental setups not possible to exclude cytoplasmic memory produced by proteins for which no reporter is at hand.
To further investigate this question of cytoplasmic and nuclear memory, we used our microfluidic device for a chromatin-wide screen that tries to highlight various genes with an influence on transcriptional regulation in general and transcriptional memory in particular. Therefore, Manolis Stavrou combined more than 500 deletion strains with two GFP reporter systems by SGA methodology. Conducting reinduction experiments using more than 500 deletions in two different backgrounds (using reporter strains with either functional or non-functional Gal1p), we were particularly interested in those strains that show altered behavior during the second induction. These strains of interest will be further studied in a second round of experiments. Poonam Bheda and Johannes Becker designed the experimental setup. Johannes Becker performed and analyzed high-throughput experiments using our microfluidic platform, while Ponaam Bheda analyzed strains of interest in more details, using cell tracking for pedigree information.
The high number of strains gives us the possibility to detect strains that are already influenced during the first induction, something that to our knowledge has not been done on a larger scale. This information is crucial, as it gives us a relative benchmark, as the behavior during

the first induction influences the second induction. Therefore, we have now the possibility to compare a specific strain not only with the wild type, but also with other strains that showed similar changes during the first induction.

4.4.2 Materials and methods

In the following, we will first explain the experimental design (construction of a library and experimental setup of the microfluidic device). Afterwards, we will describe the different steps of our data analysis. Several steps are necessary to allow for robust quantitative analysis of these data. Stringent quality control is especially necessary to control the influence of extrinsic noise. Furthermore, it is important to condense the data in an understandable manner. This condensation is necessary to assure the comparison between experimental repeats. In a last step, we describe a coherent way to detect outliers and to represent our findings.

Experimental design

Yeast library construction. Gal1-GFP reporter strains were made in parent strain Y7092 (SGA query strain) by either replacing the Gal1 ORF with GFP (RSY16) or as a C-terminal fusion with GFP (RSY17) by homologous recombination of a PCR product containing GFP with a natMX cassette for selection. The GFP is a fast maturing ("superfolder GFP") and destabilized variant knocked-in at the endogenous Gal1 locus for expression under the control of the native promoter. Gal1-GFP fusions have previously been used successfully, with no obvious effects on Gal1 expression or activity [19].
Experiments were conducted with two different sets of strains, both being produced using the aforementioned approach. Initially, a library containing 169 strains was used, both as a proof of concept and for parameter optimization. Afterwards, we used the SGA approach to create a comprehensive chromatin-associated factor Gal1p-reporter-deletion library. Approximately 500 mutants of factors associated with chromatin were crossed in both the RSY16 and RSY17 backgrounds to create ~1,000 total strains containing both a Gal1 fluorescent reporter and a single gene knockout.Selected strains were verified by junction PCRs to detect the presence of kanMX and natMX cassettes and absence of wild-type alleles
Experimental setups. A very important consideration in galactose reinduction experiments

is to find a good duration for the different media pulses. The general setup is the use of SD medium with

2% raffinose	stationary growth
2% glucose	4h start of imaging
1.5% galactose, 1.5% raffinose	Xh first galactose induction
2% glucose	4h
1.5% galactose, 1.5% raffinose	Yh second galactose induction

Two steps of four hours for glucose pulses were chosen as a previous study showed that wild-type strains maintained reinduction memory for at least four hours (around two doublings) [20].

The variable of most interest was the duration of the first galactose pulse. To facilitate comparison between the inductions, it is of importance that as many strains as possible are considerably induced after the first induction. But at the same time induction should not be too long, so that GFP intensity of most cells reduces to background levels during repression. This is important for two reasons. First, it facilitates comparison of the two inductions, because the influence of ongoing degradation of GFP proteins due to the first induction is reduced. Second, as mentioned previously, Gal1p was found to be a main factor in reinduction rates. While we are aware that the amount of GFP inside a cell is not a direct measurement for the amount of mature Gal1p, it is at least a good indicator.

We found induction times around two hours the most promising. As expected, strains without GAL1 ORF (Gal1-) induced much slower in comparison to strains with GAL1 ORF (Gal1+). To account for this difference, we decided to split the second Gal1p library into two sub-libraries, one consisting of the Gal1+ and the other of the Gal1- strains. This allowed us to adjust the galactose times for each subset individually, using 1.5 hours for Gal1+, and 2.5 hours for Gal1- strains.

We used 60x magnification for the experiments on the comprehensive data set, which allows us to image a high number of cells in each chamber with a reasonable resolution. Each experiment contained duplicates for each strain, which were spotted in separate rows to avoid experimental biases. It has to be noted that the continuous flow of media was shortly interrupted between the overnight growth with raffinose and the first glucose repression. This was necessary as the design of the microfluidic device allowed only for two different medium

sources.

Data measurements

For each chamber, we monitored a number of values over time. These values are:

- **Abundance: mean, median, standard deviation (std), noise (std^2/mean^2).** All values are collected using the standard background subtraction method and are in arbitrary units.
- **Cell active percentage.** We estimated for each time point the percentage of cells that are in an 'ON' or 'OFF' state respectively. This percentage allows us to make an estimation of the time delay to start induction in different strains. Using an abundance threshold of 150 (a.u.), we found that all strains are close to or completely inactive in glucose and completely active in steady-state galactose.
- **Gradient measurements: gradient, Pearson's correlation coefficient, adjusted standard deviation (adjStd), adjusted noise (adjNoise: adjStd^2/mean^2).** As cells inside the chambers are provided with new medium by diffusion, a gradient in nutrients has to be expected and was found in previous studies [50]. We used linear regression to adjust standard deviation and noise measurements for this gradient.
- **Cell information: growth, mean cell size.** We monitored values related to the general state of cells to gain information about the phenotypical behavior of different deletion strains. Growth values are estimated as described in section3.6.2. Cell size is in μm^2.

To facilitate the display, we added interpolated data for mean and median abundance, cell active percentage, adjusted noise, correlation coefficient and growth. This way, one time vector with steps of 10 minutes can be used for all chambers and experiments.

Quality control

Quality control was done in several steps. Imaging the whole device at 4x magnification during overnight growth in raffinose, we picked only those chambers that were completely filled

with cells by the start of the experiment as a stringent quality filter. Chambers that were not filled at this point, but whose 60x magnification images were completely filled with cells passed a less stringent filter. The reason for this late growth could be either phenotype or cell spotting related. These strains were not used in the general analysis, but could be of interest to investigate the relationship between growth deficits and Gal1p expression.

A second quality filter was added to control for possible perturbations inside the device. Air inside the flow lines or flow lines that are clogged with cells can disturb the media flow, which is not possible to observe during the experiment. Therefore, we compare the median values between good quality strains of growth and mean intensity for each row. This makes it easy to discard either single rows or complete subsets that use the same inlet. An example of both cases is shown in 4.6.

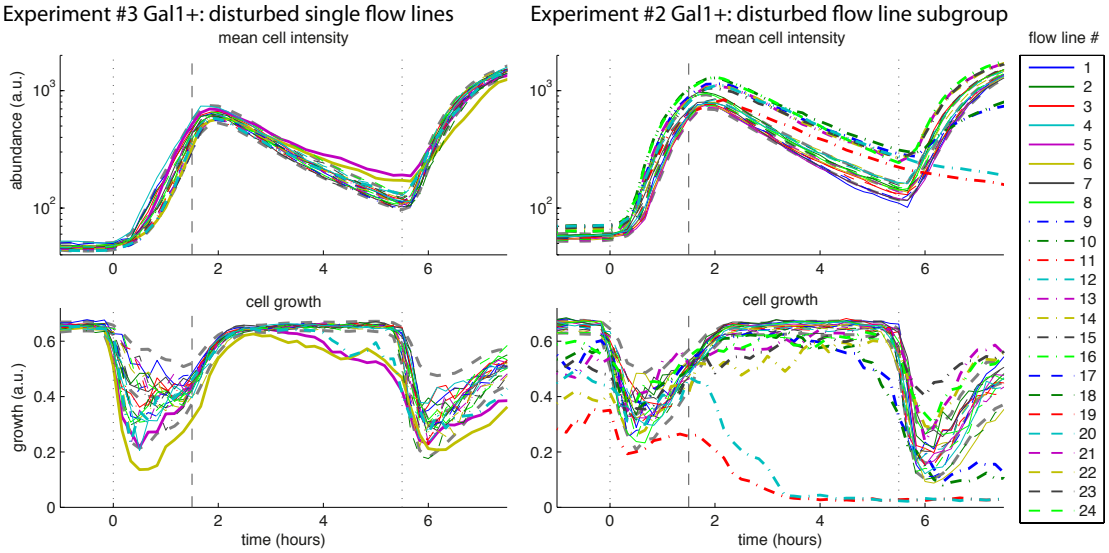

Figure 4.6: **Control of flow line quality in galactose experiments.** Two example experiments: Plot of median of all strains that were manually annotated as 'good quality'. Dashed grey lines show 25 and 75 percentile of all 'good quality' strains from nonsuspicious flow lines. Median values of perturbed flow lines are drawn thicker. Experiment #3 in Gal1+ strains showed suspicious behavior in flow lines 5, 6 and 12. Experiment #2 in Gal1+ was found to have perturbed nutrient supply in subset #2.

Comparison of independent experimental repeats

Having three independent experiments for both Gal1+ and Gal1- strains, we compared abundance and growth, to determine experiment reproducibility. We have two different possibilities to evaluate the experiment-to-experiment and intra-experiment variance. The first is to look at the behavior of the wild-type strain, which we spotted in eight repeats for each experiment. It has to be noted that 'wild-type' refers to the Gal1 reporter strain with no other deletions and is not a consensus wild-type strain, as it has besides the GFP reporter an inserted kanMX cassette due to the SGA procedure. The second possibility is to assume that most deletions have little to no influence on the GAL network, therefore making the median and inner quartiles of strains comparable to the wild type. We focused on the second approach, as numbers of good wild-type repeats for each experiment was found to alter between three and eight.
Figure 4.7 summarizes median and inner quartiles of our experiments. The most important information of this figure is that experiment-to-experiment variance is high in comparison to intra-experiment variability. This is indicated by several experiments for which there is little to no overlap for the inner quartile range. A possible reason for this high variability could be the high complexity of the GAL network and all the involved mechanisms, so that even slight changes in pressure, timing or nutrient concentrations can strongly influence the cellular response.
For example, the time window of interrupted media flow, which was needed to change the nutrient sources, was prone to vary between 20 and 40 minutes. This could be prevented in future experiments with a slightly altered chip design that allows for the simultaneous connection of three media sources. Furthermore, a recent low-throughput study that focused on growth rate changes in *E. coli* shows similar variance between experiments [113].
Especially for the Gal1- experiments we see remarkable differences between the slopes for both abundance and growth. The reason could be that Gal1- strains are more likely in a region of bistability during our experiments than Gal1+ strains. Bistability was found in a recent study for galactose concentrations of 0.1% or less [114]. Our medium has a galactose concentration of 1.5%, which is expected to be gradually reduced inside the chambers. Therefore, it is possible that this critical concentration is reached. Growth can also be influenced by the inability of cells to effectively use the Leloir pathway for galactose processing.

