

Establishing a robotic high throughput measurement system for dissection of transcriptional networks in yeast

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

By

Jenia Gutin

Supervised by Prof. Nir Friedman

September 2013

The School of Computer Science and Engineering
The Hebrew University of Jerusalem, Israel

Abstract

Understanding the transcriptional regulation networks that control the gene expression programs in a living cell is one of fundamental questions in molecular biology. Using single perturbations allows us to identify the key players in those networks. To better understand the structure, we can use double perturbations to measure genetic interactions between the components in the network. By systematically analyzing all the genetic interactions between the genes in the network we can reconstruct it and to characterize the relations between its components. However, such a systematic approach is challenging because the number of pair-wise interactions grows at the square of the number of genes. Thus, a high-throughput measurement system is needed.

In this work I describe the establishment of a robotic high throughput measurement system for dissection of transcriptional networks in the budding yeast. We adapted a novel robotic high throughput system for manipulating yeast strains in multi-well microtiter plates (in either 96 or 384 well format). We developed and calibrated a robotic growth protocol that allows us to grow the strains on such plates under optimal and homogeneous growth conditions and brings them to a pre-determined density range at the beginning of the experiment. The cells are exposed to arrays of stresses, and then the transcriptional response of the entire plate is monitored in a single cell level by an automated flow cytometry measurement system. We developed an algorithm for partitioning events from the flow cytometry data stream of the whole plate to their correct wells.

We used the system to perform a systematic inspection of the budding yeast general stress response network. A generic response to stress that is mediated by two transcription factors, Msn2 and Msn4, which regulate the transcription of hundreds of genes. We identified the main genes playing a role in this network and evaluated the effect of single and double perturbations in the network on Msn2/4 activity in response to several stresses.

Abstract.....	2
1 Introduction	4
2 Automated Flow Cytometry	8
2.1 HyperCyt System	9
2.2 Separation to individual wells	10
2.3 Possible limitations and drawbacks of the system.....	13
2.4 Automation.....	13
3 Robotic growth protocol	15
3.1 The need for automated growth protocol.....	15
3.2 General idea and proof of concept.....	17
3.3 Protocol phases.....	19
3.4 In practice.....	21
4 Data Analysis.....	23
4.1 Data correction	23
4.2 Single population analysis.....	27
5 Application – General stress response network	29
5.1 Mapping the activators and the repressors of Msn2/4.....	29
5.2 Screen of a DKO library	31
5.3 Genetic interactions analysis	32
6 Discussion.....	36
7 Acknowledgments	38
8 Bibliography	39

1 Introduction

Gene expression regulation mechanisms are one of the major and intriguing aspects of molecular biology. However, these mechanisms are far from being fully understood and there is a myriad of open questions in this field of research. All the phases of gene expression are regulated, the packaging of the chromatin, the recruitment of transcription factors, the different steps of the transcription process itself and even the degradation of the mRNA and the proteins. Thus dozens of proteins are involved in regulating the expression of a single gene and they can be divided to many sub-types; The general transcription factors that are required for the proper RNA polymerase II activity, the sequence specific binding factors (TF) that control the rate of the transcription of a specific gene or a group of genes, the chromatin remodeling proteins that control the dynamics of the chromatin structure thus modulating the binding of the TFs and the RNA polymerase II to the gene and many more (Figure 1).

Changes in the activity and the quantity of each one of those regulatory proteins can affect the transcription profile of a specific target gene. The type and the size of the effect depend on the specific role of the protein in the regulatory pathway of the target gene. Therefore altering the expression of one gene and measuring the influence of the alteration on the transcription profile of another gene is a potential method for understanding the interactions between two proteins and revealing new functions and mechanisms.

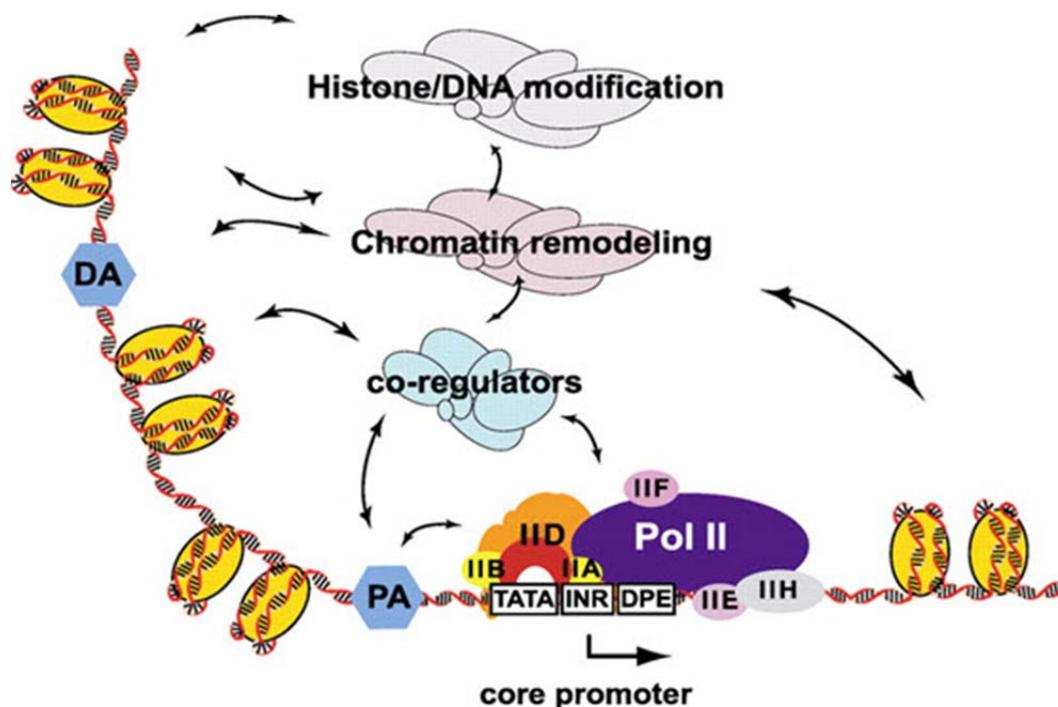


Figure 1: Illustration of different components involved in the transcriptional process. Adapted from: Lodish H, Berk A, Zipursky SL, et al. *Molecular Cell Biology*. 4th edition. New York: W. H. Freeman; 2000.

The basic mechanisms of gene expression regulation are highly conserved in eukaryotes, allowing their study in model organisms. Our model organism is the budding yeast (*Saccharomyces Cerevisiae*), it is easy to grow and manipulate, well studied and there is a rich existing infrastructure of molecular and genetic tools. Performing genetic manipulations on the yeast genome is relatively easy. One of the common manipulations is gene knockout (KO), a removal of a particular gene from the genome. An important resource that we have in the lab is the knockout (KO) library, a collection of gene deletion mutants¹. In our experiments we use KO strains that were designed in our lab along with strains from the KO library to study the function of the deleted proteins. Another important technique is fluorescent reporters. A sequence encoding for a fluorescent protein can be incorporated into the genome, fused to the protein of interest. The fluorescence level will then allow estimating the amount of mature protein in living cells². Mating a strain containing a fluorescent reporter to one of the KO strains, will allow us to examine the effect of a lack in one gene on the expression of another, when compared to the expression of the gene in a wild-type (WT) strain.

The levels of the reporter protein can be measured in a single cell level. In this work we use *flow cytometry*; a laser-based technology that allows a multi-parametric analysis of up to thousands of particles per second. The average fluorescence level in the population can be used as a measure to the reporter protein abundance. The single cell resolution can be taken in advantage to measure the cell-to-cell variability in gene expression, revealing another dimension of the regulation mechanisms^{3,4}.

Examining the effect of a single perturbation on the expression of a gene can help us to identify the players involved in its regulation. However, this is not enough to reconstruct the transcription regulation network. By examining the effect of double perturbations we can identify genetic interactions between genes. We would say that a genetic interaction exists if the phenotype of the double KO (DKO) is surprising given the phenotypes of the single KO (SKO)^{5,6}. By measuring multiple genetic interactions we can build interaction maps. These maps can be utilized to reconstruct the transcription regulation networks and to separate the different players into modules^{7,8,9}. Our vision is to measure and map hundred thousands of interactions. This requires adapting high-throughput measurement techniques and developing suitable analysis methods. In this project we describe the establishment steps of a robotic high throughput measurement system for dissection of transcription networks in yeast and the development of analysis methods that help us to interpret the data (Figure 2).

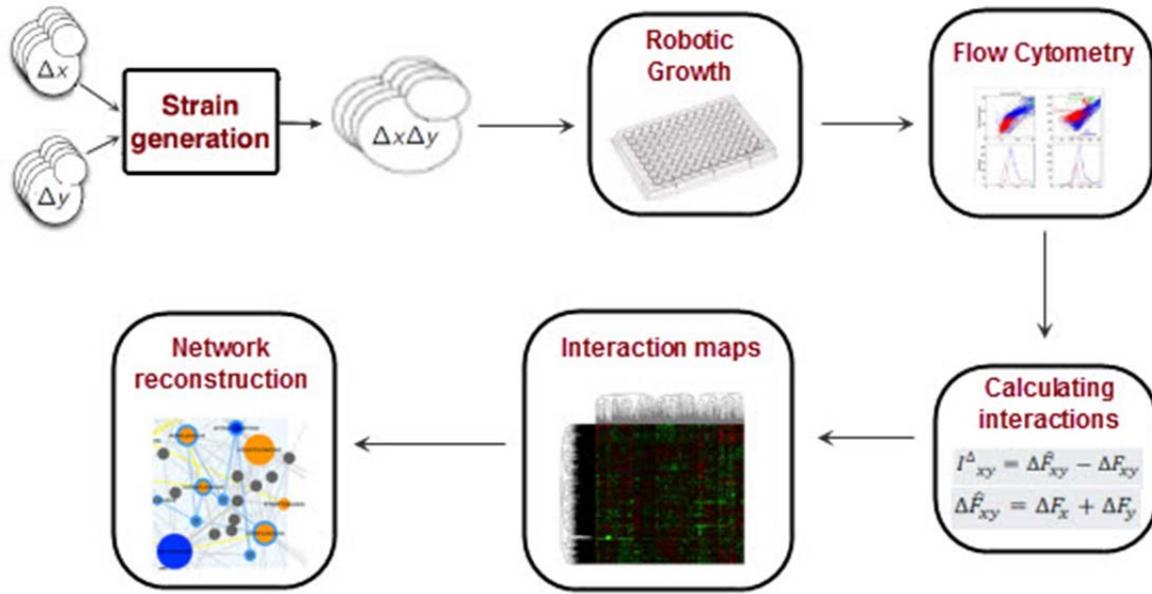


Figure 2: An illustration of our vision, a high throughput measurement system and an analysis infrastructure for reconstructing transcription networks.

To carry out high-throughput screens the measurement system should be able to work with microtiter plates, a feature which our flow cytometer (FACS) device does not have. Thus we are using an automated plate sampler that aspirates samples out of microtiter plates and transfers them sequentially to the FACS, resulting in one continuous data file for the entire plate. We developed an algorithm based on sequence alignment that partitions events from data stream of the whole plate to their correct wells. We built automation software that connected the system to an existing robotic liquid handling system, enabled the scheduling of FACS experiments and facilitated the running of time-lapse FACS measurements of an entire 96 wells microtiter plate with sampling frequency of 7 minutes per time point.

Yeast cells change their gene expression patterns according to their physiological state and to the changes in the environment¹⁰. Thus, the growth conditions of the cells prior to the measurement of gene expression can affect the results. To make sure that the measured difference in gene expression is due to the difference between strains and not a side effect of the growth process, we perform the experiment when all the strains are in the same growth phase. Together with Avital Klein, we designed and calibrated a novel automated high-throughput growth protocol. Given a 96/384 microtiter plate filled with yeast strains, each having a different growth rate, the protocol grows the plate in a way that at the beginning of the experiment all the strains on the plate are in exponential growth phase at a pre-determined density range.

The two main statistics that we extract from the flow cytometry data are the median fluorescence level and the cell-to-cell variability in fluorescence. We tested the integrity of those parameters

and showed that they are reproducible and comparable. The massive amounts of data that we deal with require a series of automatic corrections and sanity checks. We describe here the different procedures that we developed.

Once the system became functional we put it into use. In collaboration with Amit Sadeh, we focused on an important fundamental biological system, the response of a cell to environmental changes. The budding yeast model organism has a complex array of pathways to detect and respond to different environmental stress conditions. In addition it has a generic response to stress that is mediated by two transcription factors, Msn2 and Msn4, which regulate the transcription of hundreds of genes^{10,11}. Msn2/4 are part of a complex transcription regulation network and their activity is regulated by different components and pathways¹². Some of the regulation pathways have been studied and described^{13,14}, however most of the network is still uncharacterized. In this project we tried to reconstruct the part of the network upstream to Msn2/4. We created a library of ~3000 strains with double KO of different Msn2/4 related genes. The library was screened to measure the genetic interactions in the network as a response to different stress conditions. We computed the interactions that appeared in the data and used them to construct several genetic interaction maps. Many novel interactions were discovered and will be validated in further research. The maps now can be used to reconstruct regulation pathways within the general stress response network.

2 Automated Flow Cytometry

Most of the research questions in our lab, and especially those that deal with regulation of gene expression, require quantifying the expression levels of different proteins in the yeast cell. We are working with yeast strains in which the protein of interest is tagged with a fluorescent reporter, allowing us to use the fluorescence level of a specific cell as an indicator of the protein level. One of the methods for measuring the fluorescence levels of a cell is *flow cytometry*, a laser based technology that performs rapid single cell measurement in multiple fluorescent channels. We are equipped with a BD FACSCalibur flow-cytometer (FACS) with CyTek upgrade. Our FACS has 3 lasers at 488, 514, and 561nm wavelength that allow us to measure GFP, YFP and RFP/Cherry levels.

Briefly, flow cytometric measurement works as follows. Cells are placed into a test tube and drawn into the FACS. The sample stream is narrowed until the cells are forced to flow into the measurement system one at a time (Figure 3a). One or more laser beams of a single wavelength are directed on the cells, and the scattered and fluorescent lights are measured by an array of detectors measuring light intensity at different wavelengths. The signals are then analyzed and for each particle in the sample, several parameters are derived and stored. The output data file is a table. In this table the rows represent the measured particles and the columns represent the different parameters. The main parameters that we are working with are: Forward Scatter (FSC) – proportional to the diameter of the cell; Side Scatter (SSC) - proportional to the cells granularity; BluFL1 - a measure of the GFP fluorescence levels (515-575nm); RedFL2 - a measure of the Cherry/RFP fluorescence levels (590-640nm) (Figure 3b).

