# Effects of repeated osmotic stress on gene expression and growth: from cell-to-cell variability to cellular individuality in the budding yeast Saccharomyces cerevisiae

Artémis Llamosi

▶ **To cite this version:**

Artémis Llamosi. Effects of repeated osmotic stress on gene expression and growth: from cell-to-cell variability to cellular individuality in the budding yeast Saccharomyces cerevisiae. Quantitative Methods [q-bio.QM]. Université Paris Diderot, 2015. English. NNT : . tel-01253235

**HAL Id: tel-01253235**

**https://theses.hal.science/tel-01253235**

Submitted on 11 Jan 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**UNIVERSITE SORBONNE PARIS CITE**

***UNIVERSITE PARIS DIDEROT (PARIS 7)***

**THÈSE**

**Pour obtenir le grade de DOCTEUR DE L'UNIVERSITÉ PARIS DIDEROT**

**Spécialité: Biologie synthétique et systémique**

**École Doctorale Frontières du Vivant (ED 474)**

**Laboratoire Matière et Systèmes Complexes (MSC)**

**INRIA Paris Rocquencourt.**

# Effects of repeated osmotic stress on gene expression and growth: from cell-to-cell variability to cellular individuality in the budding yeast *Saccharomyces cerevisiae*.

---

Présentée par

**Artemis Llamosi**

dirigée par Pascal HERSEN et Gregory BATT

Soutenue le 15 décembre 2015

| | | |
|---|---|---|
| Marc LAVIELLE | Directeur de recherche, INRIA | Rapporteur |
| Gael YVERT | Directeur de recherche, CNRS | Rapporteur |
| Peter SWAIN | Professeur, University of Edinburgh | Examinateur |
| Heinz KOEPPL | Professeur, TU Darmstadt | Examinateur |
| Gregory BATT | Chargé de recherche, INRIA | Directeur de thèse |
| Pascal HERSEN | Directeur de recherche, CNRS | Directeur de thèse |

M. C. Escher. *Reptiles*. 1943

# Acknowledgments

Many people have contributed to the work presented here, through scientific exchange, personal support or both.

I had the opportunity to work with many interns during this project: Sebastian Jaramillo Riveri, Matt Deyell, Rémi Sieskind, Alice Llamosi, and Antonio Villarreal Larraui. Thank you all for helping me, whether by lightening my share of ungrateful repetitive tasks or in taking the risk of exploring new research directions.

I would like to thank all the members of the MSC lab and the Contraintes/Lifeware team, in particular Jean-Marc Di Meglio, François Fages, Benoit Sorre, Gaëlle Charron, Mathieu Receveur, Arnaud Grados et David Pereira.

I obviously want to thank all the members of the Lab 513, both past and present. First of all, I would like to thank Jannis Uhlendorf for teaching me so much about working with yeast, doing microfluidics, keeping calm with molecular biology, and tuning microscopes. Also, I am very grateful to Clément Vulin, a friend and a teacher to me. Thanks to you I know why yeast is awesome and I have filled many blanks in my homemade biology education. I would like to thank Jean-Baptiste Lugagne, Xavier Duportet, Zoran Marinkovic, Adrien Halou, and Zacchari Ben Meriem for the very good vibes in the lab, the fruitful discussions and the troubleshooting support.

I would like to thank my collaborators and in particular Andres Gonzalez-Vargas for working with me side by side (although most of the time at distance) with patience and strength. I thank Giancarlo Ferrari-Trecate and Eugenio Cinquemani for being our mathematical backbone and for our discussions. I thank Cristian Versari for our collaboration on Cell* and the immense amount of effort you've placed in order to create a truly great tool.

I thank again Eugenio Cinquemani along with Frédéric Devaux for our yearly advisory committee, which was always useful and benevolent.

I would like to thank Véronique Letort, Florence d'Alché-Buc and Thomas Landrain who acted as enzymes in catalyzing my transformations from engineering to biology prior to my PhD.

I thank my family for their support and Séverine, my life partner who shares the ups and downs of a PhD and is always ready for geek's chats and late philosophical and scientific debates.

Obviously, I will never be able to thank enough my PhD advisors, Pascal Hersen and Gregory Batt. Thank you for believing in me, for your multidimensional support and advice, for your patience, and for your subtle blend of profound expertise and humility.

At last, I would like to thank the billions of yeast cell, which, unwillingly, have been the core of this project. Thank you for being so fascinating and for bread,… and beer,… and wine.