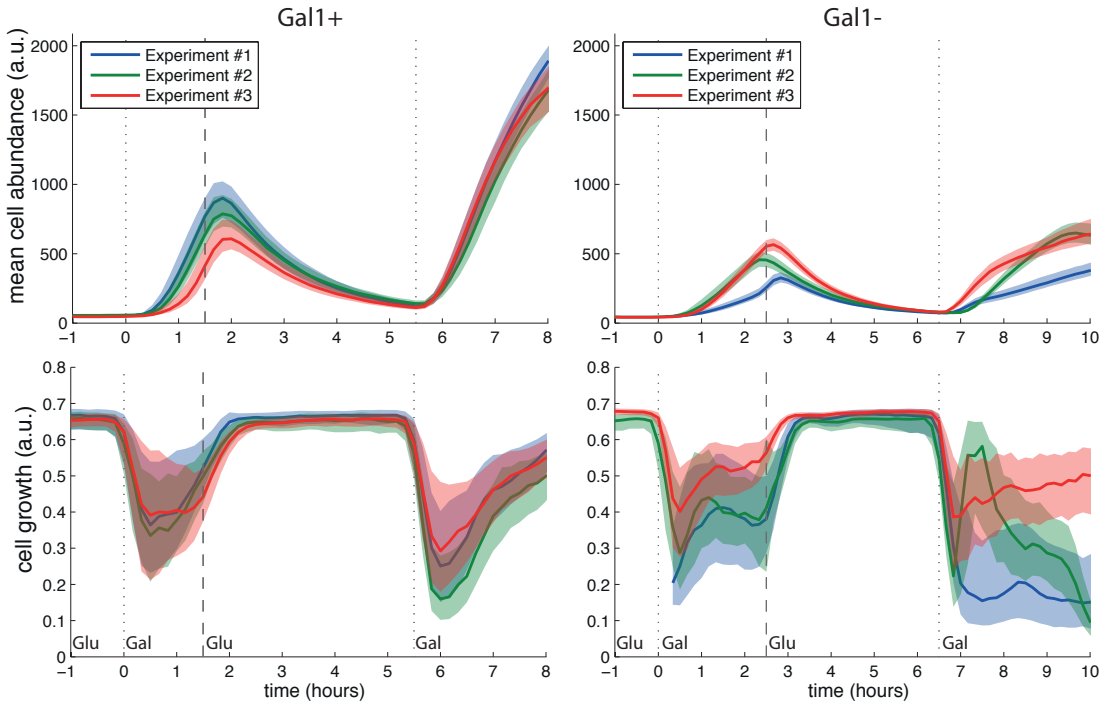

Figure 4.7: **Summary of all galactose screen experiments.** Median and 25 and 75 percentile of abundance and growth measurements for all experimental repeats in Gal1- and Gal1+.

An additional general observation for Gal1+ strains is that we find the temporary growth decrease during the second induction to be more severe than during the first induction. This is contradictory to the general assumption of memory also confirmed by previous findings [114]. But it could very well be that slight changes inside the chip over time influence this growth adaptation negatively.

Disregarding the aforementioned variance between experimental repeats, we were still able to extract valid results, as will be shown in the following sections. One reason therefore is that similar to other high-throughput experiments, we are mainly interested in relative differences between strains, which are expected to be conserved even under slightly varying conditions.

Data condensation

To further condense some of the information, we reduced the time traces of mean abundance and active percentage to single data points for each galactose induction. For mean abundance, we estimated the abundance of each chamber 30 minutes before the first galactose induction

ends and after the same duration during the second induction Thus, we estimated abundance after 1 hour of induction for Gal1+ and 2 hours for Gal1-. The estimation used a weighted linear fit. The weight w_{TP} was necessary to avoid fitting out-of-focus images and used the number of good cells NbC_{TP} at the respective time point TP

$$w_{TP} = \frac{NbC_{TP}^2}{50^2 + NbC_{TP}^2} \tag{4.2}$$

The fit uses the data acquired between 0.5 hours before and after the estimated time point, a period of time that usually includes 3 acquired images for each chamber. Manual quality control showed satisfying results for all chambers.

The information contained in mean abundance is convoluted. Changes both in induction rate as well as the influence of transcription delay can have similar effects on the intensity. And in cases where some of the cells do not return to background levels, separating the abundance of cells before and during the second induction is non-trivial. Although in no way perfect, the use of cell activation percentage can at least partially account for these problems. Estimating the time at which a certain percentage of cells becomes active gives us an idea for the delay time of a strain. And using the information that a certain percentage of cells remained active at the start of the second induction allows us to focus on the percentage of cells that became inactive. This estimation is possible because we observed during single cell tracking experiments that active cells do not turn inactive during induction.

There are two general thresholds that play a role in estimating the time of cell activation during induction. One of them is the aforementioned abundance threshold, the value that decides if a cell is assumed to be active or not. The other one is a percentage threshold, taking the time as estimation that a certain percentage of cells need to become activated. To estimate this time, while adjusting for potential outliers, we fit a smoothed cubic spline to our data values, using the Matlab function 'csaps' with weight w_{TP} as in formula 4.2. To fit the smoothed spline, we used only time points after the respective inductions. We estimated a minimum value of active cell percentage a_0, using the minimum percentage of active cells one hour before the first induction (a value that was usually close to zero percent) for the first induction, or the minimum percentage of active cells +-20 minutes around the second induction for the second induction. The threshold $actT_1$ for activation time was composed of the fixed active

percentage threshold parameter $actT_0$ and the estimated value of active cell percentage during the beginning of induction a_0

$$actT_1 = a_0 + (1 - a_0) \cdot actT_0 \tag{4.3}$$

As an example, if 50% of cells were already estimated as active at the beginning of the second induction the cubic spline needed to pass 50%+0.5*actT0 for the estimation of the second time. The spline was evaluated every 0.01 hours and the time point it passed $actT_1$ was taken as time estimation. Manual quality control showed satisfying results for all chambers. If the spline did not pass the threshold, something that only happened for a few chambers during the first induction, it was stored as a late inducing outlier, storing the maximum active percentage in addition. *Gal3*Δ and *gal4*Δ strains did not induce, in agreement with the literature [107]. Varying the abundance and percentage thresholds in a certain range, we could not observe any unexpected outliers, validating our approach as stable. An activation threshold $actT0$ of 50% was chosen for two reasons. First, 50% is as well the time point were the median crosses the activation threshold, symbolizing the time were an 'average' cell would be estimated as being activated. Second, the medium shape of the active cell percentage over time can be approximated by a logistic function. This function is the steepest for 50%, making it the threshold with the smallest theoretical estimation error.

Detection of outliers/ strains of interest

In a first step, we were interested in all strains that are outliers. For this, we looked for outliers inside the two-dimensional distributions of our condensed abundance and timing values. A non-linear correlation between both inductions and the previously mentioned experiment to experiment differences in induction behavior made it impractical to find a conventional distribution.

In addition, most strains were not available as sextuplicates, due to aforementioned reasons. Therefore, conventional statistical tests like the Z-test were not appropriate. Instead, we used a cutoff-based approach to detect strains that are repeatedly different from the norm. First, outlier cutoffs focused on the first induction. We combined a percentage-based cutoff with an interquartile range approach. The latter is used to avoid missing potential strains of interest

due to a fixed percentage cutoff. In the following, the percentage is set as the 2% and 98% quartile and r for the interquartile range is set as 1, which would be as well around 2% for a normal distribution. This 1D approach highlights those strains that show a phenotype during a single induction.

To include the second induction, we sorted the data points after their first induction values. We used a moving window of 11 data points combined with the interquartile range, to obtain moving thresholds for upper and lower outliers. We once again used 'csaps' and r of 1 to get smooth curves for these values.

Combining these two approaches, we can find 8 different potential types of outlier. For abundance, we can find strains with high or low abundance in the first induction, as well as strains with a reinduction that is higher or lower than expected. For the response time, we can identify fast or slow inducers, as well as strains for which the reinduction is faster or slower than expected. These outlier types are obviously pair wise related, as for example a faster response is expected to lead to a higher abundance.

To reduce outliers that are simply due to extrinsic or intrinsic noise, strains had to reproducibly identify as outliers. As a threshold for this repeatability filter, we chose the requirement to be the same kind of outlier in at least two different experiments and for more than 50% of repeats. Strains that are outliers in more than 50% of repeats, but only in one experiment are labeled as 'inconclusive outliers'. Figure 4.8 summarizes this method for all experiments for both abundance and cell activation.

We estimated the number of outliers we would expect in a random population of strains, simulating 10.000 random draws of our experiment. A p-value indicates the number of times that a random draw showed at least the same amount of 'reliable' outliers. While this does not indicate a p-value for each specific strain, it is at least a good indicator about the general reliability of outliers of a specific type. A summary of the detected strains and the estimated p-values for each type can be found in table 4.1.

Further condensation and visualization

To allow the merged representation of all three experimental repeats, we applied locally weighted scatterplot smoothing (LOESS). We used the mean of repeats that existed in all three

Outlier Type	high Gal1 1st induction	low Gal1 1st induction	high Gal1 2nd	low Gal1 2nd	fast Responder	slow Responder	fast 2nd ind	slow 2nd ind	no Response
Gal1+	EAF7, IKI3, MIG1, UBC4 (4/0.0024)	CHD1, CTI6, GAL3, GAL4, ITC1, PRE9 (6/ <1E-4)	(0/ 1)	CIT1, GAL83, ITC1, RTT103 (4/0.0001)	AIM4, MIG1, UBC4 (3/0.0003)	ADE1, CHD1, CHZ1, CTI6, GAL83, ITC1, NHP10, PRE9, THP2 (9/ <1E-4)	(0/ 1)	BRE5, EAF7, MLH1, RRD1, RTT103, SAP30, SET3, SWC3 (8/ <1E-4)	GAL3, GAL4 (2/<1E-4)
Gal1-	ACH1, CLA4, CRC1, CTF18, EST1, HDA3, HEL2, MIG1, MPP6, NEW1, RAD5, RAD51, RTT101, UBC4, UGA3, YBP2 (16/ <1E-4)	CHZ1, CTI6, DST1, GAL3, GAL4, GAL83, ITC1, NHP10, PRE9, SOH1 (10/ <1E-4)	DPH5 (1/0.6)	(0/ 1)	BCK2, BRE5, CLA4, CTF18, EST1, HDA3, HXT17, MIG1, RAD5, RAD51, RTT101, UBC4, UBR2, UGA3, YBP2 (15/<1E-4)	ADE1, CHZ1, CTI6, DST1, GAL83, ITC1, NHP10, PRE9, RPN4, SDS3, SEM1, SOH1 (12/ <1E-4)	(0/ 1)	(0/ 1)	GAL3, GAL4 (2/ <1E-4)
TOP3 enriched GO	ribosome (0.014)	protein complex biogenesis (0.00044), histone binding (0.0065), transcription termination, DNA-dependent (0.0072)			protein modification by small protein conjugation or removal (0.037)	protein complex biogenesis (7.8e-05), nucleus (0.0027), transcription termination, DNA-dependent (0.012)		chromosome (0.023)	

Table 4.1: **Summary of detected outliers.** Numbers in bracket indicate the number of detected strains and an estimated probability that this number could have occurred by chance. 'TOP3 enriched GO' mentions the 3 GO annotations (GO slim) with the lowest p-Value ($p<0.05$) and the respective p-Value

experiments to compute a local regression curve for each of the four data points (abundance and response time during first and second induction) in each experiment. These curves were then used to standardize all individual chambers towards one average experiment. Figures 4.9 and 4.10 illustrate this procedure for Gal1+.

The advantage of LOESS is that it reduces the noise, as averages are now taken from up to six repeats. Strains with no or little influence on galactose expression tend to regress further towards the mean. This allowed us to highlight further strains that did not pass the more stringent threshold, but are nonetheless repeatedly found at a specific position inside an experiment. Just like previously mentioned outliers with a low number of repeats, these strains were labeled as 'inconclusive outliers'.