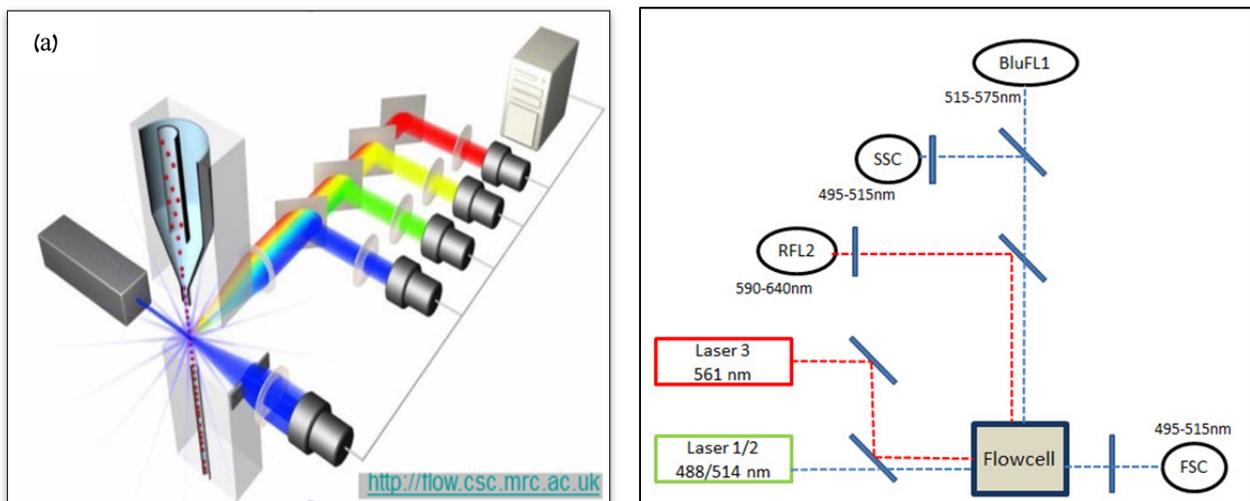


Figure 3: (a) Illustration of the flow-cytometer inner measurement system. (b) Scheme of the light path configuration of the flow cytometer device.

2.1 HyperCyt System

The FACS system itself does not meet two of the main requirements that we have of our measuring system. The first requirement is high throughput experiments. We measure thousands of different strains in each experiment, and thus we have to work in microtiter plates format throughout the experimental steps (sample storage, sample preparation and the measurement itself). However, our FACS does not have the ability to sample directly from microtiter plates. The second requirement is time lapse experiments. Most of the experiments that we perform measure the dynamic response of a cell population to stress, which requires an automated time lapse measurement of the samples.

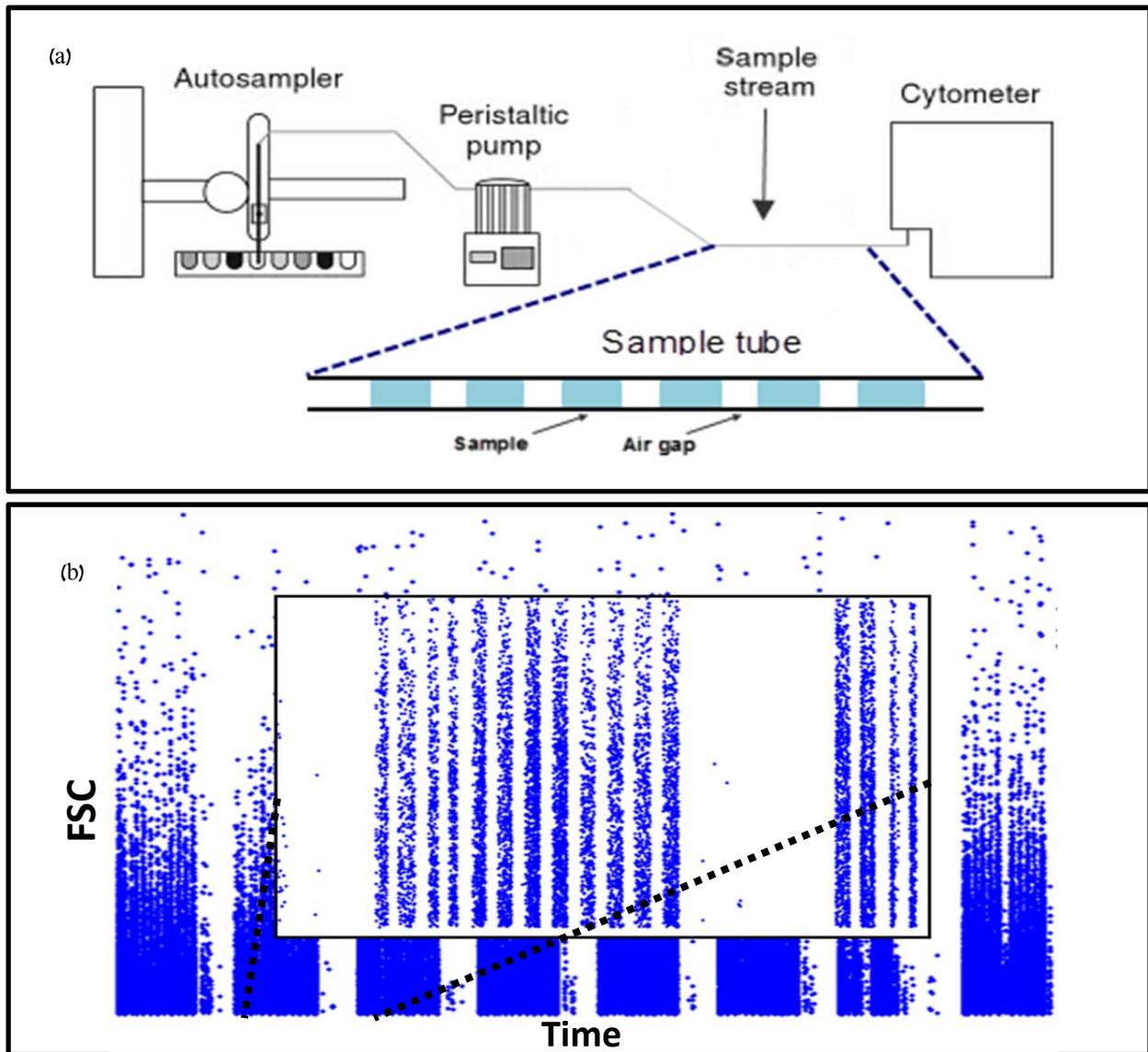


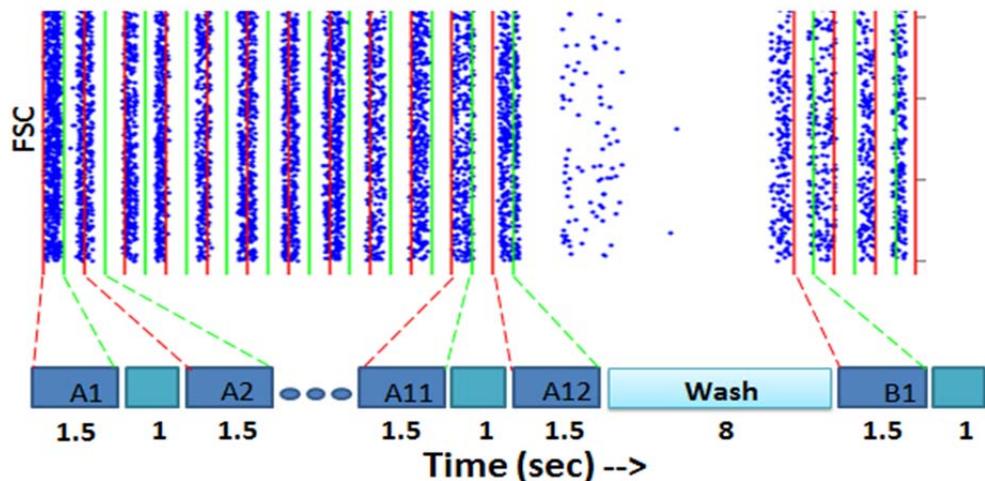
Figure 4: (a) A scheme of the HyperCyt auto sampler connected to the flow cytometer by a sample tube. (b) Example of FACS data output after a whole plate sample using the HyperCyt device. Each point represents one event from the sampled data. The short periodic gaps correspond to the air gaps between wells and the long ones correspond to the rinsing periods between plate rows. The window at the center of the plot is a zoom in to one of the rows in the plate.

To overcome these issues we are using the HyperCyt automated sampler (IntelliCyt). The HyperCyt is a small robot that aspirates cell samples out of a 96 or 384-well microtiter plates and transfers them sequentially to the FACS through a tube while the rate of the sample flow is controlled by a peristaltic pump (Figure 4a). Between every two wells the pump draws air into the tube, resulting in the formation of a series of air-gap-separated samples flowing through the tube. The HyperCyt samples the plate according to pre-determined parameters, among them: SipTime – The duration of the sampling from each well. UpTime – The pause between two wells and the length of the air gap. Rinse – The duration and the frequency of rinsing, usually there is a rinse interval after each row of wells. These parameters determine the sampling volume, the length of each experiment and the time intervals between the wells in the FACS output data. Throughout the experiment the FACS proceeds to continuously collect events into a single data file. When looking at the output data of a whole plate (Figure 4b) we can clearly see the periodic short gaps between events that correspond to the air gaps and the longer gaps that correspond to the rinse intervals between the rows of the plate.

2.2 Separation to individual wells

To analyze and comprehend the results of a whole plate experiment we should match each event in the data with the well it was aspirated from and partition the sequential event stream to individual wells. The first approach that we tried separated the event stream according to the known experimental timing as set by the run parameters (SipTime, UpTime, RinseTime). We created a temporal template according to these parameters and tried to find the best fit between the template and the data. The template only partially fitted the data and in most cases the wells boundaries were not defined precisely (Figure 5). The source of inaccuracies is temporal shifts and "warps" between the sampling from wells to the measurements in the FACS, due to physical properties of the air-gap separated flow in the HyperCyt tube.

Figure 5: At the lower part of the figure there is an illustration of a temporal template that was built according to the run parameters (1.5 seconds for SipTime, 1 for UpTime etc.). At the upper part of the figure there is an example of FACS data that was separated to wells according to this approach. A red line indicates the beginning of a well and the green line is the end of a well. We can see that the fit between the template and the data is not accurate.



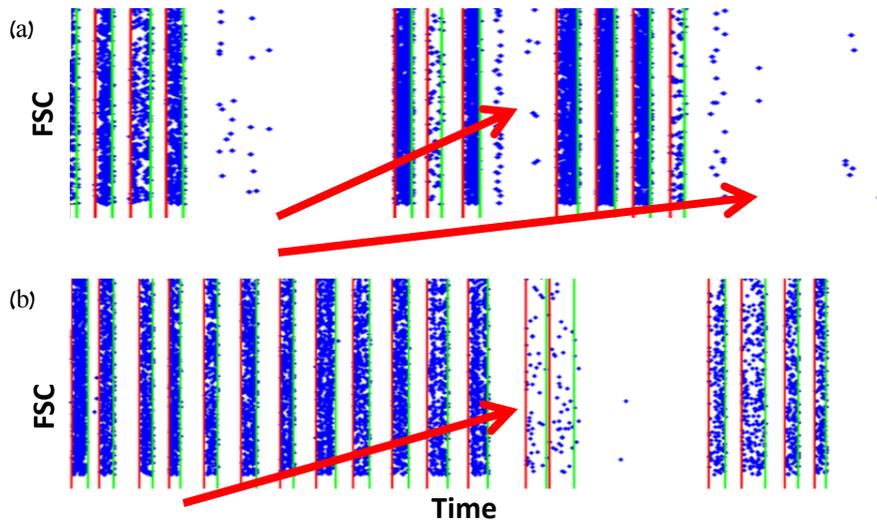


Figure 7: Second approach-separation of the data according to events density. The figures show the result of the separation to wells of a random FACS data. A red line indicates the beginning of a well and the green line is the end of a well. (a) In this example the algorithm fails to identify sparse wells due to insufficient events count in these areas. (b) In this example the algorithm identifies redundant wells at the end of one of the plate rows. The redundant wells are located at the beginning of the rinse region and contain significant amount of events, probably due to contaminated rinse media or a massive carryover from the previous wells.

The second approach was separating the data according to the density of events. We divided the event stream into small time windows and assumed that a dense window will be a part of a well and a sparse window will be an air gap or a rinse region. We traversed the list of events and for each event checked if it is located in a dense or a sparse window. A large enough group of consecutive events that were found in a dense window was defined as a well. The method was surprisingly accurate, the algorithm found all the wells that appeared in the data. The drawback of this approach was the disregarding of existing information about the specific experiment that we are dealing with: the number of wells in the plate and the different run parameters. On one hand the algorithm could not identify the existence of empty or sparse wells, which were identified as a rinse regions or air gaps and on the other hand it sometimes treated contaminated rinse regions as data wells (Figure 7). This problem had a significant influence on the data analysis.

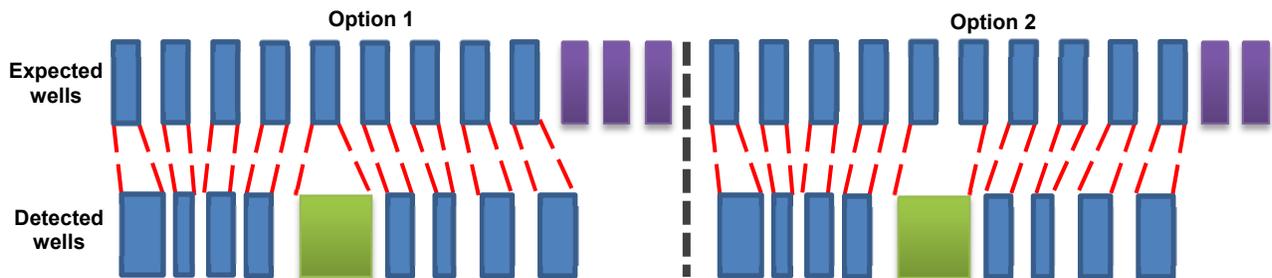


Figure 6: A simplified example of the score function considerations. In the lower part there is an illustration of the wells that were detected by the second approach. In the upper part there is a wells template that was built according to time parameters. In option 1 the green wide well is aligned with one narrow well, thus the penalty paid due to large time differences is significant, moreover at the end of the template row there are three purple wells that were not aligned (gaps). In option 2 there are only two unfitted wells at the end of the row and the green wide well is fitted to two narrow wells, the length of the two narrow wells together is close to the length of the green well, thus the penalty due to time differences is negligible. The score of option 2 is lower and the algorithm will prefer it in front of option 1.

Our solution was to combine the two approaches described above. Each one of the previous approaches outputs a sequence of well start times and end times. The first sequence is based on what we expected to see in the data according to the experiment parameters and the second is based on the event density. The idea was to find the best agreement between the detected and the expected wells. The expected sequence will be used to decide which one of the detected wells is a real well and whether there are more undetected wells. The detected sequence will determine the precise start and end times of the expected wells.