# Table of Contents

Acknowledgments ..... 5

Abstract ..... 11

Foreword: Why engineers should study cells? ..... 13

General introduction ..... 15

I. Introduction ..... 21

1. Dealing with variability and time scale in gene expression ..... 21

a. Twins are not identical ..... 21

b. What a difference a day makes? ..... 26

2. Measurements at the single cell level ..... 31

3. A synthetic and systems biology approach ..... 35

a. Cells as systems ..... 35

b. Experimenting within a cell: Synthetic biology and microfluidics ..... 38

4. *S. cerevisiae* response to osmotic stress ..... 41

a. An overview of the HOG response ..... 41

b. Yeast response to osmotic stress as a model cellular process ..... 46

c. Modelling the cellular response to osmotic stress ..... 52

5. Introduction Conclusion and Outline ..... 54

II. Long term dynamic experiments and single-cell data ..... 55

1. Single-cell measurements in precisely changing environments using microfluidics and microscopy ..... 56

a. The Truman show: the use of microfluidics ..... 56

b. Fluorescent probes to peep into cellular activity ..... 61

2. Image Analysis ..... 64

a. Segmentation and Tracking using Cell* ..... 64

b. Measures of cellular identity ..... 66

3. Measuring growth in populations and single cells ..... 69

a. Going beyond the field of view: An Eulerian measure of population growth ..... 69

b. Measuring growth at the single cell level ..... 72

4. Conclusions on: long term dynamic experiments and single cell data ..... 74

III. Individuality in the transcriptional response to osmotic stress ..... 75

1. Modelling dynamics of gene expression at the single cell level ..... 75

a. pSTL1 as a reporter of HOG transcriptional response ..... 75

b. Representing variability in pSTL1 gene expression with stochastic models....80
c. Identifiability of extrinsic and stochastic models of gene expression at the single-cell level84
2. Mixed effects models of pSTL1 expression....87
a. Building a single-cell model of pSTL1 expression including cell-to-cell variability....87
b. Representing extrinsic variability with using Mixed-effects models....89
c. Estimating population and single-cell models and validating them....90
3. Cellular identity and gene expression....97
a. Relations between gene expression and cell physiology....97
b. Inheritance of phenotype and gene expression features....98
c. Listening to the noise: harvesting natural cell to cell variability....101
4. Conclusions on: Individuality in the transcriptional response to osmotic stress....103
IV. The impact of repeated stress on cellular proliferation....109
1. An integrated view of the response to osmotic stress....111
a. How osmotic stress affects growth and division?....111
b. Proliferation quantification at the single cell level: a matter of point of view....113
2. The impact of osmotic stress on the cell cycle....116
a. Osmotic stress can trigger phase-dependent arrest of the cell cycle....116
b. Nuclear separation is perturbed by osmotic stress....119
c. Timing of cell cycle arrest and partial lock-in....123
d. Lock-in phenomenon....127
3. The impact of osmotic stress on metabolism....131
a. Metabolism shifts upon osmotic stress....131
b. Quantifying adaptation variable cost....132
c. Quantifying acclimation costs of osmotic fluctuation....136
4. Conclusion: The impact of osmotic stress on colony growth dynamics....140
V. Perspectives and final discussion....143
1. Experimental pipelines for systems biology at the single-cell level....143
2. Cellular variability and context....147
List of abbreviations....153
References....154
Appendix....167
1. List of Strains....168
2. The use of S. cerevisiae....169
3. Transcriptome time course in response to hyperosmotic stress....172

4. Custom microfluidic chips fabrication method ..... 173
5. Glucose diffusion and consumption in microfluidic chambers ..... 176
6. Single-cell parameter estimation of models of gene expression (article, submitted version) 185
7. Simulation of Eigen cell behavior ..... 230
8. Developing an Open Source, single-cell optogenetic system ..... 231

# Abstract

301 words

When shifted to a stressful environment, cells are capable of complex response and adaptations. Although the cellular response to a single stress has been studied in great detail, very little is known when it comes to dynamically fluctuating stressful environments. In addition, in the context of stress response, the role of cell-to-cell variability in cellular processes and more specifically in gene expression is still unclear.

In this work, we use a systems and synthetic biology approach to investigate osmotic stress in S. cerevisiae at the single cell level. Combining microfluidics, fluorescent microscopy and advanced image analysis, we are able to subject cells to precise fluctuating osmolarity and monitor single-cell temporal response.

While much previous research in gene expression heterogeneity focused on its stochastic aspect, we consider here long-lasting differences between cells regarding expression kinetics. Using population models and state-of-the-art statistical analysis, we manage to represent both population and single-cell dynamics in a single concise modelling framework. This quantitative approach capturing stable individuality in gene expression dynamics can define a form of non-genetic cellular identity.

To improve our comprehension of the biological interpretation of such identity, we investigate the relation between single-cell specificities in their gene expression with their phenotype and micro-environment. We then take a lineage based perspective and find this form of identity to be partially inherited.