To summarize, the further data condensation using LOESS allowed us to represent the data in a more condensed manner, while adding a continuous aspect to the previously very discrete

architecture of our outsider detection. One caveat of LOESS is that it can lead to non-trivial distortions of the data space, making the use of standard deviations for further going statistical estimations impractical.

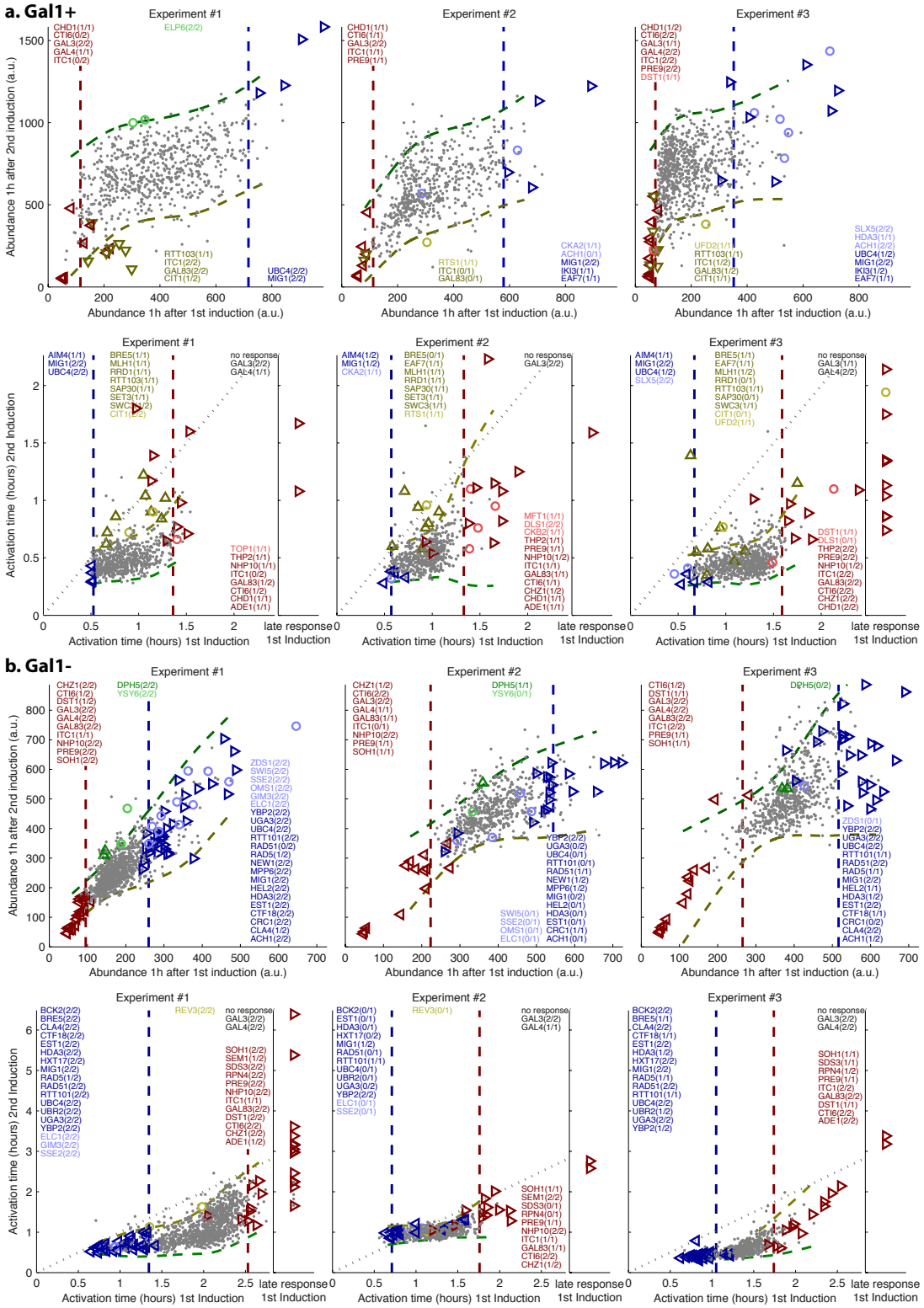

Figure 4.8: **Summary of outlier detection for (a) Gal1+ and (b) Gal1- strains.** Abundance and induction time outlier for all repeats. Numbers in brackets show number of repeats detected as outliers and number of repeats annotated 'good quality' strains. Lighter colors show inconclusive outliers.

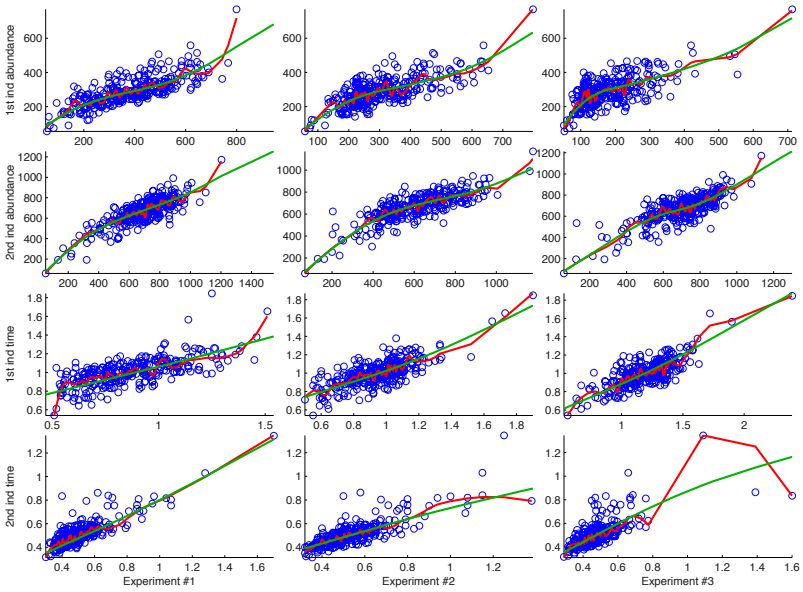

Figure 4.9: **Normalization of Gal1+ using LOESS.** y-axis shows average values of strains in abundance and induction time for both first and second induction. x-Axis shows the respective values in the experimental repeats. Red lines show the result of LOESS, green lines show the final local regression lines used for data normalization.

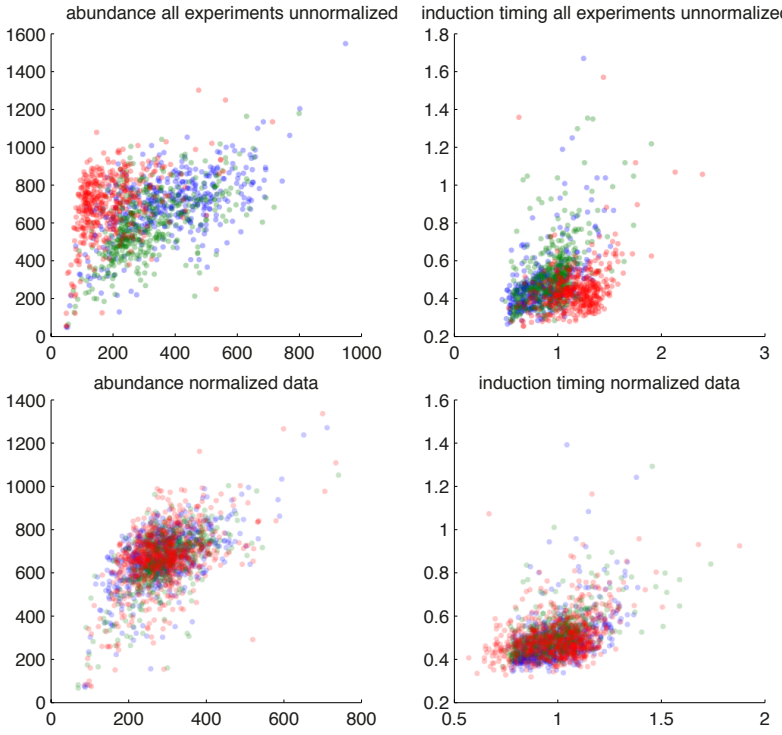

Figure 4.10: **Overlay of Gal1+ experiments after normalization by LOESS.** Upper panels show raw values for abundance and induction time estimation, lower panels normalized values. x-Axis shows the first induction, y-Axis the second.

4.4.3 Results

The main goal of a biological screen is to identify candidates, in our case genes that are likely to influence nuclear memory. The number of potential candidates is hereby often subjective. Depending on subsequent experiments and resources, outlier detection thresholds can be more or less stringent. A less stringent outlier threshold increases the number of detected strains of interest while at the same time increasing the likelihood of detecting a false outlier. In the following if not noted differently, we will use a list of our most stringent outliers to highlight several findings. This information is mainly to give a general overview of what seems to be influential on GAL1 induction and would of course need further investigation. General gene information was taken from the *Saccharomyces* Genome Database unless stated otherwise. It has to be noted that these strains are not yet controlled for genetic integrity, a problem of deletion mutants mentioned in section 4.2. Before focusing on specific strains, it is important to note several general aspects of GAL network induction dynamics.

Observation of general induction dynamics

Preliminary data acquired in a previous experiment that compared induction after overnight growth in raffinose and overnight growth in glucose gives a good estimation for upper and lower limits of response time for both Gal1+ and Gal1- strains (Figure 4.11a). Discarding 1% of all repeats that were found to be strong outliers (like *gal3*Δ and *gal4*Δ), the response time in raffinose is estimated to be around 25 minutes for all Gal1+ and Gal1- strains, with a standard deviation of less than five minutes. Response time in glucose on the other hand differs vastly between Gal1+ and Gal1- strains. This is a strong indication that Gal1p plays a yet unknown role in induction of GAL network genes, as its known role of binding the inhibitor Gal80p is most likely glucose independent.

The Gal80p-related role becomes visible if we compare the mechanisms of Gal1+ and Gal1- strains during the first induction. We plotted the first induction data point for response time against an estimation of abundance after one hour for all 6 experiments (Figure 4.11b). For all deletion genes in both Gal1+ and Gal1-, induction times were consistently slower than 25 minutes. This indicates that growth in raffinose is a lower threshold for induction response, which most likely cannot be overcome by a single deletion. The interesting finding is that Gal1+

and Gal1- strains are found on different curves, each curve itself being extremely correlated (Figure 4.11b). This shows the important role of Gal1p as part of a positive feedback loop in the GAL network [19].