To find the best agreement between the sequences of times we posed the problem as an alignment, similar to sequence alignment. We designed an algorithm that uses the dynamic programming approach and is based on the Needleman-Wunsch algorithm¹⁵ for global alignment. We made a few changes to suit the algorithm to the task. Only matches between start times or between end times were considered, thus in the dynamic programming matrix half of the slots remained empty. We built a designated score function for the problem. The matrix is filled using the values that the score function outputs for each pair of indices. The value of cell (i,j) is the score of best possible alignment that begins at indices (1,1) and ends at indices (i,j). The output of the score function given a pair of indices is determined according to two parameters. The first is based on the time difference between the compared start/end times, larger difference will incur a larger score. The second is based on the number of data events that will be discarded in case that the compared values will be matched at the final result (Figure 6). After filling the matrix we trace the best path between the upper left cell to the lower right cell. Each type of move in the chosen path symbolizes one of the possible cases in the construction of the final alignment. The different cases include splitting one of the detected wells into two, merging two detected wells, ignoring a detected well and adding a well to the final result (Figure 8). After the construction of the final result few more heuristics deal with the different cases, for example finding the best place for splitting a wide well into two. The algorithm worked very well on real data and overcame the problems that the first two approaches had. The recent program version outputs a data structure

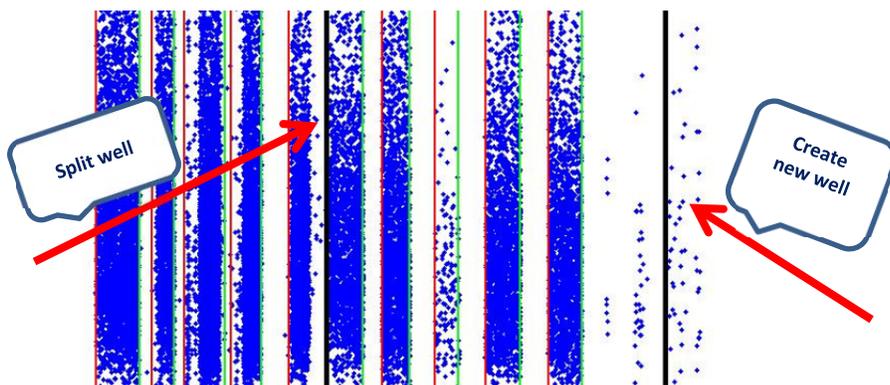


Figure 8: The black lines in the graph represent the modifications of the result that were made by the alignment algorithm. The left line exemplifies a case in which one wide well was split by the algorithm into two wells. The right line demonstrates an addition of wells in the end of the plate row, the exact place of the added wells is less important because the wells are empty.

that contains all the metadata of the experiment and for each original well plate the chunk of the data that was related to it together with a few basic values that describe the population of the cells in the well.

2.3 Possible limitations and drawbacks of the system

Along with all the benefits described above the non-conventional FACS+HyperCyt setup introduces some problems into the data. One of the main concerns in such a system is sample carryover between different wells in the plate. The carryover can be attributed to two sources, due to the movement of the sample probe between the wells and due to adhesion of cells to the walls of the sample tube. The percentage of carryover can be reduced by adding wash steps during the experiment, preserving homogeneous sample concentration between the wells in the plate and by not inserting the sample probe too deep into the wells during the sampling. More steps for dealing with carryover can be done at the data analysis level (Data analysis chapter). Another type of problems can arise due to the physical properties of the flow. An example for such a problem is fluctuations that we see in the measurement of the FSC and BluFL1 parameters. The width of the fluctuations is changing together with the rate of peristaltic pump. We believe that the peristaltic pump creates vibrations in the sample stream and those vibrations interrupt the FACS measurement system. We found that the fluctuations can be significantly reduced if the consumable parts of the HyperCyt are replaced on a regular manner. However we developed testing and correction procedures to deal with this problem (Data analysis chapter).

2.4 Automation

The last step that was left to make the system operational and start running high-throughput experiments was the automation of the system. We had a few requirements of the automation: to connect the flow cytometry system to the Tecan liquid handling robotic station that we have in the lab, which will link between the sample preparation and the measurement steps and turn the FACS into a part of the larger automated robotic system that we have in the lab. Another requirement was to have an interface for scheduling and running FACS experiments according to some predetermined parameters which will enable to carry out experiments without human intervention. The last requirement was facilitating the run of time-course experiments (i.e. repeatedly sampling the same plate with predefined intervals between the samples).

We built automation software that coordinates between the different components of the system (Figure 9b). As input the software receives the parameters of the desirable experiment design: the setup of the HyperCyt experiment and the time-course parameters. Controlling the HyperCyt can be done through an API that allows a program to control its operation. However the FACS

interface was not designated to be a part of an automated system and the acquisition software is accessible only through mouse/keyboard interface. To overcome these problems we physically connected the FACS to a relay board that can turn the FACS on and off, the relay board is connected to the network and can be controlled through the automation software. To automate the communication with the acquisition software we used the Sikuli Script¹⁶ which is an automation and testing software that uses image recognition to identify and control GUI components. We created a Sikuli script for each group of orders that we are executing during the acquisition, the script runs a macro that imitates mouse and keyboard actions. The automation software is using the Sikuli scripts to control the acquisition and synchronizedly controlling the HyperCyt through its API. The automation software also handles the execution of time-course experiments by repeating single experiments over and over again. This non-standard setup makes it difficult to monitor the run of a single experiment and to identify and handle failures on real time, making the system less robust.

In the experiments that we are already running and intend to run in the future, the preparation of the sample is crucial for the success of the experiment and usually needs more resources and time than the measurement itself (Robotic growth protocol chapter) and the timing of the measurement is defined precisely by the former steps of the protocol. Thus it is very important to physically connect between the flow cytometry system and the other robotic components in the lab and allow the measurement to be scheduled and managed together with the other steps of the protocol. One of the components of our robotic system is the KiNEDx 4-Axis robot with an electric plate gripper and 2000mm horizontal rail that can transfer plates between the HyperCyt to the liquid handling station. We defined the HyperCyt as one of the devices of the liquid handling system and created a server-client connection between the HyperCyt and the liquid handling robot computers (Figure 9b). This setup allows the management system of the liquid handling robot to directly operate our original flow cytometry automation software and schedule FACS experiments.

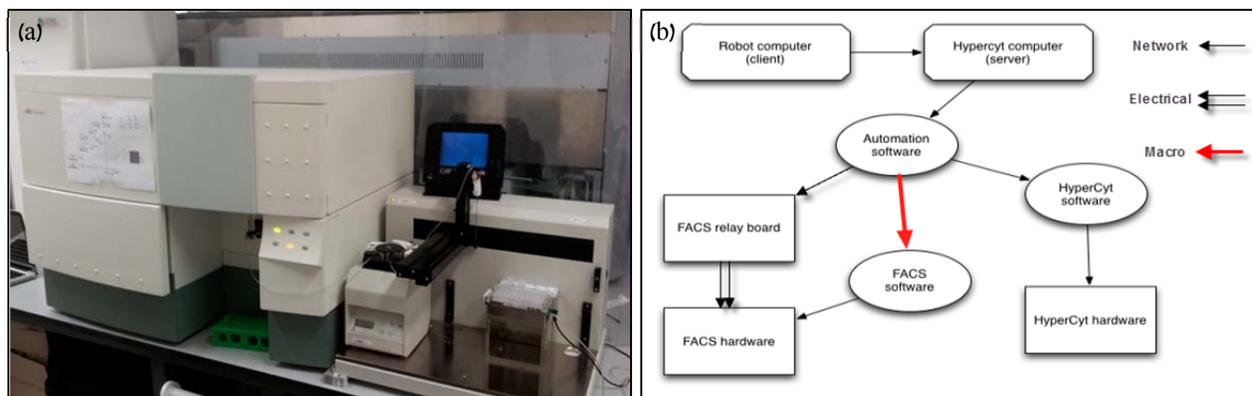


Figure 9: (a) A photo of the FACS (on the left) and the HyperCyt (on the right) during a running experiment. (b) Scheme of the automation infrastructure, the automation software in the middle of the diagram is receiving commands from the robotic system and synchronizing between the FACS and the HyperCyt devices to run complete time-course experiments.

3 Robotic growth protocol

In the majority of the experiments measuring gene expression, the physiological state of the yeast population can affect the results and their interpretation. Usually in yeast laboratories the strains are either frozen (-80°C) or refrigerated on agar (4°C) and transferred to fresh nutrient rich media prior to the experiment. In the first few hours the cells are being adjusting to the new conditions and hardly reproduce, this period is defined as the *lag-phase*. The next phase is called the *log-phase* and it is characterized by fast exponential growth. During this period the yeast utilize the glucose in the media by fermentation. We refer to the maximal duplication rate during the log-phase as the *growth rate* and the *doubling time* can be calculated accordingly. The doubling time can have a wide range of values depending on the genotype of the strain. When the glucose in the medium is depleted, the yeast shift to *respiratory growth*, characterized by a much longer doubling time. After a few days the population reaches the *stationary phase* and the growth stops.

3.1 The need for automated growth protocol

During the different growth phases the gene expression patterns change dramatically. Thus it is important to measure gene expression when all the measured strains are in the same growth phase. So we will know that the observed differences in gene expression are consequences of the differences between the strains and not a side effect of the growth process that preceded the experiment. In our lab we are running high throughput gene expression experiments that examine and exploit the stress response of yeast. This combination turns the growth protocol to be even

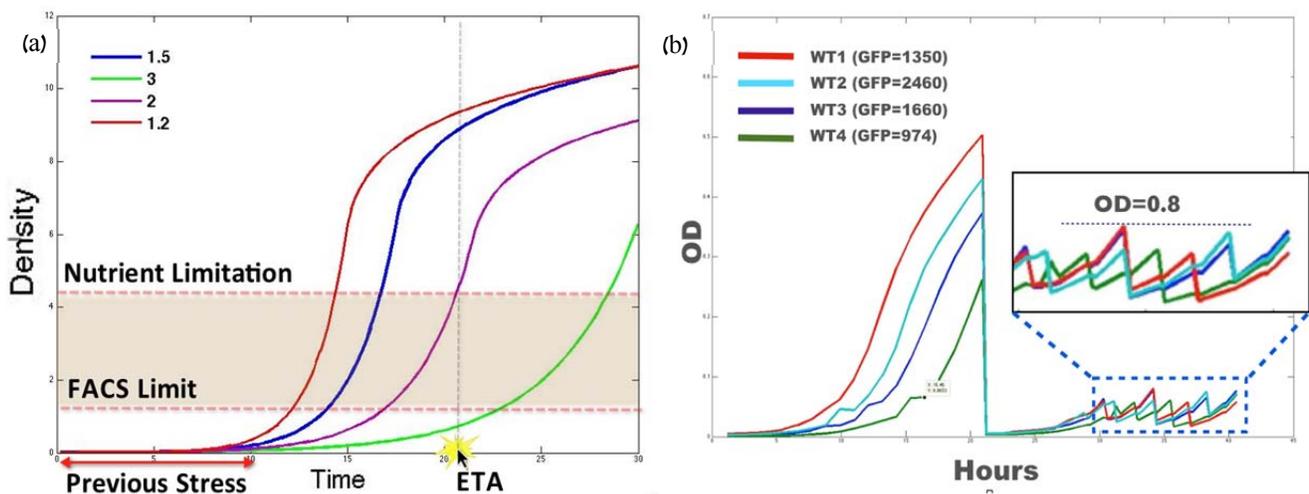


Figure 10: (a) Illustration of the constraints of the growth protocol. Previous stress – The minimal time needed to allow the population to get accustomed to optimal conditions. Nutrient limitation – The OD threshold from which the cells turn on the general stress response. FACS limit – The lower bound of sample density that is acceptable by the flow cytometry system. ETA – the scheduled start time of the experiment. The colored curves simulate the growth of KO strains with different doubling times. (b) Example of a straight forward “measure and dilute” growth protocol for 4 WT strains. The stress level of the population was measured at the end of the protocol. The timing and the amount of dilutions had a direct effect on the condition of the population.

more complicated, as it puts some strict constraints on it. When trying to characterize the response of a cell to some environmental stress we will usually wish that prior to exposure to the stress the cell will be in un-stressed optimal conditions. For this to hold we need to wait enough replication cycles between the previous exposure of the population to stress to the experiment itself, letting the cells to get accustomed to the un-stressed conditions and to assure that the cells that were exposed to stress will become a negligible fraction of the population. In practice, we wait for at least 4-6 doublings, which on average last for ~10 hours. Moreover, we observed that the yeast general stress response is activated already at the middle of the log-phase, probably due to sensing of the increasing density in the media. This observation determined the upper bound of the sample density for our gene expression experiments, setting it to 0.9 optical density units (OD) which is an early log-phase. Measuring a sample whose density is too low is not possible in our high throughput flow cytometry setup, because makes it difficult to distinguish between empty and dilute wells (Automated flow cytometry chapter). This limitation sets the lower bound of the sample density for FACS experiments to 0.3 OD. The last constraint that should be mentioned is the estimated experiment time (ETA). To successfully run high throughput experiments we have to follow a strict schedule, so the end time of the growth protocol has to be coupled with the ETA of the FACS experiment. These constraints are illustrated in Figure 10a.

Given the above constraints, a precise growth protocol is crucial for the success of experiments. Bringing a few yeast strains to the desired OD range at the ETA can be a relatively easy task. However, when we work with KO and double KO strains, we should consider the differences in

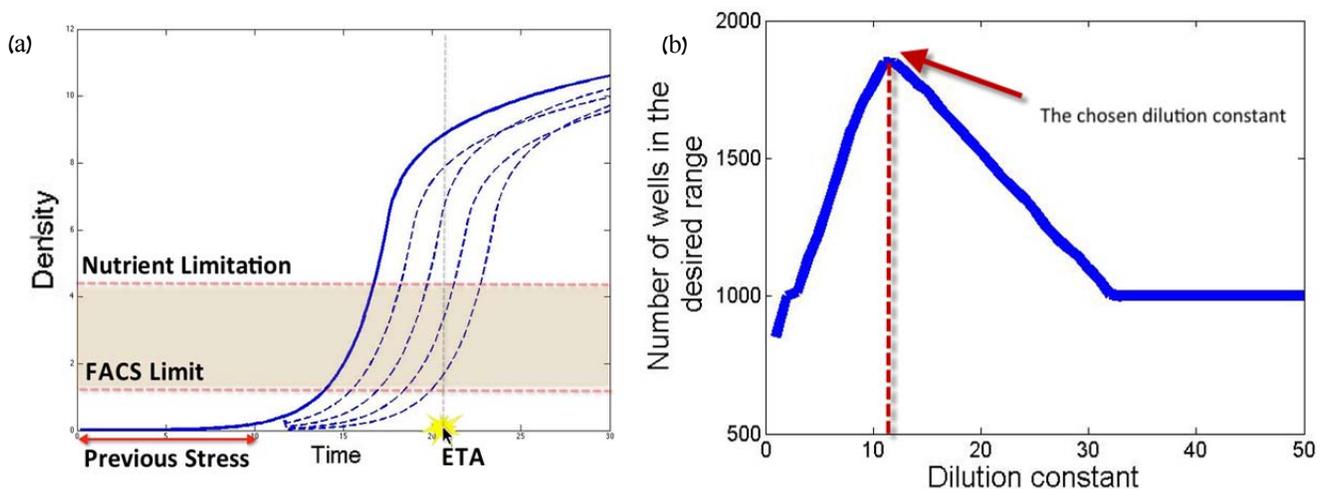


Figure 11: (a) The general idea of our growth protocol. Each strain is split into four, according to a 2-fold dilution series. If the dilution constant is chosen correctly, at least one of the diluted wells will reach the ETA at the desired OD range. (b) A simulation for choosing the dilution constant for 1000 strains with different initial parameters. For each dilution constant we calculate the number of strains that will reach the desired OD range, first assuming that all the strains have the highest possible doubling time and then assuming that they have the lowest one.

the doubling times (ranging between 1.5 to 3 hours) and in the lag-phase length. Thus, synchronizing between the final OD of the strains becomes much harder (Figure 10a). Making the protocol high throughput (work on 96/384 microtiter well plates of with different strains) becomes a real challenge. We will briefly describe one idea that we used in initial experiments. The density of the wells in a plate was measured at a few fixed time points. For each well we considered whether it should be diluted in order not to pass the upper OD bound until the next measurement and to reach the ETA in the desired OD range. The dilution constant was calculated according to a doubling time that was estimated based on density reads that preceded the dilution. Although the strains did reach the desired OD range at the end of the protocol, it had a few drawbacks. We observed big differences in the physiological condition of the strains at the end of the growth. We believe that the differences were a result of uneven number and extent of dilutions that the strain went through, and the maximal OD that it reached during the process (Figure 10b). Another drawback was the large number of dilution operations needed during the protocol. Thus we wanted to consider a protocol which will reduce the number of dilutions and will treat the different wells in a uniform manner.