Understanding the evolutionary consequences of inheritable non-genetic cellular identity requires a better knowledge of the impact of fluctuating stress on cell proliferation. Dissecting quantitatively the consequences of repeated stress on cell-cycle and growth gives us an overview of the energetic and temporal consequences of repeated stress. At last, technical and theoretical developments needed to carry this investigation further are presented. These include the use of automated experimental design, both offline and online through real-time experimental design and single-cell real-time control of gene expression.

# Foreword: Why engineers should study cells?

Before I begin exposing the research I undertook during these past years, I would like to share briefly the motivation which drove me into this project. I always have been curious to understand how things work. My first focus was on how machines works, which extended to how the physical worlds works. Studying physics and engineering allowed me to satisfy this appetite and to grasp the basics of how planes fly, computers compute and nuclear reactors delivers power.

Understanding how things work is my way of being more conscious of the world, it enhances my perception of what surrounds me and therefore, is a kind of philosophical need. Many systems designed by men might be complex, they nevertheless are well defined. Each part serves a precise set of purposes, plays a precise role. This however is not the case in the natural world. As my knowledge of physical objects and machines was increasing, I got interested in less domesticated complex systems like human organizations and economics for instance and soon enough, I realized that *living things* were the most complex, ill-defined and fascinating systems to be. Not only their raw complexity (in terms of components etc.) is gigantic, but the interactions among them, which distinguish a bag of chemical from a living thing, can only be described as vertiginous. For a long time, I had a biased and very old-fashioned view of biology. It seemed like living systems were so far from our understanding capacity that we would be forever limited to the collection of a series of empirical facts while lacking this great feeling of hidden simplicity one gets when studying physics.

While studying computer science, I got interested by complex optimization problems (termed NP hard of NP complete problems) for which computing the optimal solution is unfeasible due to the number of possible solutions to be tested. Since the exact optimal solution is beyond reach, heuristics or meta-heuristics methods are employed which seek at finding a “good enough” solution in reasonable time. Interestingly, two very performant algorithms to address such problems are inspired by natural systems. Swarm algorithms mimic how ants forage for food and genetic algorithm how natural selection acts upon individuals. Working with these algorithms, I realized that nature’s way of solving problems was not only incredibly efficient, but surprisingly general and versatile: when fancy heuristics require usually much features of the problem at hands to be included in their conception, natural algorithms can adapt rather autonomously. Also, it showed me that some fundamental knowledge of biological system was in fact accessible, and that it was a potential source of ground breaking innovations.

As I learned more on current biology, I realized this classic field of study was undergoing a profound revolution. Mathematics and quantification were slowly but steadily revealing principles in biological processes and new experimental techniques were probing the inner working of living systems always further. Although being a complete layman in biology, I realized my skills in quantitative analysis could be applied to the study of biological systems. I was also pleased to see that biological experimentation can benefits from crafting and hacking which are among my favorite activities. It is not without difficulty that I learned biology from scratch. But learning the basics while reading the latest research has an interesting consequence: it allows avoiding many *unlearning* steps

related to the fact textbooks and school programs are rapidly outdated by recent findings (just think of how many unlearning a student in physics is subjected to). In an interdisciplinary endeavor, such as understanding a cell, I think it is essential not only to discuss a common problem among different experts, but for all collaborators to learn from each other field and if possible, have hands on practice. From my experience of diving into biology, I would say that besides the hardships, there is not a single day where I regret it, for every difficulty is well paid off by the enlightenment the study of cells brings me.

More generally, as an engineer, my purpose is also to interact with the world, to modify, to design tools and systems. Acting upon complex systems is a delicate endeavor; a good design should not require more energy than necessary and must avoid unforeseen consequences. In order to do that, engineers use top down design principles which are challenged by how evolution has shaped living systems. Evolution works rather bottom-up, generating diversity which is most of the time useless or even harmful, until it somehow stumbles upon some improvement. Natural systems can also be made adaptable, which is a fairly rare property in the realm of machines. At last, many of the current engineering challenges have to do with life cycle management of products, seeking environmentally compatible alternatives to common products. In this respect, all the natural cycles (oxygen, water, carbon etc.) constitute the best examples to get inspiration from. By studying biology, engineers can learn new design principles; they can adapt natural tools to find other solutions. For all these reasons engineers can both participate to the current revolution in Biology and benefit from this knowledge to improve traditional engineering.

# General introduction

In this section the model organism used for this project will be presented briefly. This description aims at sharing a conception of what a cell is and does from the point of view of a physicist or engineer.