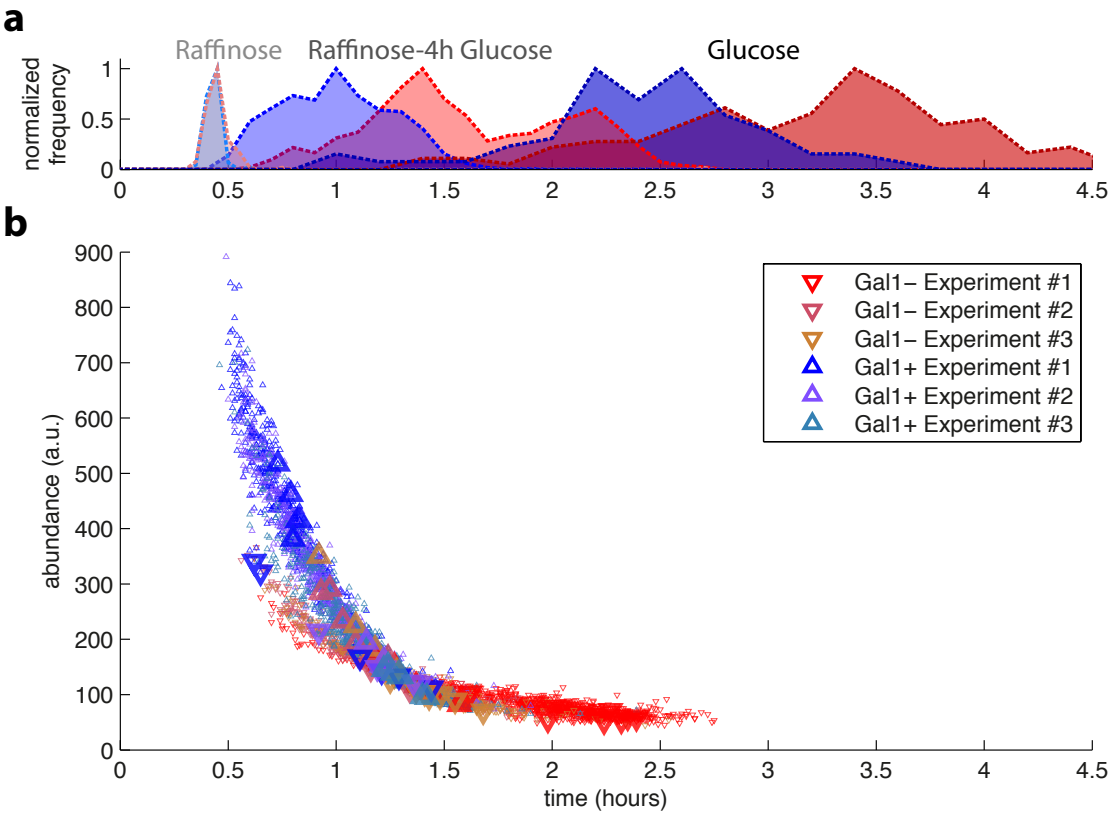

Figure 4.11: **Comparison of Gal1+ and Gal1- strains under different induction conditions. a. Comparison of induction time in glactose after different medium histories.** Gal1+ strains (blue) Gal1- (red) strains show little difference in induction timing if induced after growth in raffinose (light colors). The induction time and variance increases from 'growth in raffinose – 4h glucose' repression to 'growth in glucose' repression (dark colors). **b. Induction timing and rate comparison between experimental repeats.** Gal1+ strains (△) are found to induce on a different curve than Gal1- strains (▽). Wild-type repeats symbolized by bold triangles. Both Gal1+ and Gal1- wild-type are spotted in octuplicates in each experiment.

To summarize, Gal1+ and Gal1- strains are on separated curves, each curve highly correlated with no strain showing a remarkable deviation. This information indicates that no previously unknown gene plays a crucial role in the GAL network itself, as this would have led to a strong deviation from the curve, similar to the deletion of GAL1. Instead, we can assume that the predominant roles of all strains influencing the first induction (either by measurement of abundance or response time) are during the early stage of galactose induction.

Comparing the first induction time of our standard experimental setup (overnight raffinose to four hours glucose) to that of strains that were grown in glucose overnight we see a longer delay for the latter. This indicates that some mechanisms of glucose repression need several cell cycles to be completely effective. A possible reason could be a slightly raised Gal4p concentration. GAL4 expression in glycerol shows 5-fold higher expression as in glucose [115]. Gal4p is found to be a stable protein with a half-life comparable to the cell cycle [116], thus making a still slightly increased concentration likely. This once again indicates the sensitivity of the GAL network induction.

Screen for strains of interest

Using the aforementioned methods and thresholds, we identified numerous strains as outliers (see Table 4.1). Many of these strains have known relations to the GAL network control, validating our experiments and process. For example, deletion of MIG1 leads to faster induction, something that is expected due to its function as transcriptional repressor of GAL1. Interestingly, we found *gal83*Δ deletion, a gene that is part of GAL repression in glucose conditions [108], to be one of the slowest inducing genes. This behavior is completely diametric to its expected behavior and will need further in-depth observation.

Looking at table 4.1 and figures 4.12,4.13,4.14, we can observe several general trends. First, we find decisively more strains that show remarkably high abundance during first induction for Gal1- strains. One explanation could be the predominant role of Gal1p in induction, which conceals minor effects of other genes. Furthermore, the response times for Gal1+ strains is close to those of strains grown in raffinose. A possibility to adjust for this might be to lengthen the first repression period in glucose for Gal1+ strains.

Furthermore, we found a large percentage of strains with increased size to be fast inducer. This was the case for 27% of strains in Gal1+ and 52% of strains in Gal1-. One explanation therefore could be that at least some of the strains with increased cell size are aneuploid, even though a previous study did not find indications for galactose induction related changes in aneuploid strains [76].

Including those strains that are inconclusive outliers, five deletions were found to have a fast first induction in both backgrounds (*eaf7*Δ, *hda3*Δ, *mig1*Δ, *rtt101*Δ, *ubc4*Δ). Hda3p is a

subunit of the HDA1 histone deacetylase complex, making it likely that its deletion results in a more accessible chromatin formation. Eaf7p is a subunit of the NuA4 histone acetyltransferase complex and both RTT101 and UBC4 are involved in ubiquitin-conjugation, a reaction that targets a protein for degradation via the proteasome.

For *rtt101*Δ and *eaf7*Δ, but not for *ubc4*Δ, we can see an increased GFP abundance during the first glucose repression (Figures 4.13,4.14) and the deletion of RTT101 is found to have an increased size in both backgrounds. As we see a general correlation between GFP abundance during repression and the first induction (data not shown), it is reasonable to assume that there are strains with a slightly increased Gal1p concentration during glucose repression, consequently allowing these strains to induce faster. Interestingly, *mig1*Δ mutation is not found to have increased Gal1p concentration during the first glucose repression. This indicates that Mig1p repression is only of primary importance if glucose and galactose are available at the same time.

Numerous strains show a strong delay for both Gal1 backgrounds. Including only those strains that pass the stringent outlier filter for slow responder, 7 strains are slow for both backgrounds (*ade1*Δ, *chz1*Δ, *cti6*Δ, *gal83*Δ, *itc1*Δ, *nhp10*Δ, *pre9*Δ). *Ade1*Δ is of interest, as it is the only strain which repeatedly and for both backgrounds shows a stress pattern during glucose growth. ADE1 is required for 'de novo' purine nucleotide biosynthesis and its deletion shows decreased growth rates in both backgrounds.

CHZ1, CTI6, ITC1, NHP10 (and several other genes that were found under less strict restrictions, e.g. CHD1 and ISW2) are all related to the SWI/SNF complex, for which a deletion of SWI2 was previously found to decrease memory [20]. This gives a case to the argument, that the deletion of SWI/SNF related genes decreases already GAL network protein genes during the first induction, therefore additionally influencing cytoplasmic memory.

Pre9p forms the only non-essential part of the 20S proteasome, α3 subunit. Deletion of PRE9 leads to replacement by the α4 subunit (Pre6p) [117]. It is assumed that this creates a more active proteasome isoform. Thus, it is very well possible that a *pre9*Δ mutant has the effect of increased protein degradation. Interestingly, no strain detected as potential slow inducer was found to have increased cell size.

In comparison to the vast number of strains detected during the first induction, few strains were found to change their response for the second induction. Only one of the deletions

(*dph5*Δ) was found reliably at a place inside the distribution that would indicate memory gain in comparison to the background, an outcome that is estimated to have an 85% probability of being randomly generated. Changing the thresholds to look for more outliers with potential memory gain yielded only a low number of additional hits, while the probability of being the result of a random event always remained above 20%. The distribution of wild-type strains for our measurements was a further indication that for our experiments , outliers in this direction were likely a result of extrinsic noise (see figure 4.12).

Looking at the induction time during the second induction, it seems that many strains reach a plateau that is close to the response time of ~25 minutes in raffinose. Previous studies found that translation in galactose reinduction after four hours of glucose is faster than translation response after raffinose [20], a result that could be conceived as being contradictory to our observation. But these studies do not adjust their second induction for potential cytoplasmic memory and furthermore use a different technique. A study that uses similar reporter genes as our studies also note 'when pregrown in no glucose media, such as in raffinose media, yeast cells respond to galactose for the first time with a graded and very rapid kinetics, masking the accelerated second response after consequent growth in glucose (data not shown)' [19], indicating a similar finding. In this case, lengthening the second repression period in glucose might lead to an increase in outliers of interest.

While we are not able to detect deletion strains that indicate a gain in memory, several strains can be found for which the amount of memory is potentially reduced. For Gal1+, we found 8 deletion strains (*bre5*Δ, *eaf7*Δ, *mlh1*Δ, *rrd1*Δ, *rtt103*Δ, *sap30*Δ, *set3*Δ, *swc3*Δ) that repeatedly were detected as outliers for the slow responder threshold, a number that we estimated to have less than 0.01% probability of occurring by chance. As changes in reinduction analyzed relative to the first induction, these changes can be of very different character.

The mutations *bre5*Δ, a ubiquitin protease factor, and the aforementioned *eaf7*Δ both show increased GFP-levels during glucose repression, a fast response time during the first induction and an unexpected slow response time during the second induction (Figure 4.13). Interestingly, a similar tendency is found in Gal1- strains (Figure 4.12). A possible explanation for this behavior could be that the effect the deletion had onto the Gal1p concentration was advantageous for a fast first induction. But at the same time, the same mechanism could be either neutral or even disadvantageous during the second induction. In addition, *bre5*Δ was

found to be increased in size for both experiments.

Four mutations, *mlh1*Δ, *rrd1*Δ, *sap30*Δ and *swc3*Δ, show no increased abundance in glucose repression, a relatively normal response time for first induction and a slower response time during the second induction. Taking a closer look at the relative abundance traces (Figure 4.13), we can see that the underlying dynamics are vastly different. While the deletion strain *mlh1*Δ induces relative fast during the first induction, the level of GFP decreases more than expected during the second glucose repression, similar to *bre5*Δ. Interestingly, both Bre5p and Mlh1p are proteins that are linked to DNA damage [118, 119].

Two deletions of genes with seemingly similar trajectories, *rrd1*Δ and *sap30*Δ, are involved in very different processes. While Rrd1p is found to be involved in DNA repair and G1 phase, Sap30p is a component of Rpd3L histone deacetylase complex.

The deletion of SWC3, a component of the SWR1 chromatin-remodeling complex, is found to respond normal, albeit slightly slow during the first induction and repeatedly slower in second induction. The function of the SWR1 complex, exchanging histone variant H2AZ (Htz1p) for chromatin-bound histone H2A, could very well be a possible candidate for loss in inherited nuclear memory [120].

The last two deletions, *set3*Δ and *rtt103*Δ, are both examples for strains that show already a slow response during the first induction.