3.2 General idea and proof of concept

To avoid these problems, we developed an alternative approach. In short, we are choosing a dilution factor given some initial and measured parameters. Each plate is duplicated into four plates in a dilution series based on the chosen factor (Figure 11a). We let the plates grow without intervention for a pre-defined amount of time. At the end of the growth we measure OD in the four plates and combine them into one, choosing for each well the source whose OD is closest to the optimal OD.

The challenging part in this approach is choosing the dilution factor. At the beginning of the protocol we do not have any information about the doubling times of the specific strains that we are dealing with. However, we can assume that the times are within some range. The lower bound of this range will be close to the doubling time of the wild type (WT) strain, since only a small percent of KO mutations increase the fitness of the strain relative to the WT¹⁷. The upper bound of this range is the doubling time of the sickest strain that we are willing to work with. So we would like to find a dilution factor that will suit all strains with doubling time within the given range. We designed a function which receives as parameters the initial OD of the strains and the expected range of their doubling times. For each possible dilution factor (between 1 to 50), the function calculates the number of strains that will have at least one source in the desired OD range at the ETA, first assuming that all the strains have the highest possible doubling time and second assuming that they have the lowest one. Formally: given OD_w (Current OD of well w),

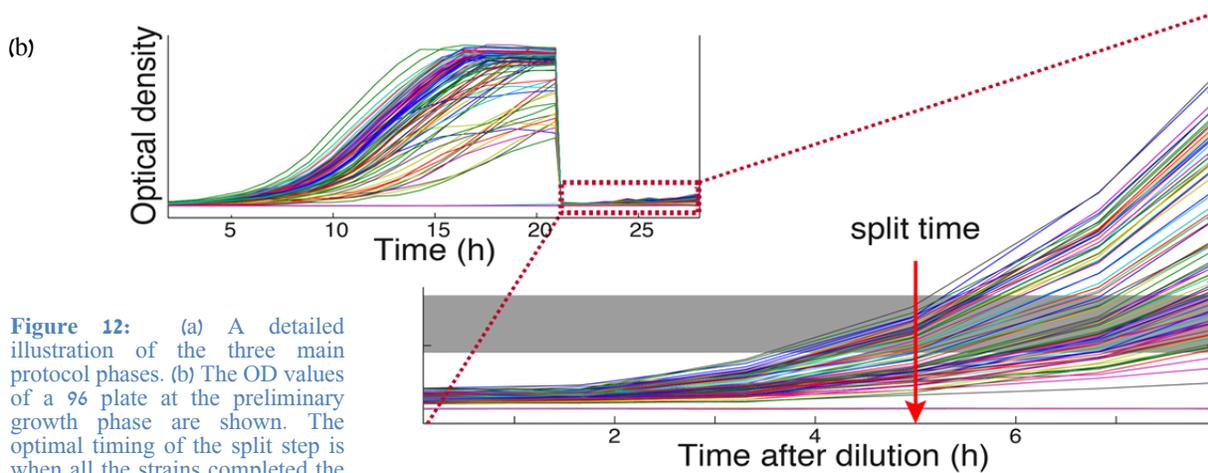
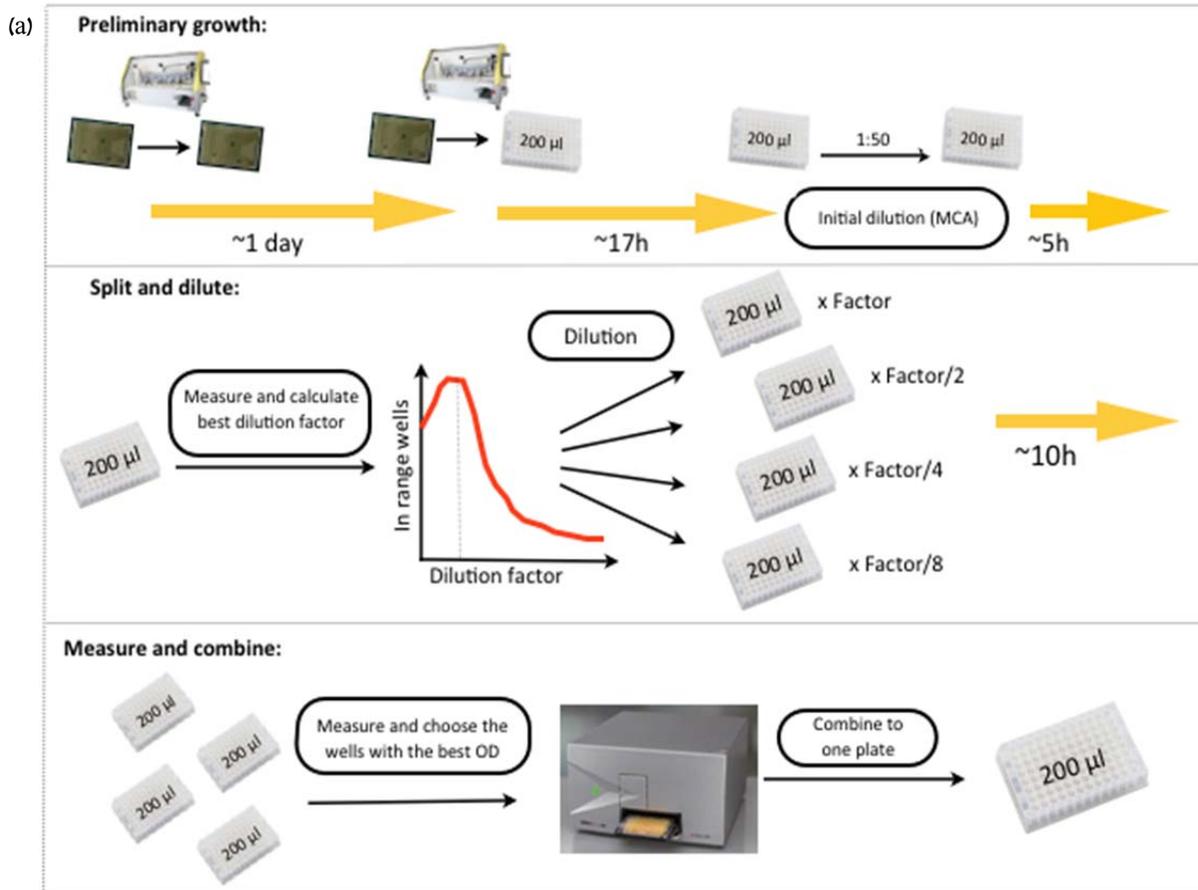
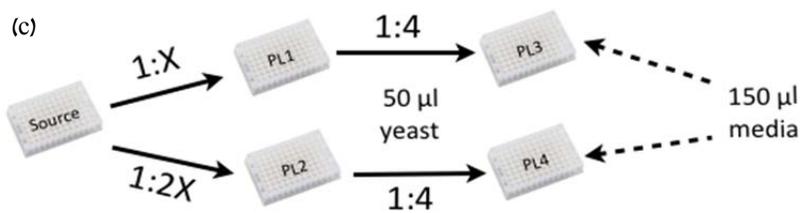


Figure 12: (a) A detailed illustration of the three main protocol phases. (b) The OD values of a 96 plate at the preliminary growth phase are shown. The optimal timing of the split step is when all the strains completed the lag-phase and the differences in OD between the strains are minimal. (c) The dilution steps at the split and dilute phase where X is the chosen dilution factor. The dilution scheme was planned to minimize the amount of tips replacement operations and to work with the largest volumes possible to get precise dilutions.



OD_{Max} , OD_{Min} (The desired range of final OD values), DT_{UB} , DT_{LB} (The upper and lower bounds of the doubling times) and GT (Time to ETA) the function is defined as:

$$F: [1 , 50] \Rightarrow [0 , 2 * NumWells]$$

$$F(x) = \sum_w^{NumWells} \sum_{DT \in \{DT_{UB}, DT_{LB}\}} \max_{i \in \{1,2,4,8\}} 1\{OD_{Min} \leq \frac{OD_w}{i * X} * 2^{\frac{GT}{DT}} \leq OD_{Max} \}$$

The chosen dilution factor is the one that maximizes the function (Figure 11b). We wrote a simple simulation to check whether the proposed design is feasible given the constraints of the robotic system and the properties of the yeast libraries that we are working with. Our main conclusion was that if we will be able to reach the split step when the differences in the OD between the strains are not large enough (maximal ratio of four), theoretically we can get close to 100% success.

3.3 Protocol phases

The first phase of the protocol is the preliminary growth (Figure 12a). This phase has two purposes. The first is to bring the yeast to well defined growth conditions before the final key steps of the protocol and escape the lag-phase period whose length can significantly vary between different strains. The second is to synchronize between the OD values in the plate prior to the split step and avoid dependence on the precise measurement of small OD values. The yeast libraries in our lab are usually held on 96/384 colonies agar plates in the refrigerator. We use a robotic colony copier (Singer Instruments RoToR) to replicate from agar plates to agar/liquid plates. The first step of this phase is replication of the agar plate into a new agar plate dedicated to the experiment. After approximately one day the fresh agar plate is replicated into a liquid plate (filled with minimal growth medium (SD)). This replication is carried out through an intermediate liquid plate, using calibrated parameters, to minimize the initial OD values in the plate. The liquid plate is then inserted into the Tecan Freedom Evo 2000 liquid handling station. We initialize the new growth protocol instance in our scheduling software. Then we input the necessary information regarding the specific plate, the timing of the different protocol phases and the continuing FACS experiment. From this point on, the protocol is fully automated.

The liquid plate is grown inside a robotic incubator at 25°C, for 16-24 hours. Then the plate is diluted uniformly 1:50 (Figure 12b). After this step we expect the OD values in the plate to be relatively homogeneous since the 16-24 hours growth synchronized the OD values around the characteristic stationary OD value for SD medium (usually 8-10 OD).

The next phase of the protocol is the “split and dilute” phase (Figure 12a). As explained above, the key to the success of the growth protocol is the uniformity of the OD values at the split step. Thus the timing of the split step should be chosen carefully. On one hand we have to wait a few hours before the split because the yeast are in lag-phase after the uniform dilution. On the other hand if we will wait too long the differences between the OD values will increase due to varied doubling times. We calibrated the timing of the split to be 5-7 hours after uniform dilution, depending on the library that we are working with (Figure 12b). At the split step the plate OD values are measured and the optimal dilution constant is calculated by optimizing the function that was described above. The plate is split to 4 according to X: 2X: 4X: 8X dilution series, where X is the chosen dilution constant. The dilution is executed by a series of dilution steps, as described at Figure 12c. The dilution scheme was designed to minimize the amount of tips replacement and wash operations and to work with as large volumes as possible to get precise dilutions. After the dilution the plates are inserted into the incubator for a final growth period, which typically lengths 10 hours, but can be defined differently at the initialization step.

The next phase of the protocol is the “measure and combine” phase (Figure 12a). The density of the four plates is measured in a serial manner (Figure 13a). For each well we choose the source that best fits the desired OD range. Wells, for which no appropriate source was found, remain empty. A robotic script that executes a series of pipetting steps is being created on the fly. This script then performs the combine procedure. Finally, the OD values of the combined plate are measured.

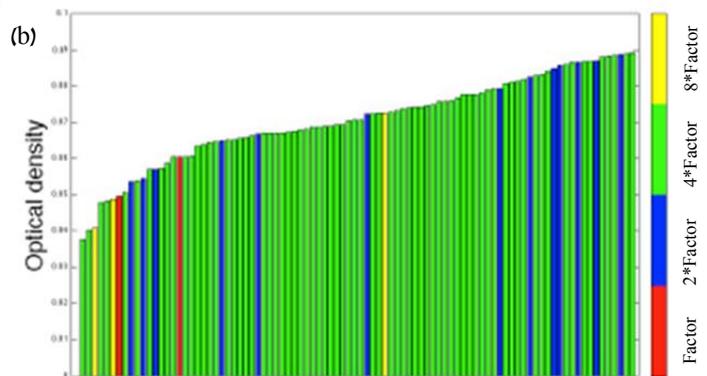
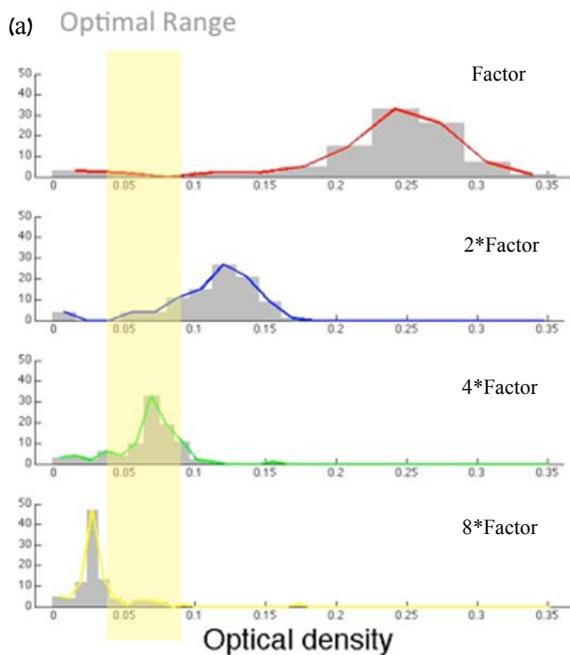


Figure 13: Illustration of the plate OD values at the “measure and combine” phase. The four plates are measured and combined into one, choosing for each well the source whose OD is closest to the optimal OD (a) Distributions of the OD values in each plate at a successful run of the protocol. The optimal OD range is highlighted in yellow. There is a 2-fold difference between the median values of the distributions, as expected. (b) A bar graph that shows the measured OD values in the combined plate, the color of the bar indicates the source plate of the specific well. All the wells are in the desired OD range at the end of the protocol.

Unless there were pipetting or measurement errors, the absolute majority of the wells in the final plate are in the desired range (Figure 13b). The final phase of the protocol depends on the specific design of the experiment that we wish to perform. Usually we expose the yeast to some environmental stress and send it to the automated flow cytometry system.

3.4 In practice

We successfully utilized the protocol over a few different yeast libraries, non-mutated, KO and double KO. Our throughput during the experiments was four 96-well plates per day, while each plate spent around 48 hours in the system. However the bottleneck in those experiments was the availability of the automated flow cytometry system. Ignoring this limitation we can achieve a throughput of 6-8 plates per day. In this case the bottleneck will be the working time of the liquid handling station and the availability of positions for plates and consumables.