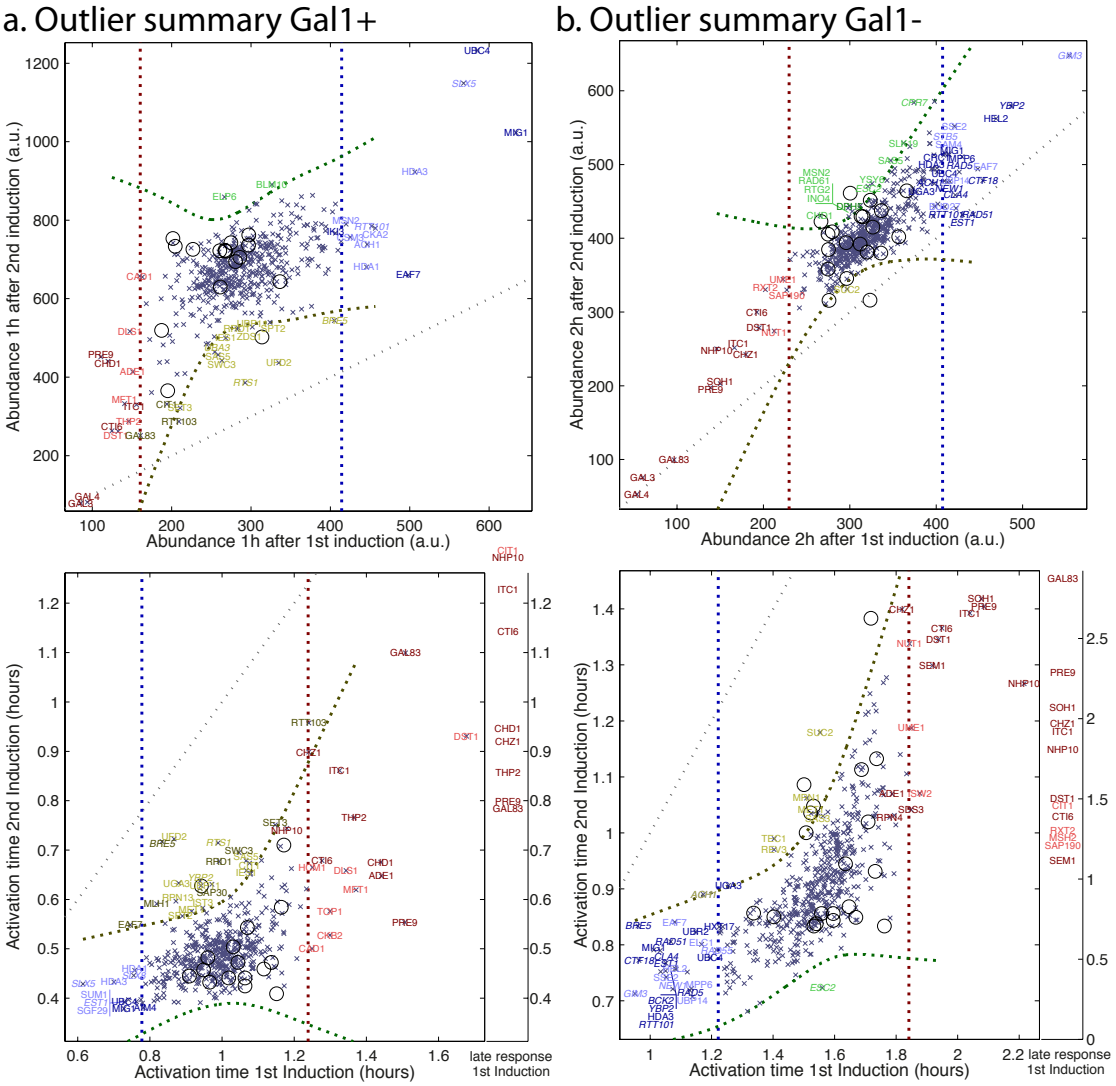

Figure 4.12: **Summary of outlier detection in (a.) Gal1+ (left) and (b.) Gal1- (right).** Light strain names indicate 'inconclusive outliers'. Strain names in italic indicate strains with increased cell size. Strain names can be slightly moved for readability. Wild type repeats are shown with black circles.

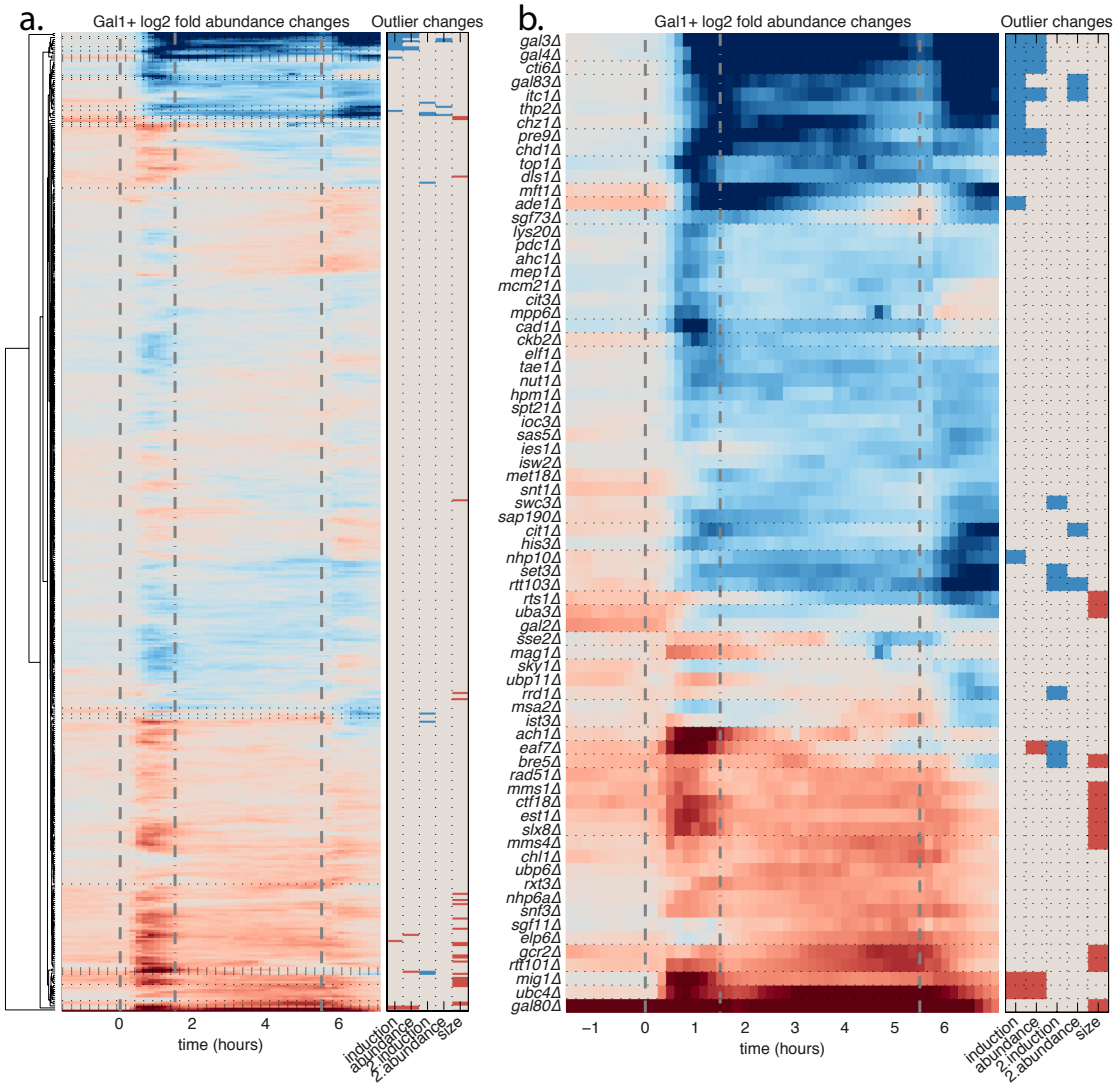

Figure 4.13: **Clustergram of Gal1+ abundance changes. a. Traces of all strains.** Fold abundance change is normalized for each experiment individually, using the medium value at each time point. Blue indicates decreased and red increased abundance. Outlier changes indicate those strains that are conclusive slow/low (blue) or fast/high outliers during the first or second induction. The last colum points out strains with increased cell size. Dotted lines show the borders of different clusters. **b. Focus on strains in Clusters of interest.** Bigger clusters were removed for clarity.

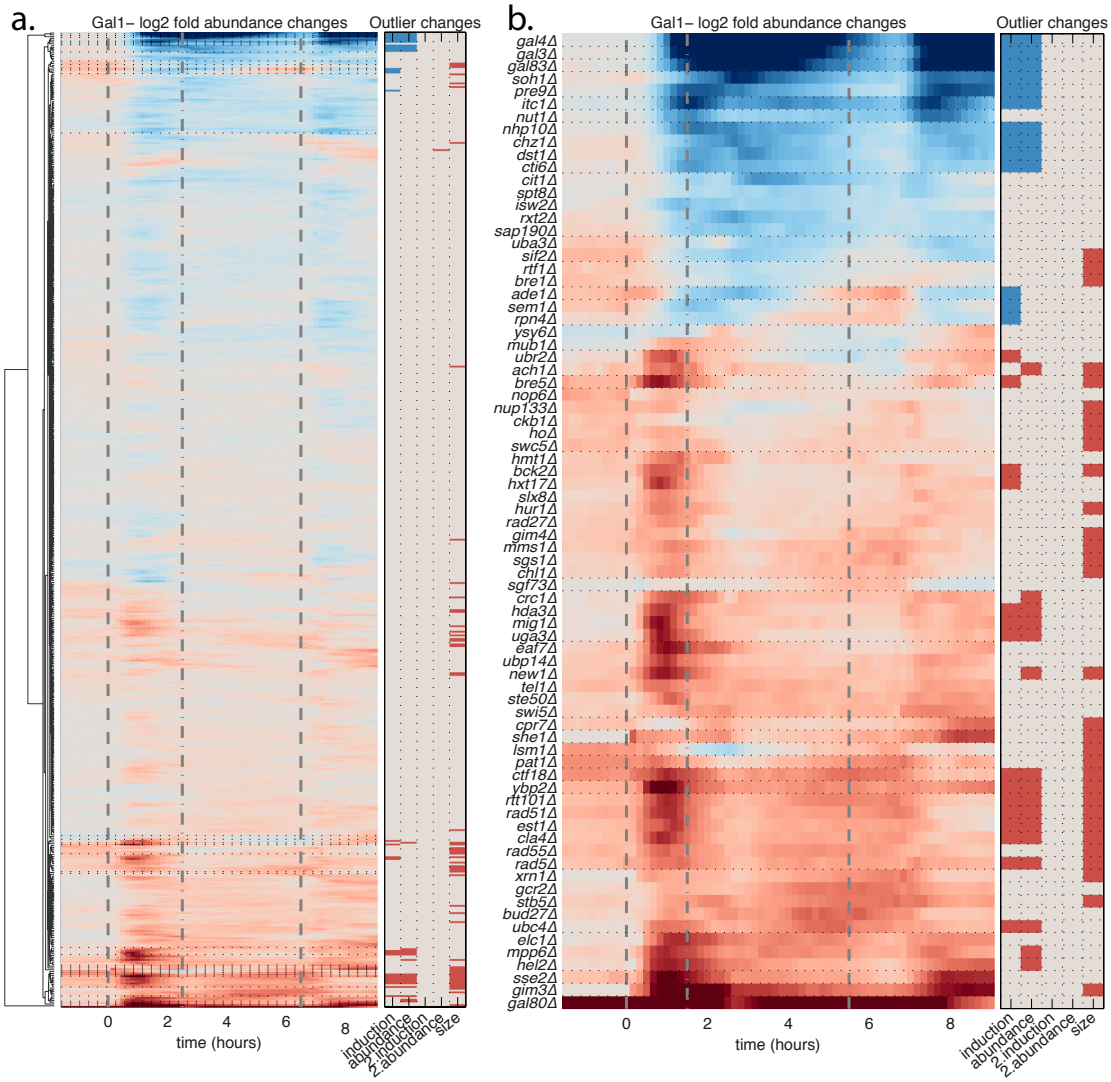

Figure 4.14: **Clustergram of Gal1- abundance changes. a. Traces of all strains.** Fold abundance change is normalized for each experiment individually, using the medium value at each time point. Blue indicates decreased and red increased abundance. Outlier changes indicate those strains that are conclusive slow/low (blue) or fast/high outliers during the first or second induction. The last colum points out strains with increased cell size. Dotted lines show the borders of different clusters. **b. Focus on strains in Clusters of interest.** Bigger clusters were removed for clarity.