The percentage of success of the protocol was high, when on average 95% of the strains reached the ETA in the desired OD range. The strains that did not pass the protocol were usually extremely sick strains that had a very long doubling time or lag-phase. In some cases strains failed to pass the protocol due to technical problems like bias in OD measurement and imprecise dilutions. We identified these problems and fixed the implementation of the protocol accordingly.

It is left to show that the protocol that we designed and implemented indeed meets our preliminary goal, to grow an entire plate to optimal un-stressed conditions. As part of our experiments we used the protocol to grow ~3000 double KO strains (32 plates) in two biological repeats. All the strains contained an endogenous GFP reporter fused to the Hsp12 protein, a marker for the general stress

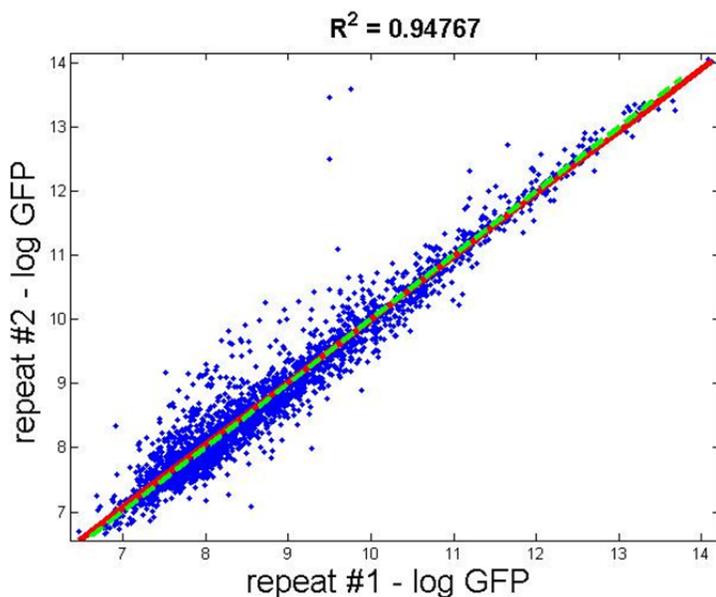


Figure 14: ~3000 DKO yeast strains were grown using our growth protocol in two biological repeats. The strains contain an endogenous GFP reporter fused to the Hsp12 protein, a marker for the general stress response. The strains fluorescence level was measured at the end of the growth protocol. The correlation between the median fluorescence values of the repeats is shown. The green line is the diagonal and the red line is the fit.

response¹³. The strains differed by the combination of KO mutations they had. Most of the KO mutations we selected were of genes related to the process of response to stress. The fluorescence level of the strains was measured at the end of the growth protocol. We observed good correlation between the median fluorescence levels of the biological repeats (Figure 14). The reproducibility of the results shows that the measured level of stress response activation is due to inherent properties of the strain.

4 Data Analysis

In this chapter we will describe the basic data analysis procedures that are relevant for all the data produced by our automated flow cytometry system. One can separate these steps into two tasks. The first is the correction and the gating procedures that are needed to reduce biases and variability in the data. The second is choosing and calculating the best descriptive statistics that we will work with in the advanced data analysis.

4.1 Data correction

One of the biases in the data is brought about by the flow cytometry technology. Ideally we would like to collect only data from viable single cells. However, often cellular debris and cell aggregates are counted as regular single cells and we can always expect some percent of dead cells in the sample. One of the possible solutions is gating the data – filtering the cells in the data

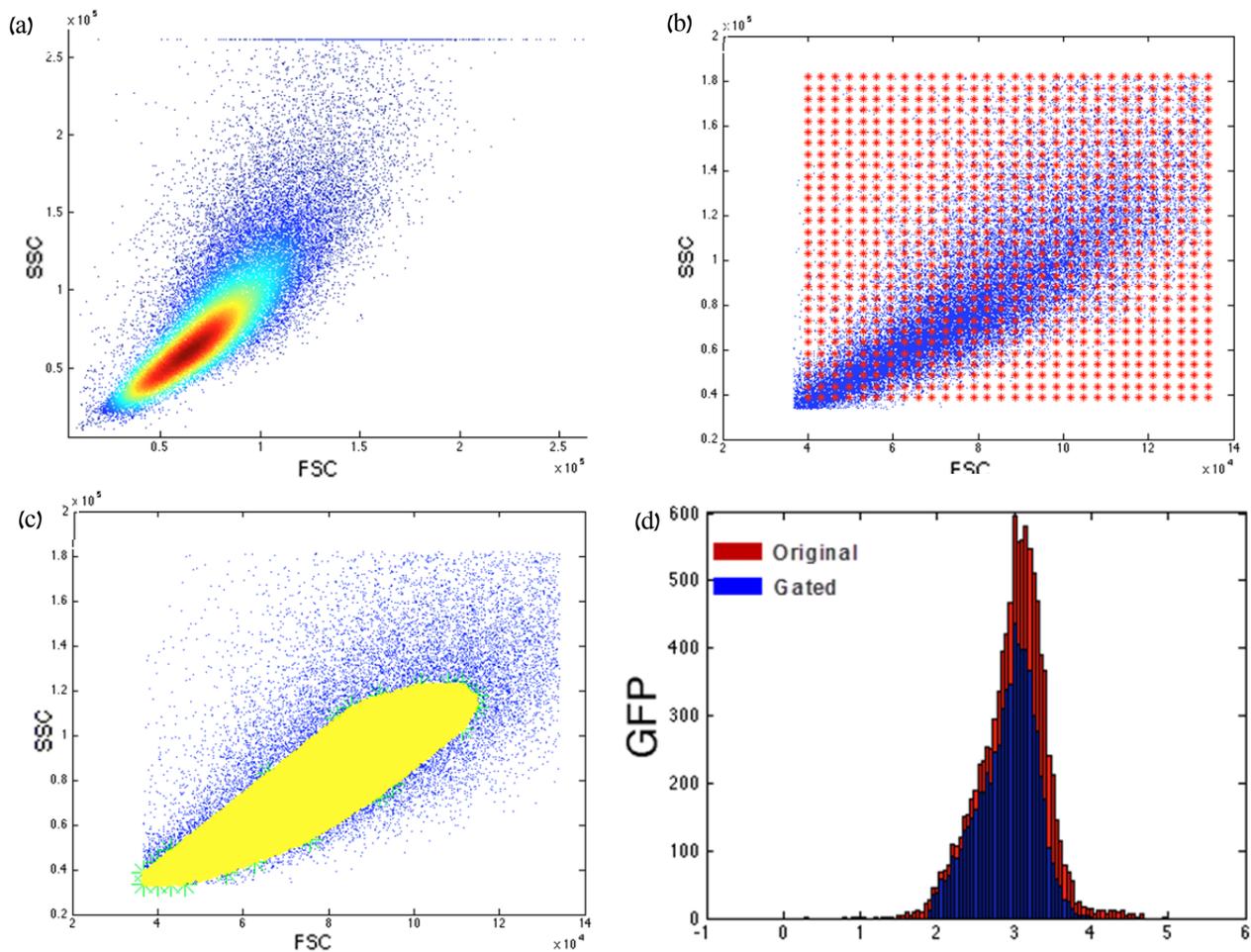


Figure 15: (a) Scatter plot of SSC against FSC in a random well. The graph is colored according to the read density at each point. (b) The area of the scatter plot is divided into $R \times R$ squares, when R is the resolution parameter given by the user. (c) The yellow area is the part of the plot that is bounded by the convex hull of the chosen coordinates. (d) Histogram of the GFP values. In red the original data and in blue the data after the gating procedure.

according to some FSC and SSC limits³. In a sample from a homogeneous population we expect the FSC and SSC parameters of the population to be normally distributed. Thus, when we look at the scatter plot of FSC vs. SSC, we expect to see a dense area that contains most of the population and around it a sparse scattering of events that are not representing the population (among them the cell debris, aggregates and dead cells) (Figure 15a). We wrote a gating procedure that automatically identifies the dense area in the plot and eliminates all the cells that remain outside of this area, given the percent of cells that we would like to retain after the filtering.

The gating procedure receives as an input two parameters. The resolution parameter R is a tradeoff between the accuracy of the procedure to its running time. The percent parameter P determines the percent of cells that should be filtered. First we eliminate the top and the bottom several percent (defined by the user) of FSC, SSC and GFP in order to get rid of the saturated values and the clear outliers. Then the area of the FSC vs. SSC scatter plot is divided into R*R regions (Figure 15b). We count the number of events in each region. We then greedily collect regions from the most occupied onwards until we collect the target percent of events (P). After the selection of the regions the combined area might not be convex and contain holes, thus the convex hull of the chosen coordinates is computed (Figure 15c). Only cells whose FSC and SSC are within the polygon are retained in the data, all the other cells are eliminated. The gating process usually entirely dismisses the cells that are situated in the tails of the GFP distribution. These are apparently the cell debris, cells aggregates and dead cells (Figure 15d).

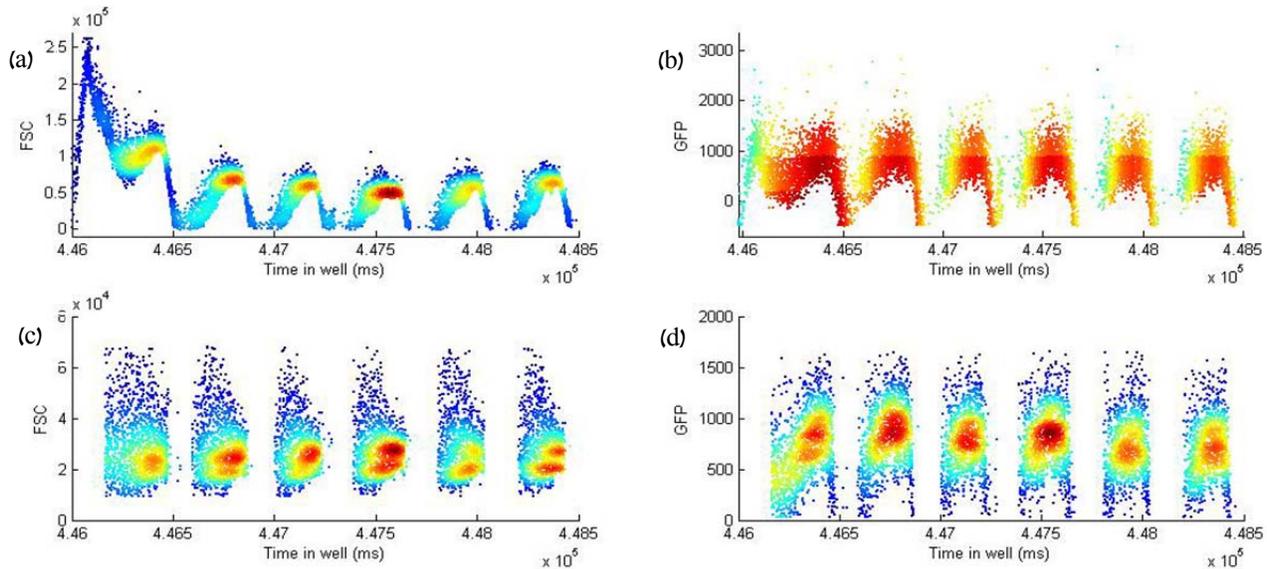


Figure 16: Demonstration of the fluctuations in GFP and FSC parameters. The parameters value change significantly in respect to the time in well. The graphs are colored according to the density of reads at each point. (a) FSC as a function of time in well (b) GFP as a function of time in well (c) FSC as a function of time in well in the normalized data (d) GFP as a function of time in well in the normalized data

Some biases in the data are introduced by the HyperCyt technology, one of them is fluctuations in FSC and BluFL1 parameters values that arise due to the physical properties of the flow (Figure 16a,b). We developed a procedure for correcting these fluctuations. First we normalize the FSC parameter by calculating its local regression (loess) and subtracting it. The FSC parameter is used only for processing the results of the current well and not for comparison between wells thus this type of normalization suited us. Since we are using the GFP parameter for comparing between different wells, we could not change its values, and had to eliminate all the areas with the biased GFP. We noticed that the time areas with the biased values are also the less crowded ones, thus we completely eliminated the sparse areas in the well. Finally we eliminated all the time areas with the extremely high or low GFP values (Figure 16c,d). We compared the mean GFP values that we got after the normalization to the mean GFP values that we got when we measured the same population through the FACS directly (without the effect of the peristaltic pump on the data). The results were almost identical.

This type of normalization is quite problematic because in the end we lose a significant part of the events in each well. Thus we made a lot of efforts to prevent the fluctuations from the beginning. We found out that a routine replacement of the consumables of the HyperCyt system minimizes the severity of the fluctuations. To make sure that the fluctuations are indeed minimized we designed a score function that receives the raw data of a single well and outputs a score that indicates the severity of the fluctuations in the well (Figure 17a,b). The score is based on the assumption that the values of a parameter (FSC, BluFL1) inside a single well should be homogenously distributed through the well. We calculated the fit of the data using the sliding window approach and defined the fluctuations score to be the standard deviation of the fit divided

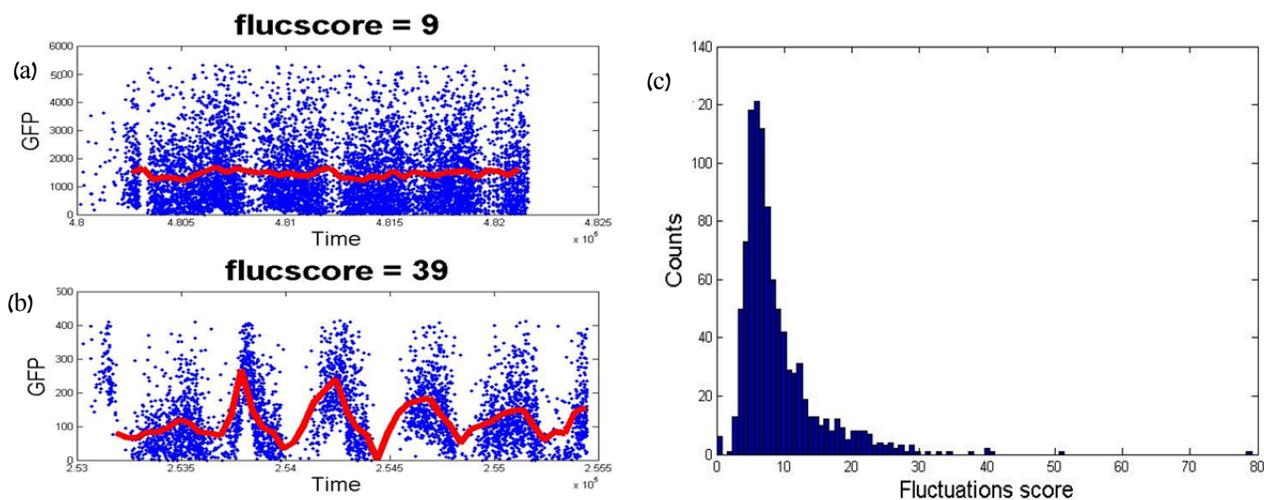


Figure 17: (a) An example of a well with a low fluctuation score. Each blue dot is a particle from the sample. The red line is the sliding window fit of the data. (b) An example of a well with high fluctuation score. (c) Histogram of the fluctuation score of all the wells in a time course experiment (96*12 wells). There is a small amount of wells with high score, these wells are checked individually.

by its mean. Checking the fluctuations score of the wells in an experiment helps us to make sure that the data is reliable and to identify cases in which we should use the normalization procedure (Figure 17c).