5 Discussion of the results and outlook

5.1 Results overview

Our microfluidic live-cell imaging platform opens new possibilities for the study of large collections of yeast mutant strains. The aim of this thesis was to explore several of these possibilities, while focusing on approaches of quantitative analysis.

Chapter 3 describes our approach to determine protein localization. We showed that our approach of a supervised classification of six spatial patterns could be used to quantify changes in protein localization. This quantification helped to identify 111 proteins changing their localization upon MMS treatment. Furthermore, it allowed us to draw comparisons for localization changes under different conditions. For example, we could show that the formation of P-Bodies has significantly different temporal dynamics in MMS, HU and UV. Proteins involved in the MCM complex on the other hand showed very similar timing under all cell damaging conditions, indicating that mechanisms of cell cycle arrest are more context-independent.

Using two different sets of reporter-deletion mutants, we show in chapter 4 that the possibilities of our microfluidic platform can transcend those of a typical screen, adding valuable dynamic information.

Section 4.3 describes the evaluation of a set of different reporters and deletions in response to a stress condition. In this case, our analysis was used to merge information on localization and abundance for each reporter with information on phenotype for each deletion. We highlight the likely relationship between the localization of Rnr4p and the respective phenotype of a

strain, as we found that all strains for which Rnr4p does not re-localize under UV irradiation are lethal. The relationship between the formation of P-Bodies and a growth related phenotype on the other hand was found to be indirect, as strains with defective P-Body formation were still found to be viable. This illustrates the value of our capability to simultaneously measure protein- and phenotype-related information.

Section 4.4 underlines the possibilities of our microfluidic device for new approaches in high-throughput screening. For this study, the adequate temporal resolution provided by our platform, as well as the temporal continuity was of importance. While not genome-wide, this screen is still broad enough to outline a general map of genes that are involved in galactose induction. Furthermore, it provides a distinct number of strains that show a reinduction-specific delay and could play an important role in epigenetics.

We can argue that our microfluidic device is in some regards better suited for the evaluation of complex mechanisms like transcriptional reinitiation than conventional low-throughput approaches. For example, the high number of strains gave us a comprehensive distribution of responses during the first induction, which allowed us to get relative fix points for all strains during the second induction. This is not possible for low throughput experiments, were it can be convoluted or impossible to estimate the effects that small changes during the first induction are expected to have on the second induction.

These results show the diverse aspects of systems biology for which our platform and analysis can obtain valuable results. Nonetheless, both the experimental and analytical parts of our work are still facing limitations.

5.2 Limitations and improvements

Both data analysis and setup of our microfluidic device are still subject to limitations. The limitations of device robustness, robustness of cell recovery, sensitivity and single-cell tracking have been mentioned previously [32]. Especially the robustness of the setup and strain recovery after spotting requires a well-trained operator. Therefore, quality of results can differ between investigators.

Differences in growth and protein expression between flow lines was perceived to be an increased problem during the experiments with galactose, which may be due to the complete

arrest of cell growth. Furthermore, a change in medium and cell cycle arrest can change background intensity, something for which an adjustment can be non-trivial.
The novelty of our platform is not readily approachable with existing statistics. Together with the vast amount of data extracted during an experiment, it can become time consuming to extract informative values with adequate statistical significance. Our approach to find a good combination of extracted data and its analysis was highly iterative. Further investigation will be necessary to assure that all steps of the analysis are as robust and sensitive as possible. However, this work will significantly reduce the effort for future applications.
A caveat that holds true for high throughput technologies in general is the question of genetic integrity of a library of modified strains. However, our device is well capable of highlighting potential candidates and can even indicate further biological causation. This causation can be of great value in comparison to other screening techniques. For example, the integrated information about cell growth allows us to directly evaluate if an increase in protein abundance is a direct or indirect result of a deletion. Nonetheless, it is not capable of replacing traditional biological techniques to confirm these results.

5.3 Outlook

As mentioned in the previous section, further investigations will need to be done to achieve a more universal deployable tool for data analysis. Data handling, analysis, and visualization, which are currently applied independently, need to be further evaluated and combined.
There is no limit of potentially interesting experiments for our microfluidic device. But their generation and analysis is time-consuming and it is therefore advantageous to focus the effort on those experiments that make an exhaustive use of the capabilities of our device. Our device was found to operate optimally for the observation of around 50 to 576 strains, a range that makes it necessary to use high-throughput technologies, but still allows for duplicates during the same experiment.
Furthermore, it is important to consider the time interval of 20 minutes and potential small differences between flow lines. Therefore, a condition change should not cause a too strong reaction to allow for the detection of more subtle deletion specific differences. At the same time it should not be too subtle to avoid uncertainties due to extrinsic noise. For example, the

use of UV irradiation had the disadvantage that P-Body formation occurred so rapidly that it was impossible to detect more elaborate strain-to-strain differences.

One interesting application for a more detailed observation of P-Body formation could be the use of a pulse width flow gradient generator, which was developed by Sylvain Bernard in the Maerkl lab. Using this gradient generator to form a stepwise gradient between standard medium and nutrient-starvation would allow for a good temporal control of P-Body network response. It would also be advantageous to include reporter-deletion constructs of genes known to be degraded during starvation. This could lead to a deeper insight into the influence of P-Body related gene deletions.

The response dynamics of the GAL network are well suited for our microfluidic device. Important for the further investigation of the GAL network is the decoupling of different aspects of induction response time. For example, histone modifications precede the increase in Gal1p and cytoplasmic memory, making it likely that nuclear memory forms before cytoplasmic memory. In addition, cytoplasmic memory and nuclear memory are expected to behave differently. Cytoplasmic memory relies on measurable increase of protein concentrations, something that can be well described in differential equations. Nuclear memory on the other hand is expected to be inherently stochastic.

Therefore, an optimal experimental setup would be to compare the following two induction responses:

Setup 1:	
2% raffinose	stationary growth
1.5% galactose, 1.5% raffinose	<20min
2% glucose	Xh
1.5% galactose, 1.5% raffinose	until galactose steady state is reached

Setup 2:	
2% raffinose	stationary growth
2% glucose	Xh
1.5% galactose, 1.5% raffinose	until galactose steady state is reached

This setup is advantageous, as galactose induction is found to respond very homogeneous. This would allow us to achieve a uniform induction of all cells and most deletions, simultaneously decreasing the amount of cytoplasmic memory. Preliminary results suggest that this

holds true for Gal1+ and Gal1- strains, which are known to differ in cytoplasmic memory, but not expected to differ in nuclear memory. Varying the time intervals for first induction and repression time in a setup where the cytoplasmic memory is not predominant can help to separate the aforementioned differences in memory characteristics.

A List of features for the classification of protein localization

Table A.1: **List of features.**

#	FEATURE TAG	DESCRIPTION
Histogram-based features		
1	top5vs20	mean(highest 5 pixels) / mean(highest 20 pixels)
2	top5vs50	mean(highest 5 pixels) / mean(highest 50 pixels)
3	top20vs50	mean(highest 20 pixels) / mean(highest 50 pixels)
4	top5vsMed	mean(highest 5 pixels) / median
5	top20vsMed	mean(highest 20 pixels) / median
6	top50vsMed	mean(highest 50 pixels) / median
7	histo1ratio	frequency of highest pixel bin (pixel values 240-255) / frequency of bottom half (pixel values 0-127)
8	histo2ratio	frequency of 2nd highest pixel bin (pixel values 224-239) / frequency of bottom half (pixel values 0-127)
9	histo3ratio	frequency of 3rd highest pixel bin (pixel values 208-223) / frequency of bottom half (pixel values 0-127)
10	histoHLratio	frequency of top half (pixel values 128-255) / frequency of bottom half (pixel values 0-127)
11	bin1vs2	93.75th percentile / 87.5th percentile
12	bin1vs3	93.75th percentile / 81.25th percentile

13	bin2vs3	87.5th percentile / 81.25th percentile
14	bin1vsMed	93.75th percentile / median
15	bin2vsMed	87.5th percentile / median
16	bin3vsMed	81.25th percentile / median
17	binHLratio	Upper quartile / lower quartile

Spatial distribution features

18	central_signal	mean($P(x \leq x_i, y \leq y_i)$) / total mean, ($x_i$ = cos(t)·3·cell width/12, y_i = sin(t)·3·cell height/12)
19	middle signal	mean($P(x_i < x \leq x_m, y_i < y \leq y_m)$) / total mean, ($x_m$ = cos(t)·5·cell width/12, y_m = sin(t)·3·cell height/12)
20	boundary_signal	mean($P(x_m < x \leq x_b, y_m < y \leq y_b)$) / total mean, ($x_b$ = cos(t)·7·cell width/12, y_b = sin(t)·7·cell height/12)

Morphological features [44, 51]

21	convex_hull _overlap	SLF 1.14, Convex hull area / cell area (with binary threshold at 0.5·P_{max})
22	convex_hull _roundness	SLF 1.15, The roundness of the convex hull (with binary threshold at 0.5·P_{max})
23	edges_fraction	SLF 1.9, The fraction of the nonzero pixels that are along an edge (with binary threshold at 0.5·P_{max})
24	edges_homogeneity	SLF1.10, Measure of edge gradient intensity homogeneity
25	edges_direction _homogeneity1	SLF1.11, Measure of edge direction homogeneity 1
26	edges_direction _homogeneity2	SLF1.12, Measure of edge direction homogeneity 2
27	edges_direction _difference	SLF1.13, Measure of edge direction difference

Granulometries [52]

28	gray_open_1	mean intensity of ($I - IO^{d1}$), IO^{dr} = grayscale opening of image I with disk of radius r

29	gray_open_2	mean intensity of $(IO^{d1} - IO^{d2})$
30	gray_open_4	mean intensity of $(IO^{d2} - IO^{d4})$
31	gray_open_7	mean intensity of $(IO^{d4} - IO^{d7})$
32	gray_open_12	mean intensity of $(IO^{d7} - IO^{d12})$
33	gray_close_1	mean intensity of $(I - IC^{d1})$, IC^{dr} = grayscale closing of image I with disk of radius r
34	gray_close_2	mean intensity of $(IC^{d1} - IC^{d2})$
35	gray_close_4	mean intensity of $(IC^{d2} - IC^{d4})$
36	gray_close_7	mean intensity of $(IC^{d4} - IC^{d7})$
37	gray_close_12	mean intensity of $(IC^{d7} - IC^{d12})$
Threashold adjacencies statistics (TAS) [53]		
38-46	tas_T35_pk	Threshold at 0.35·P_{max}, pixel count with k neighbor above threshold / pixel count above threshold
47	tas_T35_binRatio	Threshold at 0.35·P_{max}, pixel count above threshold / total pixel count
48-56	tas_T50_pk	Threshold at 0.5·P_{max}, pixel count with k neighbor above threshold / pixel count above threshold
57	tas_T50_binRatio	Threshold at 0.5·P_{max}, pixel count above threshold / total pixel count
58-67	tas_T65_pk	Threshold at 0.65·P_{max}, pixel count with k neighbor above threshold / pixel count above threshold
67	tas_T65_binRatio	Threshold at 0.65·P_{max}, pixel count above threshold / total pixel count
Threashold adjacencies statistics (TAS) - inverted image		
68-97	tas_inv_Txx_pk	Same as 38-67with inverted image, $P_{inv}(x,y)$ = P_{max}-$P(x,y)$

Bibliography

[1] Ideker, T. & Krogan, N. J. Differential network biology. *Molecular systems biology* **8**, 565 (2012).

[2] Spellman, P. T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell* **9**, 3273–97 (1998).

[3] Gasch, a. P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell* **11**, 4241–57 (2000).

[4] Tarassov, K. *et al.* An in vivo map of the yeast protein interactome. *Science* **320**, 1465–70 (2008).

[5] Huh, W.-K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–91 (2003).

[6] Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–41 (2003).

[7] Newman, J. R. S. *et al.* Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441**, 840–6 (2006).

[8] Cai, L., Dalal, C. K. & Elowitz, M. B. Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature* **455**, 485–90 (2008).

[9] Di Talia, S., Skotheim, J. M., Bean, J. M., Siggia, E. D. & Cross, F. R. The effects of molecular noise and size control on variability in the budding yeast cell cycle. *Nature* **448**, 947–51 (2007).

[10] Tkach, J. M. *et al.* Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. *Nature cell biology* **14**, 966–76 (2012).

[11] Breker, M., Gymrek, M. & Schuldiner, M. A novel single-cell screening platform reveals proteome plasticity during yeast stress responses. *The Journal of Cell Biology* **200**, 839–850 (2013).

[12] Bennett, M. R. *et al.* Metabolic gene regulation in a dynamically changing environment. *Nature* **454**, 1119–22 (2008).

[13] Taylor, R. J. *et al.* Dynamic analysis of MAPK signaling using a high-throughput microfluidic single-cell imaging platform. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 3758–63 (2009).

[14] Dénervaud, N. *et al.* A chemostat array enables the spatio-temporal analysis of the yeast proteome. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 15842–7 (2013).

[15] Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology* **2**, 2006.0008 (2006).

[16] Kim, D.-U. *et al.* Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. *Nature biotechnology* **28**, 617–23 (2010).

[17] Winzeler, E. a. *et al.* Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901–6 (1999).

[18] Tong, a. H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–8 (2001).

[19] Zacharioudakis, I., Gligoris, T. & Tzamarias, D. A yeast catabolic enzyme controls transcriptional memory. *Current biology : CB* **17**, 2041–6 (2007).

[20] Kundu, S., Horn, P. J. & Peterson, C. L. SWI/SNF is required for transcriptional memory at the yeast GAL gene cluster. *Genes & development* **21**, 997–1004 (2007).

[21] Hong, J. W. & Quake, S. R. Integrated nanoliter systems. *Nature biotechnology* **21**, 1179–83 (2003).

[22] Whitesides, G. M. The origins and the future of microfluidics. *Nature* **442**, 368–73 (2006).

[23] McDonald, J. C. *et al.* Fabrication of microfluidic systems in poly(dimethylsiloxane). *Electrophoresis* **21**, 27–40 (2000).

[24] Unger, M. a. Monolithic Microfabricated Valves and Pumps by Multilayer Soft Lithography. *Science* **288**, 113–116 (2000).

[25] Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–6 (2002).

[26] Suter, D. M. *et al.* Mammalian genes are transcribed with widely different bursting kinetics. *Science* **332**, 472–4 (2011).

[27] Di Carlo, D., Wu, L. Y. & Lee, L. P. Dynamic single cell culture array. *Lab on a chip* **6**, 1445–9 (2006).

[28] Cookson, S., Ostroff, N., Pang, W. L., Volfson, D. & Hasty, J. Monitoring dynamics of single-cell gene expression over multiple cell cycles. *Molecular systems biology* **1**, 2005.0024 (2005).

[29] Ryley, J. & Pereira-Smith, O. M. Microfluidics device for single cell gene expression analysis in Saccharomyces cerevisiae. *Yeast (Chichester, England)* **23**, 1065–73 (2006).

[30] Maerkl, S. J. & Quake, S. R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233–7 (2007).

[31] Narayanaswamy, R. *et al.* Systematic profiling of cellular phenotypes with spotted cell microarrays reveals mating-pheromone response genes. *Genome biology* **7**, R6 (2006).

[32] Dénervaud, N. A Microfluidic Live-cell Imaging Platform to Study Large Collections of Microbial Genotypes (2012).

[33] Whitesides, G. M., Ostuni, E., Takayama, S., Jiang, X. & Ingber, D. E. Soft lithography in biology and biochemistry. *Annual review of biomedical engineering* **3**, 335–73 (2001).

[34] Thorsen, T., Maerkl, S. J. & Quake, S. R. Microfluidic large-scale integration. *Science* **298**, 580–4 (2002).

[35] Danuser, G. Computer vision in cell biology. *Cell* **147**, 973–8 (2011).

[36] Bush, A., Chernomoretz, A., Yu, R., Gordon, A. & Colman-Lerner, A. Using Cell-ID 1.4 with R for microscope-based cytometry. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* **Chapter 14**, Unit 14.18 (2012).

[37] Carpenter, A. E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology* **7**, R100 (2006).

[38] Bengtsson, E. & Wahlby, C. Robust cell image segmentation methods. *Pattern Recognition and Image Analysis* **14**, 157–167 (2004).

[39] Meijering, E. Cell Segmentation: 50 Years Down the Road **29**, 140–145 (2012).

[40] Delgado-Gonzalo, R., Denervaud, N., Maerkl, S. & Unser, M. Multi-target tracking of packed yeast cells. In *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, vol. 24, 544–547 (IEEE, 2010).

[41] Thévenaz, P., Delgado-Gonzalo, R. & Unser, M. The ovuscule. *IEEE transactions on pattern analysis and machine intelligence* **33**, 382–93 (2011).

[42] Picotti, P., Bodenmiller, B., Mueller, L. N. & Domon, B. Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* **138**, 795–806 (2009).

[43] Boland, M. V., Markey, M. K. & Murphy, R. F. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry* **33**, 366–375 (1998).

[44] Boland, M. V. & Murphy, R. F. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics* **17**, 1213–1223 (2001).

[45] Nanni, L. & Lumini, A. A reliable method for cell phenotype image classification. *Artificial intelligence in medicine* **43**, 87–97 (2008).

[46] Chen, S.-C., Zhao, T., Gordon, G. J. & Murphy, R. F. Automated image analysis of protein localization in budding yeast. *Bioinformatics (Oxford, England)* **23**, i66–71 (2007).

[47] Huh, S., Lee, D. & Murphy, R. F. Efficient framework for automated classification of subcellular patterns in budding yeast. *Cytometry. Part A : the journal of the International Society for Analytical Cytology* **75**, 934–40 (2009).

[48] Handfield, L.-F., Chong, Y. T., Simmons, J., Andrews, B. J. & Moses, A. M. Unsupervised Clustering of Subcellular Protein Expression Patterns in High-Throughput Microscopy Images Reveals Protein Complexes and Functional Relationships between Proteins. *PLoS Computational Biology* **9**, e1003085 (2013).

[49] Peng, T. *et al.* Determining the distribution of probes between different subcellular locations through automated unmixing of subcellular patterns. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 2944–9 (2010).

[50] Rajkumar, A. S., Dénervaud, N. & Maerkl, S. J. Mapping the fine structure of a eukaryotic promoter input-output function. *Nature genetics* **45**, 1207–15 (2013).

[51] Huang, K. & Murphy, R. F. From quantitative microscopy to automated image understanding. *Journal of biomedical optics* **9**, 893–912 (2004).

[52] Walter, T. *et al.* Automatic identification and clustering of chromosome phenotypes in a genome wide RNAi screen by time-lapse imaging. *J Struct Biol* 1–9 (2010).

[53] Hamilton, N. A., Pantelic, R. S., Hanson, K. & Teasdale, R. D. Fast automated cell phenotype image classification. *BMC Bioinformatics* **8**, 110 (2007).

[54] Hastie, T., Tibshirani, R. & Friedman, J. The Elements of Statistical Learning **27**, 83–85 (2001).

[55] Minka, T. P. Estimating a Dirichlet distribution. *Annals of Physics* **2000**, 1–13 (2003).

[56] Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society* **35**, 99–109 (1943).

[57] Rauber, T. W., Braun, T. & Berns, K. Probabilistic distance measures of the Dirichlet and Beta distributions. *Pattern Recognition* **41**, 637–645 (2008).

[58] Hao, N. & O'Shea, E. K. Signal-dependent dynamics of transcription factor translocation controls gene expression. *Nature structural & molecular biology* **19**, 31–9 (2012).

[59] Sbia, M. *et al.* Regulation of the yeast Ace2 transcription factor during the cell cycle. *The Journal of biological chemistry* **283**, 11135–45 (2008).

[60] Buchan, J. R., Nissan, T. & Parker, R. Analyzing P-bodies and stress granules in Saccharomyces cerevisiae. *Methods in enzymology* **470**, 619–40 (2010).

[61] Jung, J.-H. & Kim, J. Accumulation of P-bodies in Candida albicans under different stress and filamentous growth conditions. *Fungal genetics and biology : FG & B* **48**, 1116–23 (2011).

[62] Thumm, M. Structure and function of the yeast vacuole and its role in autophagy. *Microscopy research and technique* **51**, 563–72 (2000).

[63] Parker, R. RNA degradation in Saccharomyces cerevisae. *Genetics* **191**, 671–702 (2012).

[64] Mitchell, S. F., Jain, S., She, M. & Parker, R. Global analysis of yeast mRNPs. *Nature structural & molecular biology* **20**, 127–33 (2013).

[65] Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**, 637–43 (2006).

[66] Pu, S., Wong, J., Turner, B., Cho, E. & Wodak, S. J. Up-to-date catalogues of yeast protein complexes. *Nucleic acids research* **37**, 825–31 (2009).

[67] Kim, T.-Y., Ha, C. W. & Huh, W.-K. Differential subcellular localization of ribosomal protein L7 paralogs in Saccharomyces cerevisiae. *Molecules and cells* **27**, 539–46 (2009).

[68] Hao, N. *et al.* Regulation of cell signaling dynamics by the protein kinase-scaffold Ste5. *Molecular cell* **30**, 649–56 (2008).

[69] Blomberg, A. Measuring growth rate in high-throughput growth phenotyping. *Current opinion in biotechnology* **22**, 94–102 (2011).

[70] Nobs, J.-B. & Maerkl, S. J. Long-term single cell analysis of S. pombe on a microfluidic microchemostat array. *PloS one* **9**, e93466 (2014).

[71] Jorgensen, P., Nishikawa, J. L., Breitkreutz, B.-J. & Tyers, M. Systematic identification of pathways that couple cell growth and division in yeast. *Science* **297**, 395–400 (2002).

[72] Hillenmeyer, M. E. *et al.* The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **320**, 362–5 (2008).

[73] Tong, A. H. Y. *et al.* Global mapping of the yeast genetic interaction network. *Science* **303**, 808–13 (2004).

[74] Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–31 (2010).

[75] Giaever, G. & Nislow, C. The Yeast Deletion Collection: A Decade of Functional Genomics. *Genetics* **197**, 451–465 (2014).

[76] Hughes, T. R. *et al.* Widespread aneuploidy revealed by DNA microarray expression profiling. *Nature genetics* **25**, 333–7 (2000).

[77] Zhang, J. *et al.* Genomic Scale Mutant Hunt Identifies Cell Size Homeostasis Genes in S. cerevisiae. *Current Biology* **12**, 1992–2001 (2002).