Another important bias that should be considered in our system is sample carryover between different wells in the plate. According to the system specifications and some published results the expected carryover in a no-rinse setup is between 1-3 percent¹⁸. Moreover, the carryover can be decreased by a correct experimental setup that includes long and frequent rinse steps together with appropriate sample preparation. We noticed that if the percent of carryover in the experiment is known in advance, a simple correction procedure can be used to discard the effect of the carryover. To calculate the un-biased average of a parameter in the i -th well in the plate (A_i), all we need to know is the average and the sample size of the previous well (A_{i-1}, N_{i-1}), the biased average of the current well (\hat{A}_i) and its sample size (N_i). Thus assuming that the first well in the plate is free of carryover and the percent of carryover is X . We can iteratively correct the values of the whole plate using the following formula:

$$A_i = \frac{N_i * \hat{A}_i - C_{i-1} * A_{i-1}}{N_i - C_{i-1}}$$

where $C_{i-1} = \frac{X * N_{i-1}}{100}$ is the size of the carryover from the previous well. We tried the procedure on a few data samples using different carryover percent and saw that even if we assume that the percent of carryover is 3% (the upper limit of the published percent for no-rinse setup, while our experiment setup includes rinse after each plate row), the expected change in the mean intensity

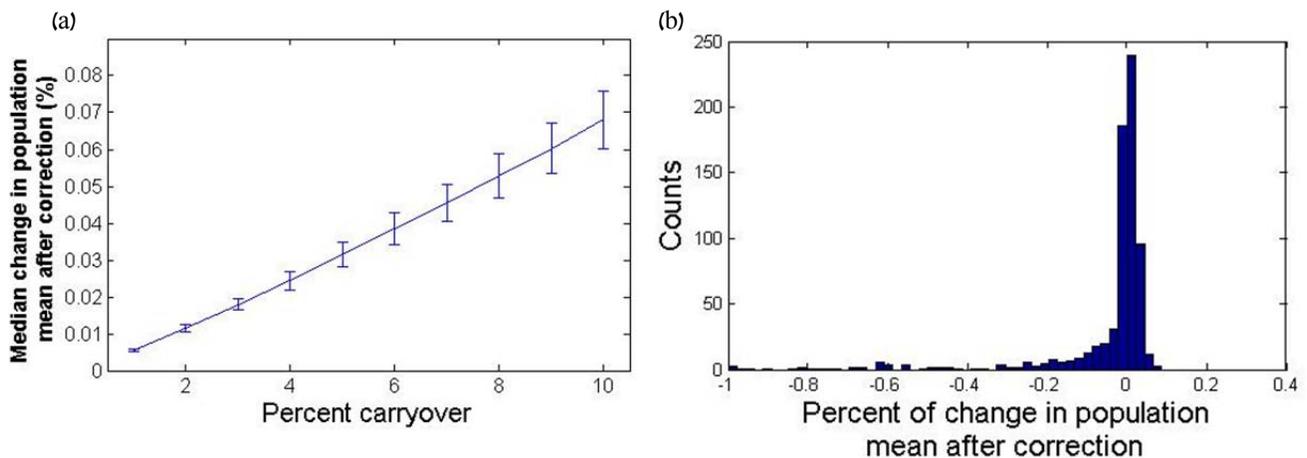


Figure 18: (a) The median change in population mean after the correction procedure was computed for 12 different plates. The mean of the 12 plates is shown versus the carryover percent that was used during the correction. The error bars show the standard deviation between the 12 plates. (b) The histogram of the change in the population mean after correction in ~1000 wells is shown, assuming that the percent of carryover is three. Most of the wells in which the change is significant relative to their mean have very low population size or very low mean fluorescence level.

caused by the carryover is approximately 2 percent of the original value (Figure 18a). If we use the correction procedure when the expected carry over percent is 3 and look at the distribution of the expected change over a whole time course experiment (~1000 wells), there are some wells in which the change is significant relative to their mean. These are the wells that have very low population size or very low mean fluorescence level (Figure 18b).

4.2 Single population analysis

After correcting for biases at the plate level, we analyze the data of each single well. If we look at the distribution of the fluorescence in a population before and after the gating process (Figure 15d), we can see that most of the extreme values were cut off and the remaining population is relatively homogeneous. Thus it makes sense to work with the BluFL1 median as a main indicator of the expression level in a population. We wanted to make sure that this statistic is reproducible and comparable, so we compared between biological repeats, between strains from different genetic backgrounds and between different fluorescent reporters (Figure 19a). When the measured population is homogeneous, the median is indeed a reliable and robust indicator for the measured parameter. However, in some cases we observe a bimodal or multimodal distribution in some of the parameters, which is usually caused by severe carryover or fluctuations. In rarer cases it can also indicate a problem with the sample or a real biological phenomenon. Therefore we would like to automatically identify such cases. To do so we built a procedure that fits the data to one and two Gaussian functions and compares between the likelihood of the fits. We treat the samples whose bimodal fit has a higher likelihood individually and try to find the source of the bimodality.

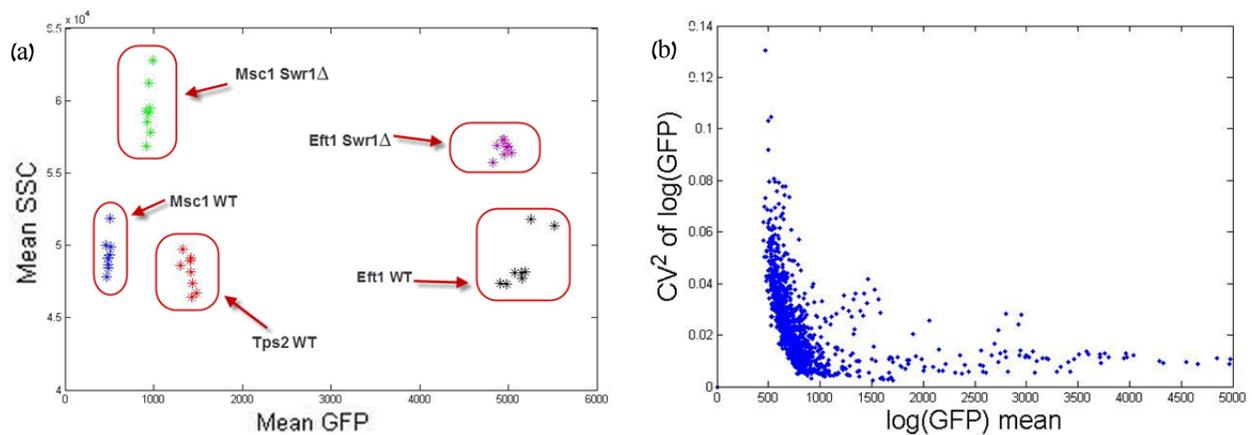


Figure 19: (a) The mean SSC as a function of the mean GFP for 8 biological repeats of 5 different strains. Repeated measurements of the same population are gathered together. The measurement system allows to distinguish between different populations and to identify repeats. (b) The noise (CV^2) of the GFP parameter as a function of the mean in ~180 populations of strains. Each strain has a GFP reporter on a different gene. Even reporters with very close mean GFP can have a significantly different noise and vice versa.

Although the median is a good indicator for the behavior of the entire population, working in single cell resolution allows us also to measure the cell-to-cell variation in the expression level of the proteins³. We calculated the expression noise – squared coefficient of variation (CV^2) - which is the variance of the log of BluFL1 parameter divided by its squared mean. We noticed that the expression noise can vary significantly between different proteins, even if the mean expression in the population is almost the same (Figure 19b). It seems like the cell-to-cell variation in the expression levels is a property that differs between proteins and may be related to their function or to the way they are regulated as suggested in previous works^{3,4,19}.

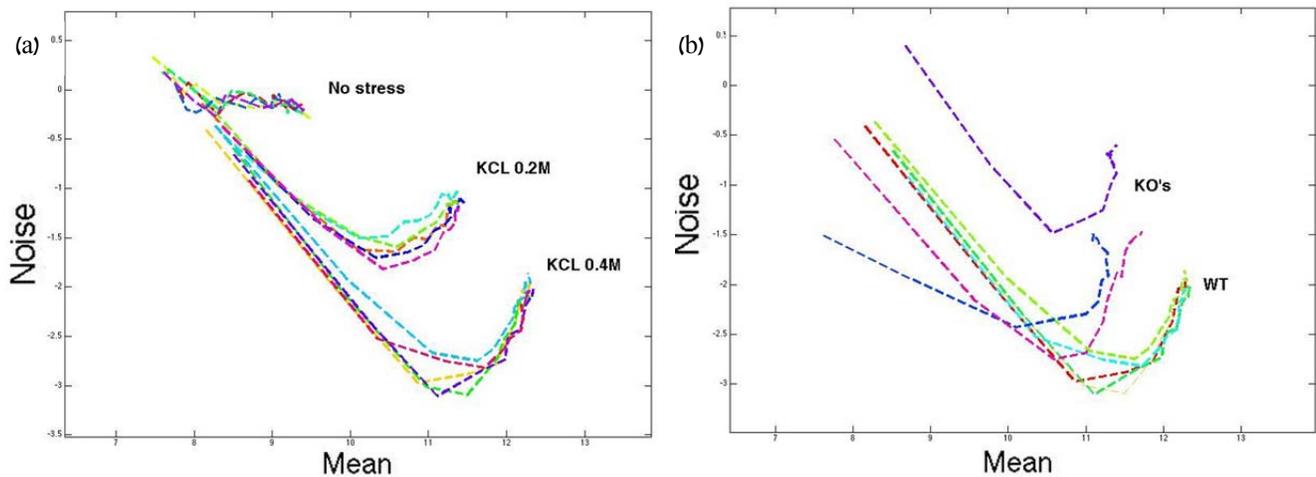


Figure 20: Each dotted line connects the results of 12 consecutive time points after exposure to stress (samples were taken in 10 minute intervals) (a) Mean GFP vs. noise in 3 time course experiments with 6 biological repeats in each. The noise parameter is reproducible between experiments. (b) Comparison of mean GFP vs. noise throughout the time course in strains from different genetic backgrounds (Knock out vs. WT strains). The noise parameter allows distinguishing between different samples.

To gauge the reliability of our estimates of cell to cell variability we evaluated their reproducibility. We used strains that contain the same stress-induced fluorescence. We exposed six biological repeats to severe and mild stress and measured a two hour time-course in 10 minute intervals. The results precisely repeated themselves (Figure 20a). In a different experiment with the same setup, we measured the fluorescence level of the same reporter strains with different genetic background and showed that the noise parameter shows different behavior in different samples and can be used to distinguish between strains (Figure 20b).

5 Application – General stress response network

In this chapter we will demonstrate the implementation of the robotic high throughput measurement system and show how it is used for dissection of transcription regulation networks in yeast. The biological mechanism that we chose to focus on is the yeast general stress response, a significant change in the gene expression program of the cell through the repression and induction of hundreds of different genes. The mechanism is activated upon the exposure of the cell to diverse environmental stress conditions, including heat, osmotic and oxidative stresses¹⁰. The partially redundant transcription factors Msn2/4 were identified as the master regulators of the process¹². Under regular conditions, Msn2/4 are located in the cytoplasm and upon stress they are shuttled into the nucleus and control the transcription of their target genes¹³ (Figure 21). Msn2/4 are located in the center of a complex regulatory network and their activity is regulated by different components and pathways. Some of the regulation pathways have been studied and described, however most of the network components and interactions are still partially characterized or completely uncharacterized. In this project we tried to map the regulation network upstream to Msn2/4 and understand how so many different stress conditions eventually converge on the same target.

5.1 Mapping the activators and the repressors of Msn2/4

To find the components in the general stress response network and to characterize their interaction with Msn2/4, we examined the effect of different genetic perturbations on Msn2/4 activity. We chose ~120 gene deletions from the yeast DaMP and KO libraries^{20,1}. Most of the genes were involved in signaling and stress response pathways. Few of the chosen genes had well characterized interactions with Msn2/4¹⁴. The others were predicted based on the literature and online interaction databases (BioGRID²¹ and String²²).

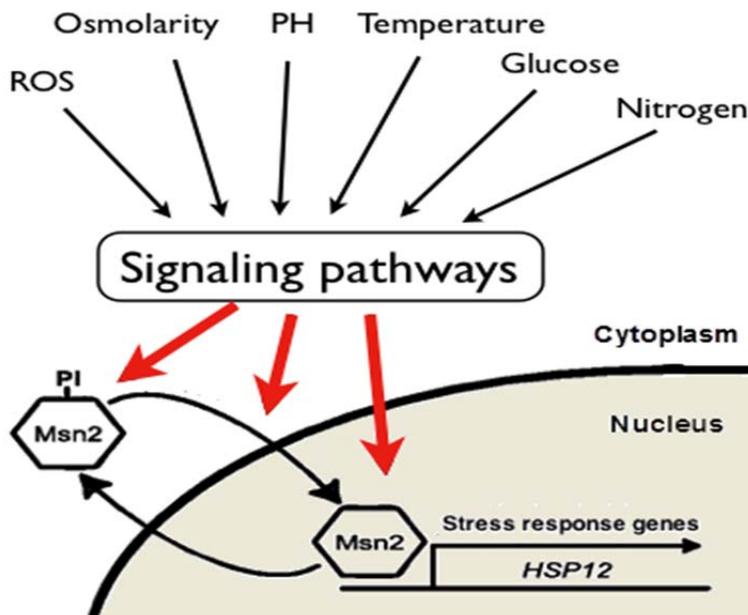


Figure 21: Msn2/4 are the master regulators of the general stress response. Under normal conditions Msn2/4 are localized to the cytoplasm and upon stress they are shuttled into the nucleus and control the transcription of hundreds of genes. In this project we tried to map the regulatory network upstream to Msn2/4.

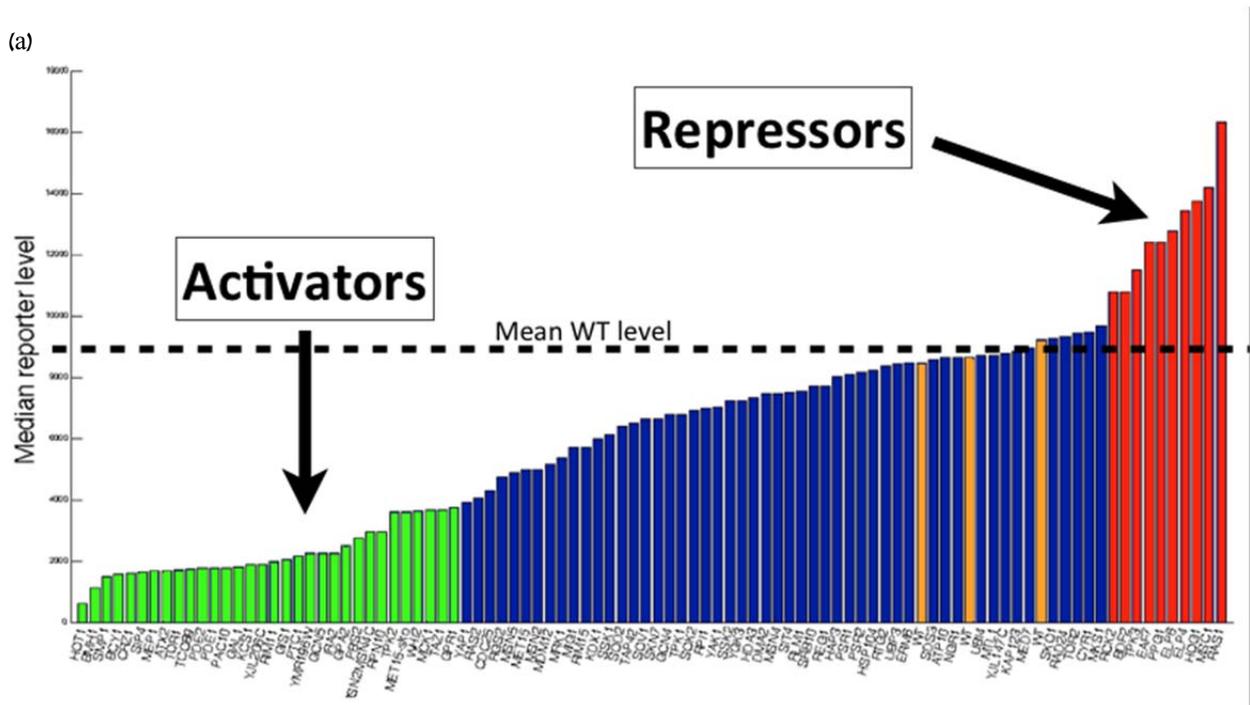
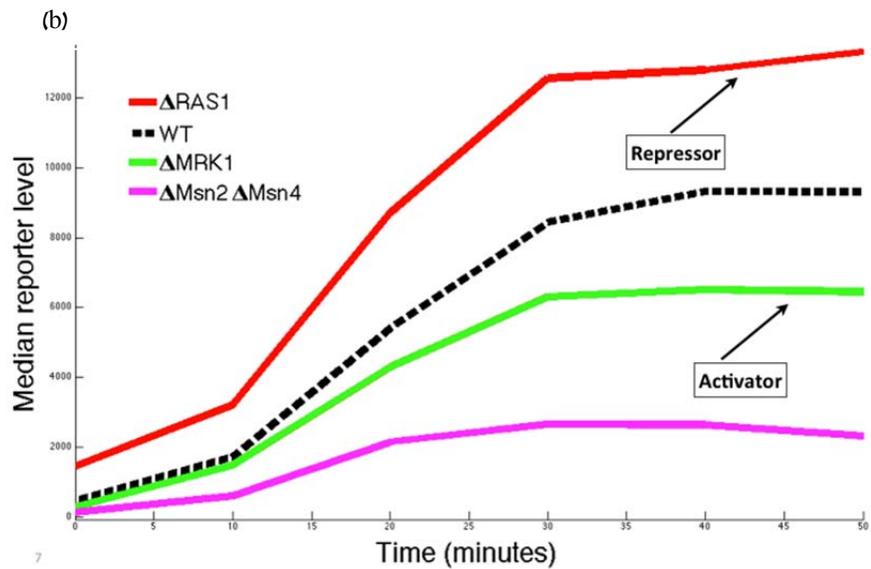


Figure 22: (a) The bar graph shows the median fluorescence level of 96 KO strains containing a GFP reporter for Msn2/4 activity two hours after the exposure to osmotic stress. The WT values are shown in orange. Based on these results, the knocked out genes can be roughly separated into three groups. Activators of Msn2/4 (green), repressors of Msn2/4 (red) and neutral to Msn2/4 (blue). (b) The plot shows the dynamics of the median fluorescence level of 4 strains throughout a two hour time-course, after the exposure to osmotic stress at t=0. The double KO of Msn2/4 shows a minimal induction of HSP12.



We mated the KO strains with a strain containing the HSP12 gene fused to green fluorescent protein (GFP). HSP12 is a general stress response gene encoding a plasma membrane protein, and its expression depends, almost exclusively on Msn2/4 function¹⁴ (Figure 22b). Thus the levels of the protein in the cell are a good measure for the recent activity levels of Msn2/4. We exposed the strains to osmotic stress (0.4M KCL) and measured them using our flow cytometry system in a two hour time-course. We calculated the median reporter level of each population at the end of the experiment and compared to the result of the WT strain (Figure 22a,b). Decrease or increase in the median reporter level relative to the WT strain identified proteins that either activate or repress

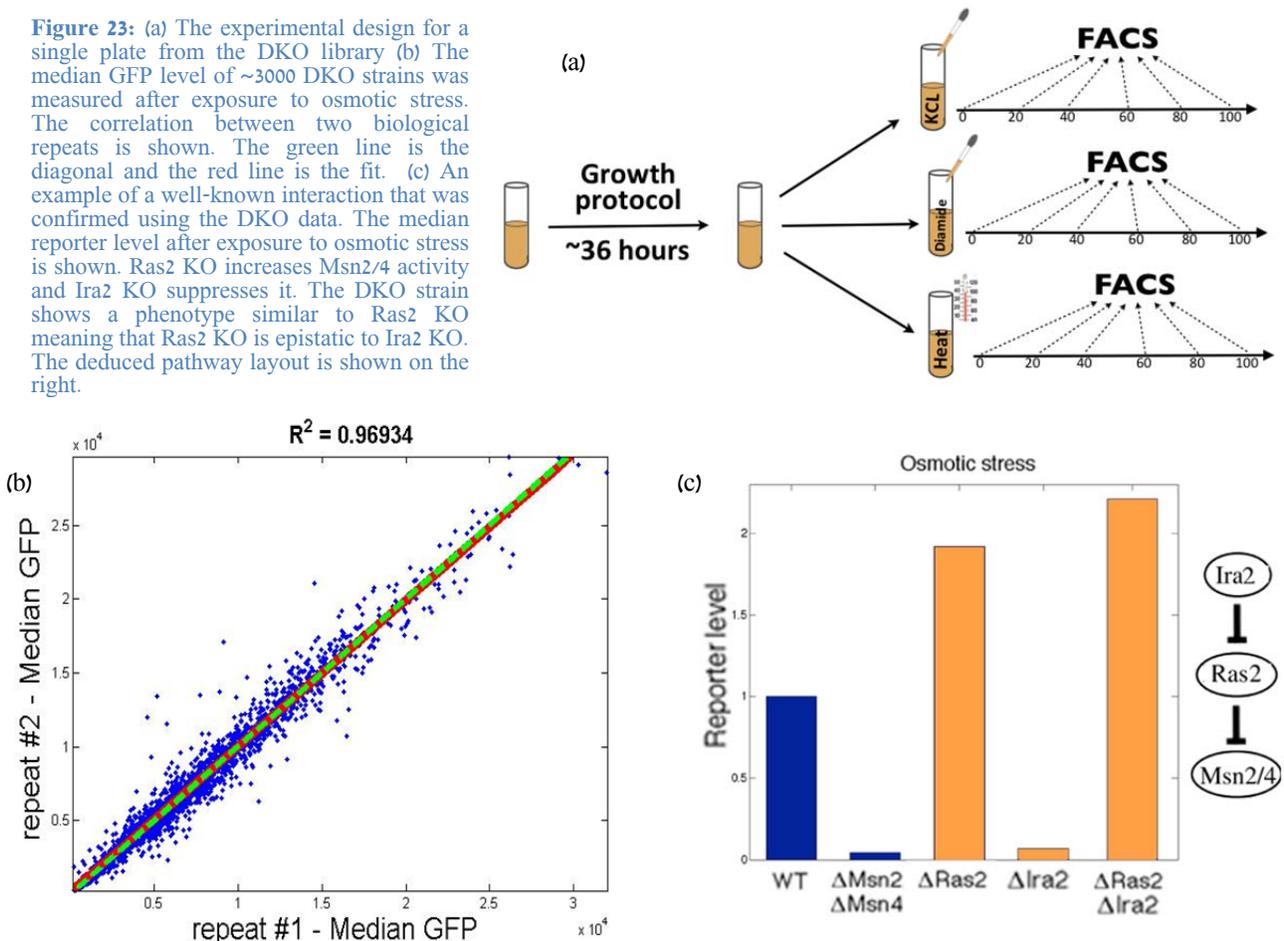
Msn2/4 activity following exposure to osmotic stress. The results were compatible with the literature and some novel interactions were discovered.

5.2 Screen of a DKO library

Learning the effect of a single gene on the activity of Msn2/4 is not sufficient for inferring the structure of the regulation network. For example a gene whose knockout reduces HSP12 expression might be a direct activator of Msn2/4, a repressor of an activator of Msn2/4, or many other alternative explanations. To gain additional insight into the structure of the network, we decided to find the genetic interactions between the genes. The genetic interaction can be measured by comparing the effect of the two SKO strains on the phenotype to the effect of the DKO. If the phenotype of the DKO strain can be derived from the phenotypes of the SKO strains using some neutrality function, we assume that there is no interaction between the proteins. Otherwise we can calculate the strength of the interaction and its sign^{5,23,24,25}.

We chose 32 out of the 120 SKO strains and crossed them against 96 SKO strains using the synthetic gene array (SGA) methodology²⁶, resulting in a DKO library of ~3000 strains. We

Figure 23: (a) The experimental design for a single plate from the DKO library (b) The median GFP level of ~3000 DKO strains was measured after exposure to osmotic stress. The correlation between two biological repeats is shown. The green line is the diagonal and the red line is the fit. (c) An example of a well-known interaction that was confirmed using the DKO data. The median reporter level after exposure to osmotic stress is shown. Ras2 KO increases Msn2/4 activity and Ira2 KO suppresses it. The DKO strain shows a phenotype similar to Ras2 KO meaning that Ras2 KO is epistatic to Ira2 KO. The deduced pathway layout is shown on the right.



defined an experimental scheme for a plate in the library (Figure 23a). Each plate passed through the robotic growth protocol (Robotic growth chapter) and was split to four 96 well microtiter plates. Each plate was exposed to different environmental stress, osmotic (0.4M KCL), oxidative (1.5mm diamide), heat (30 minutes in 37°C) and control. The osmotic and oxidative stress plates were sent immediately to the automated flow cytometry system and measured in a two hour time-course in 10 minute intervals. The heat stress and control plates were measured in a single point at the end of the experiment. We executed the entire experimental scheme for each plate in 2-3 biological repeats. The results of these biological replicate experiments were in good agreement (Figure 23b).

A simple initial analysis comparing the phenotypes of SKO strains to the phenotype of the DKO strains was sufficient to reveal many interesting interactions. One type of interactions which can be easily identified by the simple analysis are cases in which the phenotype of the DKO is similar to the phenotype of one of the SKO and different from the other SKO. These interactions are defined as *complete epistasis*; the effect of one mutation is masked by the effect of another²⁷. One of the epistasis cases that is compatible with the literature is described in Figure 23c.

5.3 Genetic interactions analysis

To identify interactions whose effect on the phenotype is more delicate and to be able to connect between pathways and components in the network, we took a more analytical approach. We decided to quantify the genetic interaction between each pair of genes as the difference between the observed phenotype of the DKO to its expected phenotype in case of a neutral interaction^{6,24,25}. Recent genetic interaction studies used different neutrality functions to calculate the expected phenotype, usually implementing a multiplicative or an additive one⁵. However, it is not obvious what the most suitable neutrality definition to use is, when the measured phenotype is single cell gene expression. We observed that the choice among alternative definitions can have profound consequences on the calculated interactions.

What directed us in the search for the right definition was the expectation that most of the gene pairs have weak interactions (i.e. if we plot the expected versus the observed values most of the points should lay around the diagonal). According to the additive definition the effect of gene X deletion is defined as $\Delta F_x = F_x - F_{wt}$, where F_x and F_{wt} are the fluorescence of the strains X and WT. We can then define the expected value of X,Y double deletion is $\Delta \hat{F}_{xy} = \Delta F_x + \Delta F_y$. That is, the increase (or decrease) in fluorescence brought by each mutation are summed to give the expected fluorescence in the double mutant strain. According to the multiplicative definition

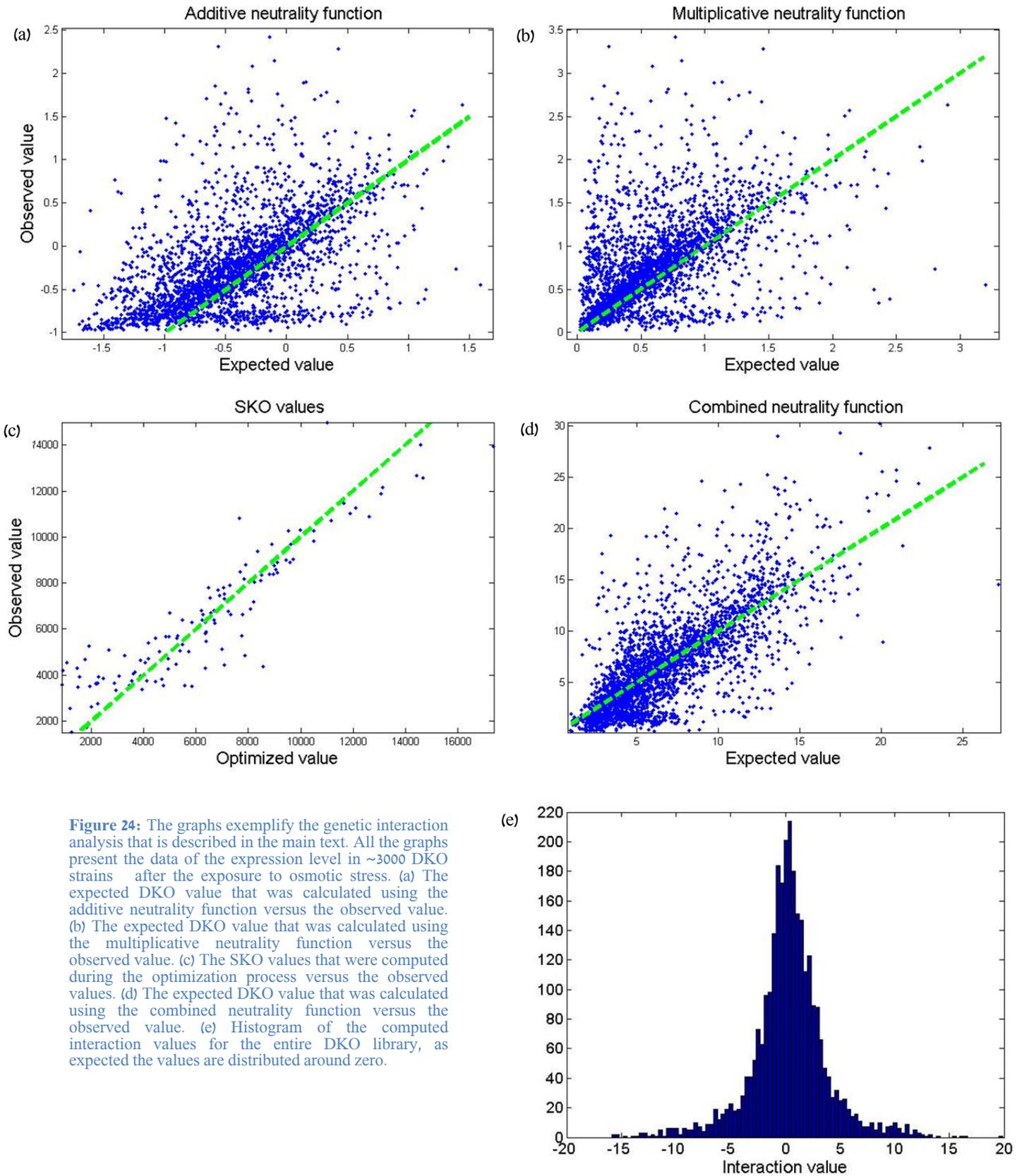


Figure 24: The graphs exemplify the genetic interaction analysis that is described in the main text. All the graphs present the data of the expression level in ~3000 DKO strains after the exposure to osmotic stress. (a) The expected DKO value that was calculated using the additive neutrality function versus the observed value. (b) The expected DKO value that was calculated using the multiplicative neutrality function versus the observed value. (c) The SKO values that were computed during the optimization process versus the observed values. (d) The expected DKO value that was calculated using the combined neutrality function versus the observed value. (e) Histogram of the computed interaction values for the entire DKO library, as expected the values are distributed around zero.

the effect of gene X deletion is defined as $\partial F_x = \frac{F_x}{F_{wt}}$ and the expected value of X,Y double deletion is $\partial \hat{F}_{xy} = \partial F_x * \partial F_y$. The interaction is then computed as $I^{\Delta}_{xy} = \Delta \hat{F}_{xy} - \Delta F_{xy}$ or $I^{\partial}_{xy} = \partial \hat{F}_{xy} - \partial F_{xy}$.