[78] Fátima Vaz, M. & Fortes, M. Grain size distribution: The lognormal and the gamma distribution functions. *Scripta Metallurgica* **22**, 35–40 (1988).

[79] Cagney, G. *et al.* Functional genomics of the yeast DNA-damage response. *Genome biology* **7**, 233 (2006).

[80] Cadet, J. & Douki, T. Oxidatively generated damage to DNA by UVA radiation in cells and human skin. *The Journal of investigative dermatology* **131**, 1005–7 (2011).

[81] Dahle, J. & Kvam, E. Induction of delayed mutations and chromosomal instability in fibroblasts after UVA-, UVB-, and X-radiation. *Cancer research* **63**, 1464–9 (2003).

[82] Eulalio, A., Behm-Ansmant, I. & Izaurralde, E. P bodies: at the crossroads of post-transcriptional pathways. *Nature reviews. Molecular cell biology* **8**, 9–22 (2007).

[83] Teixeira, D. & Parker, R. Analysis of P-body assembly in Saccharomyces cerevisiae. *Molecular biology of the cell* **18**, 2274–2287 (2007).

[84] Parker, R. & Sheth, U. P bodies and the control of mRNA translation and degradation. *Molecular cell* **25**, 635–46 (2007).

[85] Eulalio, A., Behm-Ansmant, I., Schweizer, D. & Izaurralde, E. P-body formation is a consequence, not the cause, of RNA-mediated gene silencing. *Molecular and cellular biology* **27**, 3970–81 (2007).

[86] van Hemert, M. J., Steensma, H. Y. & van Heusden, G. P. 14-3-3 proteins: key regulators of cell division, signalling and apoptosis. *BioEssays : news and reviews in molecular, cellular and developmental biology* **23**, 936–46 (2001).

[87] Morano, K. a., Grant, C. M. & Moye-Rowley, W. S. The response to heat shock and oxidative stress in Saccharomyces cerevisiae. *Genetics* **190**, 1157–95 (2012).

[88] Yao, R. *et al.* Subcellular localization of yeast ribonucleotide reductase regulated by the DNA replication and damage checkpoint pathways. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 6628–33 (2003).

[89] Elledge, S. J. Identification of RNR4 , encoding a second essential small subunit of ribonucleotide reductase in Saccharomyces cerevisiae . Identification of RNR4 , Encoding a Second Essential Small Subunit of Ribonucleotide Reductase in Saccharomyces cerevisiae. *Microbiology* **17** (1997).

[90] Megee, P. C., Morgan, B. a. & Smith, M. M. Histone H4 and the maintenance of genome integrity. *Genes & Development* **9**, 1716–1727 (1995).

[91] Prasanth, S. G., Méndez, J., Prasanth, K. V. & Stillman, B. Dynamics of pre-replication complex proteins during the cell division cycle. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **359**, 7–16 (2004).

[92] Mutvei, A., Dihlmann, S., Herth, W. & Hurt, E. C. NSP1 depletion in yeast affects nuclear pore formation and nuclear accumulation. *European journal of cell biology* **59**, 280–295 (1992).

[93] Tong, A. H. Y. & Boone, C. Synthetic genetic array analysis in Saccharomyces cerevisiae. *Methods in molecular biology (Clifton, N.J.)* **313**, 171–192 (2006).

[94] Birrell, G. W., Giaever, G., Chu, a. M., Davis, R. W. & Brown, J. M. A genome-wide screen in Saccharomyces cerevisiae for genes affecting UV radiation sensitivity. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 12608–13 (2001).

[95] Bennett, C. B. *et al.* Genes required for ionizing radiation resistance in yeast. *Nature genetics* **29**, 426–34 (2001).

[96] Weinert, T. & Hartwell, L. The RAD9 gene controls the cell cycle response to DNA damage in Saccharomyces cerevisiae. *Science* **241**, 317–322 (1988).

[97] Franks, T. M. & Lykke-Andersen, J. The control of mRNA decapping and P-body formation. *Molecular cell* **32**, 605–15 (2008).

[98] Noueiry, A. O., Diez, J., Falk, S. P., Chen, J. & Ahlquist, P. Yeast Lsm1p-7p/Pat1p deadenylation-dependent mRNA-decapping factors are required for brome mosaic virus genomic RNA translation. *Molecular and cellular biology* **23**, 4094–106 (2003).

[99] Mulder, K. W., Winkler, G. S. & Timmers, H. T. M. DNA damage and replication stress induced transcription of RNR genes is dependent on the Ccr4-Not complex. *Nucleic acids research* **33**, 6384–92 (2005).

[100] Lee, Y. D. & Elledge, S. J. Control of ribonucleotide reductase localization through an anchoring mechanism involving Wtm1. *Genes & development* **20**, 334–44 (2006).

[101] Tang, H.-M. V., Siu, K.-L., Wong, C.-M. & Jin, D.-Y. Loss of yeast peroxiredoxin Tsa1p induces genome instability through activation of the DNA damage checkpoint and elevation of dNTP levels. *PLoS genetics* **5**, e1000697 (2009).

[102] Zaidi, I. W. *et al.* Rtt101 and Mms1 in budding yeast form a CUL4(DDB1)-like ubiquitin ligase that promotes replication through damaged DNA. *EMBO reports* **9**, 1034–40 (2008).

[103] Mir, S. S., Fiedler, D. & Cashikar, A. G. Ssd1 is required for thermotolerance and Hsp104-mediated protein disaggregation in Saccharomyces cerevisiae. *Molecular and cellular biology* **29**, 187–200 (2009).

[104] Lawrence, C. L., Botting, C. H., Antrobus, R. & Coote, P. J. Evidence of a new role for the high-osmolarity glycerol mitogen-activated protein kinase pathway in yeast: regulating adaptation to citric acid stress. *Molecular and cellular biology* **24**, 3307–23 (2004).

[105] Berg, J. M., Tymoczko, J. L. & Stryer, L. Biochemistry. 5th edition. In *Biochemistry textbook*, 1120 (2006).

[106] Frey, P. A. The Leloir pathway: a mechanistic imperative for three enzymes to change the stereochemical configuration of a single carbon in galactose. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **10**, 461–70 (1996).

[107] Stockwell, S. R., Landry, C. R. & Rifkin, S. a. The yeast galactose network as a quantitative model for cellular memory. *Mol. BioSyst.* (2014).

[108] Flick, J. S. & Johnston, M. Two systems of glucose repression of the GAL1 promoter in Saccharomyces cerevisiae. *Molecular and cellular biology* **10**, 4757–69 (1990).

[109] Brownell, J. & Allis, C. Special HATs for special occasions: linking histone acetylation to chromatin assembly and gene activation. *Current opinion in genetics & development* 176–184 (1996).

[110] Lee, D. Y., Hayes, J. J., Pruss, D. & Wolffe, A. P. A positive role for histone acetylation in transcription factor access to nucleosomal DNA. *Cell* **72**, 73–84 (1993).

[111] Belle, A., Tanay, A., Bitincka, L., Shamir, R. & O'Shea, E. K. Quantification of protein half-lives in the budding yeast proteome. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 13004–9 (2006).

[112] Kaufmann, B. B., Yang, Q., Mettetal, J. T. & van Oudenaarden, A. Heritable stochastic switching revealed by single-cell genealogy. *PLoS biology* **5**, e239 (2007).

[113] Lambert, G. & Kussel, E. Memory and Fitness Optimization of Bacteria under Fluctuating Environments. *PLoS genetics* **10**, e1004556 (2014).

[114] Venturelli, O. S., El-Samad, H. & Murray, R. M. Synergistic dual positive feedback loops established by molecular sequestration generate robust bimodal response. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E3324–33 (2012).

[115] Griggs, D. W. & Johnston, M. Regulated expression of the GAL4 activator gene in yeast provides a sensitive genetic switch for glucose repression. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 8597–601 (1991).

[116] Nalley, K., Johnston, S. A. & Kodadek, T. Proteolytic turnover of the Gal4 transcription factor is not required for function in vivo. *Nature* **442**, 1054–7 (2006).

[117] Velichutina, I., Connerly, P. L., Arendt, C. S., Li, X. & Hochstrasser, M. Plasticity in eucaryotic 20S proteasome ring assembly revealed by a subunit deletion in yeast. *The EMBO journal* **23**, 500–10 (2004).

[118] Bilsland, E., Hult, M., Bell, S. D., Sunnerhagen, P. & Downs, J. a. The Bre5/Ubp3 ubiquitin protease complex from budding yeast contributes to the cellular response to DNA damage. *DNA repair* **6**, 1471–84 (2007).

[119] Strand, M., Prolla, T. A., Liskay, R. M. & Petes, T. D. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* **365**, 274–6 (1993).

[120] Halley, J. E., Kaplan, T., Wang, A. Y., Kobor, M. S. & Rine, J. Roles for H2A.Z and its acetylation in GAL1 transcription and gene induction, but not GAL1-transcriptional memory. *PLoS biology* **8**, e1000401 (2010).

Curriculum Vitae

Johannes Becker

Born April 19th, 1983

EPFL SV IBI-SV UPNAE
AAB 0 41 (Bâtiment AAB)
Station 15
CH-1015 Lausanne

Phone: +41 21 69 37204 or +41 78 6967 967

Mail: Johannes.becker@epfl.ch

Education

09.2010-ongoing	École Polytechnique Fédérale de Lausanne (EPFL), PhD in Computational Biology
04.2010-06.2010	École Polytechnique Fédérale de Lausanne, Internship in Computational Biology
10.2003-07.2009	Technische Universität Darmstadt, degree Diplom-Mathematiker (comparable to Masters degree in Applied Mathematics) Diploma thesis “Global error control for semi-explicit differential algebraic equations of index 1”
08.1992-05.2002	Archigymnasium Soest, degree Abitur

Employment History

Tutor for bachelor and master students

2010-2014	mathematical and computational modelling in biology, EPFL
10.2008-02.2009	introduction to numerical mathematics, Department of Mathematics
10.2007-01.2008	differential equations, Department of Mathematics linear algebra I, Department of Mathematics
04.2007-07.2007 04.2008-07.2008	numerical analysis for engineers, Department of Mechanical & Engineering

Miscellaneous

07.2007-09.2007	French workshop: kitchen assistant, restaurant ’Le Sorbier’, Périgueux
09.2002-06.2003	civilian service, assisting blind and visually handicapped people, Von-Vincke-Schule Soest

Skills and Qualifications

- Language: Fluency in English, French and German
- Operating Systems: Knowledge of common Windows, Unix and Macintosh environments
- Programming: Regular use of Matlab and Python; hands-on experience in C++, Java, Perl, R
- experienced with writing documents in markup languages (i.e. HTML, LaTex)
- Data Analysis Related Methods: Regression analysis, Bayesian probability, Clustering, Classification, numerical estimation of differential equations, data visualization, image analysis, broad spectrum of methods for mathematical

optimization
- Able to adapt quickly to new situations and problems, with the capacity to find satisfying solutions
- Ability to explain a wide range of ideas and topics to people from various backgrounds

Hobbies

- Playing organized team sports: Football, Basketball. Canoe Polo, Underwater rugby
- Blogging about data in sports: http://www.sportstribution.blogspot.com

Publication

N. Dénervaud, J. Becker**,** R. Delgado-Gonzalo, P. Damay, A. S. Rajkumar, M. Unser, D. Shore, F. Naef, S.J. Maerkl,
"A Massively Parallel Microchemostat Array Enables the Spatio-Temporal Analysis of the Yeast Proteome on the Single Cell Level", published in *Proceedings of the National Academy of Sciences*, 09.2013