Following these definitions, to calculate the interactions in the data we need to know the phenotypes of the SKO strains. If the phenotypes were already measured and the results are reliable we can use them directly. Otherwise we can find the values by solving an optimization problem (i.e. find the 96+32 set of SKO values that will minimize the interactions in the data). When we employed the additive and multiplicative definitions on our data, using the measured SKO values, we received a large amount of interactions, implicating that the chosen neutrality function does not fit the data (Figure 24a,b). We then tried to combine the two definitions assuming that each gene has an additive (F^+_x) and a multiplicative (F^*_x) contribution to the phenotype (i.e. $F_x = F^+_x + F^*_x$). The expected value of the double KO is calculated as follows: $\hat{F}_{xy} = \frac{F^*_x * F^*_y}{F_{wt}} + sig(F^+_x + F^+_y - F^+_{wt})$. Where $sig(x)$ is a sigmoid type function (i.e. $(x) = (a_1 - a_2) \frac{1}{1 + a_3^{-(x+a_4)}} + a_2$) that accounts for the expected saturation in the extreme expression values of the reporter. We performed a constrained nonlinear optimization using the interior point algorithm²⁸, to find the sigmoid parameters and the multiplicative and additive portions of each SKO that minimize the differences between the expected and the observed values of the DKO. We first made sure that the optimized SKO values make sense and are correlated with the measured values (Figure 24c). This method identified a larger number of neutral interactions compared to the previous definitions and helped to emphasize the significant interactions in the data (Figure 24d). We computed the final interaction value for each DKO as described above and as expected the interaction values in the data are normally distributed around zero (Figure 24e). We repeated the analysis for all the different phenotypes that we measured in the screen, to produce interaction maps of the ~3000 strains (Figure 25).

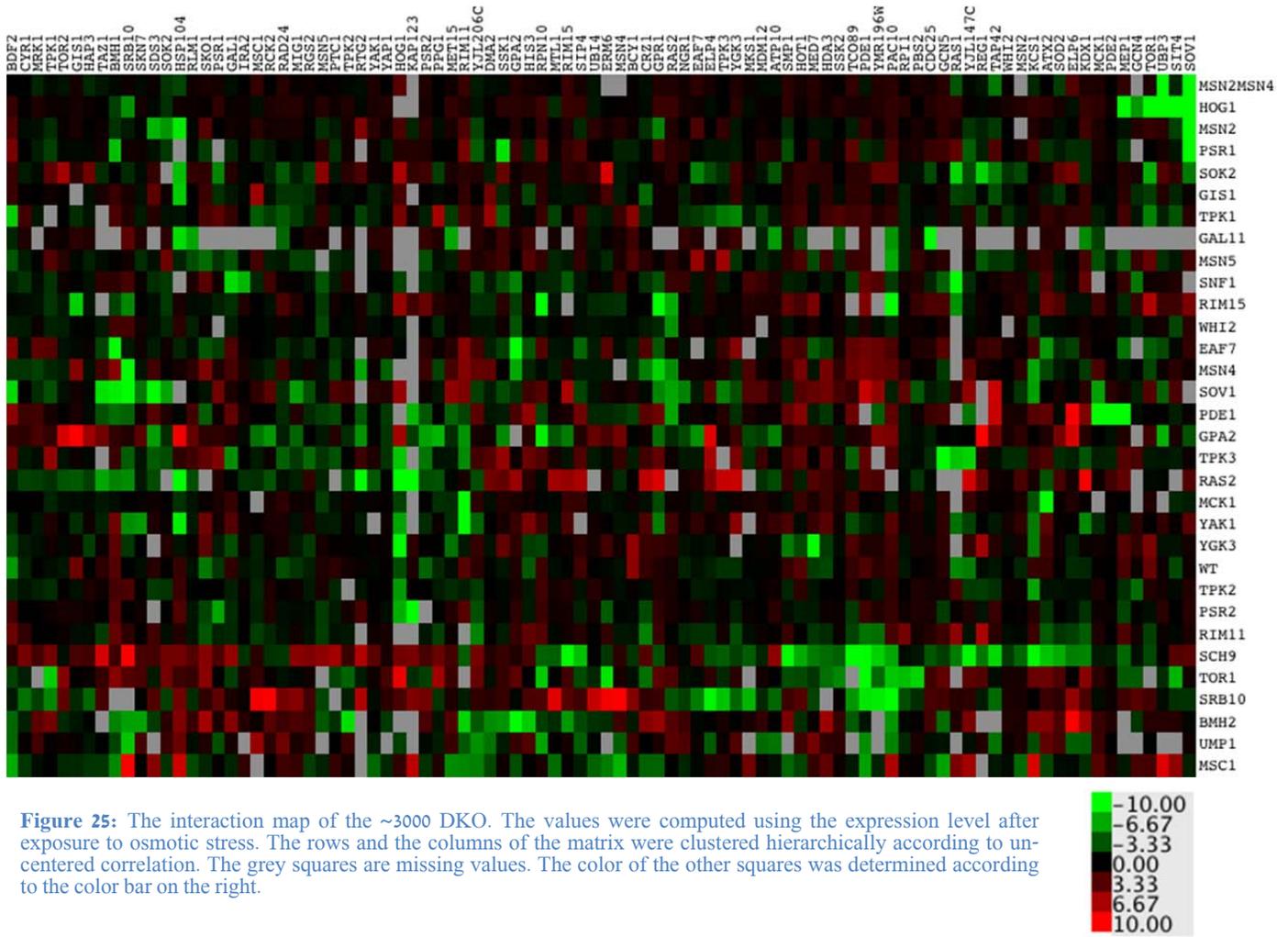


Figure 25: The interaction map of the ~3000 DKO. The values were computed using the expression level after exposure to osmotic stress. The rows and the columns of the matrix were clustered hierarchically according to uncentered correlation. The grey squares are missing values. The color of the other squares was determined according to the color bar on the right.

6 Discussion

In this work we established a robotic measurement system that allows high-throughput inspection of genetic interactions in the budding yeast. The interactions are calculated based on single cell gene expression data. One can separate the setup of the system into three main steps: Preparation of the sample, the measurement of gene expression, and the initial analysis of the data.

To systematically measure gene expression of yeast strains, we adapted an automated high throughput flow cytometry system. The system facilitates the running of time-lapse flow cytometer (FACS) measurements of an entire 96 wells microtiter plate with sampling frequency of 7 minutes per time point. We built a designated algorithm which performs the initial processing of the data and outputs it in a format suitable for further analysis. We connected the system to an existing robotic liquid handling system and enabled the scheduling of FACS experiments.

A key to the success of gene expression experiments, especially those that examine stress response, is making sure that all the measured strains are found in the same growth phase and that the last few hours of their growth were in optimal conditions. Thus, we designed and calibrated a novel automated high-throughput growth protocol. The protocol is addressing the challenge of bringing all the strains to the same, narrow density range at the beginning of the measurement experiment. So far the percentage success of the protocol is very high (~95%), even in plates with a wide range of doubling times (1.5-3.5 hours).

When dealing with massive amounts of data, automatic normalization, corrections and sanity checks are crucial. We presented here a series of procedures that we developed to handle these problems. These procedures include gating, fluctuations tests, bimodality checks and more. We are executing these procedures on the data on a regular manner as part of the data extraction process.

We showed that the system provides reproducible results when biological repeats are measured and some of the results are already shown to be compatible with the literature. The measurement system was successfully utilized over a few different yeast libraries. The throughput was four 96-well plates per day. We are currently working on moving the system to work with 384-well microtiter plates, which will significantly increase the throughput.

Our vision was to use this system to gain a better understanding of transcription regulation networks. We chose the yeast general stress response network as a first project. We created a collection of 96 SKO strains and ~3200 DKO strains that contain a fluorescent Msn2/4 activity reporter gene (HSP12-GFP). We followed the dynamics of Msn2/4 activity at a single cell level in

response to different environmental conditions in all the perturbed strains in our library. We determined the genetic interactions definition that best corresponded to the data and computed the interactions for all the couples.

We created interaction maps for the different phenotypes that we measured. In a further analysis we plan to utilize existing computational approaches^{7,9} to analyze these interaction maps, organize the components into modules and understand their hierarchy. It is quite possible that the existing approaches will not suit the data that we have due to the large amount of interactions in the subset of genes that we chose. In this case, we will need to develop novel computational approaches. Comparing between the outcomes of the analysis for different interaction maps will allow us to identify interactions and pathways that function only in part of the environmental conditions.

The data analysis methods that were described in chapter 5, used the median fluorescence level as the main phenotype. However, our measurement system allows us to work with more complex phenotypes. We can take advantage of the time-course data to extract parameters that describe the dynamics of the transcription response. We can measure cell-to-cell variability in a single time point or look at its dynamics. Examining these phenotypes can help us to identify more delicate effects of single and double perturbations. Moreover, these phenotypes can be used as well to compute genetic interactions in the network and to find different types of epistatic effects.

7 Acknowledgments

I would like to dedicate a few words of thanks to all the people who assisted and aided me throughout the last few years in the lab.

First many thanks to my supervisor, Nir Friedman, for all the assistance and guidance, for encouraging and challenging, for the short conversations and the long scientific discussions and for teaching me something new on a regular basis.

Thanks to Ayelet Rahat the lab manager, for the help and support through every single thing.

Thanks to Avital Klein for the joint work on the establishment of the robotic protocols and the constant willingness to aid.

Thanks to Amit Sadeh for the collaborative work on the general stress response project (and many more), always with a good sense of humor and invigorating spirit.

Thanks to Ariel Jaimovich who guided me at my first days in the lab.

I wish to thank all the members of the Friedman lab, for the wonderful atmosphere, good spirit and the nutritional salads.

8 Bibliography

- ¹ Giaever et al. "**Functional Profiling of the Saccharomyces Cerevisiae Genome.**" Nature 418, 2002
- ² Huh et al. "**Global Analysis of Protein Localization in Budding Yeast.**" Nature 425, 2003
- ³ Newman et al. "**Single-cell Proteomic Analysis of S. Cerevisiae Reveals the Architecture of Biological Noise.**" Nature 441, 2006
- ⁴ Rinott et al. "**Exploring transcription regulation through cell-to-cell variability.**" PNAS 108(15), 2011
- ⁵ Mani et al. "**Defining genetic interaction.**" Proc Natl Acad Sci. 105(9), 2008
- ⁶ Segre et al. "**Modular epistasis in yeast metabolism.**" Nat Genet. 37(1), 2004
- ⁷ A Beyer, S Bandyopadhyay, and T Ideker. "**Integrating physical and genetic maps: from genomes to interaction networks.**" Nat Rev Genet. 8(9), 2007.
- ⁸ E Segal, M Shapira, et al. "**Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.**" Nat Genet. 34(2), 2003
- ⁹ Jaimovich A et al. "**Modularity and Directionality In Genetic Interaction Maps**". Bioinformatics 26(12), 2010
- ¹⁰ Gasch et al. "**Genomic expression programs in the response of yeast cells to environmental changes.**" Mol Biol Cell 11, 2000
- ¹¹ Berry DB, Gasch AP "**Stress-activated genomic expression changes serve a preparative role for impending stress in yeast.**" Mol Biol Cell. 19(11), 2008
- ¹² Boy-Marcotte et al. "**Msn2p and Msn4p control a large number of genes induced at the diauxic transition which are repressed by cyclic AMP in Saccharomyces cerevisiae.**" J Bacteriol 180, 1998
- ¹³ Gorner et al. "**Nuclear localization of the C2H2 zinc finger protein Msn2p is regulated by stress and protein kinase A activity.**" Genes Dev 12, 1998
- ¹⁴ Sadeh et al. "**Fine-tuning of the Msn2/4-mediated yeast stress responses as revealed by systematic deletion of Msn2/4 partners.**" Mol Biol Cell. 22(17), 2011
- ¹⁵ Needleman, S., and C. Wunsch. "**A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins.**" Journal of Molecular Biology 48.3, 1970
- ¹⁶ <http://www.sikuli.org>
- ¹⁷ Costanzo et al. "**The genetic landscape of a cell.**" Science 327(5964), 2010
- ¹⁸ Ramirez, Sklar and Edward, "**High-Throughput Flow Cytometry: Validation in Microvolume Bioassays**" Cytometry Part A 53A, 2003
- ¹⁹ Bar-Even et al. "**Noise in Protein Expression Scales with Natural Protein Abundance.**" Nature Genetics 38.6, 2006

-
- ²⁰ Breslow et al., “**A comprehensive strategy enabling high-resolution functional analysis of the yeast genome.**” Nat. Methods. 5(8), 2008
- ²¹ Stark C et al. “**BioGRID: A General Repository for Interaction Datasets**”. Nucleic Acids Research 34 , 2006
- ²² Jensen et al.” **STRING 8--a global view on proteins and their functional interactions in 630 organisms.**” Nucleic Acids Res. 37 , 2009
- ²³ Jonikas MC et al. “**Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum**”. Science 323(5922) , 2009
- ²⁴ St Onge et al. “**Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions**” Nature Genetics 39, 2007
- ²⁵ Schuldiner, M. et al. “**Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile.**” Cell 123, 2005
- ²⁶ A. Tong and C. Boone “**Synthetic Genetic Array (SGA) Analysis in Saccharomyces cerevisiae.**” Methods Mol Biol. 313, 2006
- ²⁷ Cordell, Heather J. “**Epistasis: what it means, what it doesn’t mean, and statistical methods to detect it in humans**”. Human Molecular Genetics 20 , 2002
- ²⁸ Waltz, R. A., J. L. Morales, J. Nocedal, and D. Orban, “**An interior algorithm for nonlinear optimization that combines line search and trust region steps**” Mathematical Programming 107, 2006.