*Saccharomyces cerevisiae* is a species of yeast which can be naturally found on mature grape fruits among other environments. It is a unicellular organism which can be probably considered as the first microbe which has been domesticated by mankind (although unconsciously and in an unspecific manner at first since the proper isolation of this species occurred during the XIX$^{th}$ century). Its role in bread, wine or beer making (thanks to fermentation) explains why it was already used in ancient times. It was probably domesticated separately in various regions of the globe whereas when it comes to winemaking, it seemingly originated from Mesopotamia around 10 000 years ago and its evolution has been coupled to human's since (1). Its significant contributions to humanity explain why it is also called Baker's yeast or Brewer's yeast. Not surprisingly, baker's yeast is one of the most common model organisms in Biology. Accordingly, there is a significant amount of literature available on *S. cerevisiae* describing decades of research on countless aspects of this yeast. As of today it is still a model organism of choice for new fields of studies (2). Readers unfamiliar with budding yeast (as it is also called) can refer to a short presentation in Annex 2 which also mentions some advantages of using this organism in research.

We will now present briefly cells in general and baker's yeast in particular through an engineer perspective: that is, comparing cells to machines. Such description rely on many estimated of measured values which, when applicable, will be referenced through their Bio Numbers IDentifiers (BNID)[1]. A cell can be very crudely described as a tiny bag (or bubble) made out of fat (a bi-lipid layer) which encloses a bunch of chemicals. For a yeast cell, we speak of a 4 µm diameter object (BNID 108258). To be even blunter, a yeast cell composition is roughly water (60% of mass) and $C{:}H_{1.61}{:}O_{0.56}{:}N_{0.16}$ (BNID 103689, 101801 and (3)). This very rude chemical description exemplifies the role of the observation scale one can take when looking at cells. Now that molecular biology made it possible to think precisely of each of the molecules composing a cell and that super resolution imaging allows for single molecule dynamic *in vivo* measurements, a cell never seemed so big and complex.

At the molecular level, this bag of chemical hosts an incredible amount of chemical reactions. Metabolism, which accounts for a large part of the normal day to day chemical life of cells, is a ballet featuring 584 types of dancers (metabolites) involved in 1175 choreographies (chemical reactions) (BNID 100647). Several molecules play special roles in generating this busy chemical activity. Enzymes for instance are molecules which act as catalyzers. This means they somehow *channel*

[1] Measuring or estimating any biological quantity is usually a very difficult task but we will not quote all the studies were these numbers come from, but instead their Bio Numbers identifier will be provided. This is intended both for readability and because we would like to highlight the importance of the "bionumbers" initiative (http://bionumbers.hms.harvard.edu/) which is a great tool to rapidly put a number, however crude it is, on many aspects of biology.

chemical reactions without being transformed themselves. In the other hand, metabolites are transformed one into another. The most fundamental characteristic of cells is their ability to replicate. A cell imports from its environment several molecules such as sugars or oxygen which will be used to increase its metabolites pools. From metabolites pools, a cell will, among other things, produce more cells. Several metabolites can be considered final in the sense they cannot be converted anymore. These are therefore considered as waste from the cellular perspective and will usually be released in the environment. When we consider cellular metabolism as a process which takes inputs, creates products and waste, we can see cells as microscopic factories. Interestingly, when compared to actual factories, cells appear as very efficient ones. In fact, metabolites pools (*i.e.* intermediate stocks) are maintained at very low levels but are renewed completely in a matter of seconds (BNID 109701). Minimizing intermediate stocks is a founding concept in lean manufacturing which is advocated by many as the most efficient manufacturing organization. In actual factories, achieving such an efficient production pipeline requires specific coordination mechanisms. In cells such coordination is achieved through two sets of mechanisms. In one hand, several cellular components are capable of information processing and actively coordinate cellular activity. In the other hand, physical constrains along with economic considerations of the cellular context impose indirect co-regulations.

***Gene expression and cellular information processing***

In yeast as in any cell, a large part of the cellular dry mass (40 to 50%, BNID 108200) is accounted as proteins. For simplicity, we can consider proteins as macromolecules composed of chains of elementary building block molecules (amino acids). Thanks to dedicated molecular machinery, cells can synthesize proteins from amino acids and breakdown existing proteins into these building blocks again. The synthesis of a given functional protein requires a precise sequence of amino acids to be assembled. The *blueprint* for a protein is stored in a coded form within the famous DNA molecule which comprises on a single molecule thousands of blueprints which are called genes. The information about the production of a given gene which is stored in DNA is first copied (transcribed) into another molecule: mRNA. It is this copy of the genetic information which is actually used as a template to produce proteins having a sequence of amino acids determined by the gene (4). Overall we speak of gene expression to describe the production of proteins from genes.

Proteins in turn serve a multitude of functions for the cell and can be compared to actuators and operators in a regular factory. In Figure 1 we report a synthetic representation of baker's yeast proteome (*i.e.* the ensemble of all its proteins) where both abundance and function of most proteins are represented. The different panels allow us to progressively zoom in and out from the functional territories composing the proteome. To some extent, this representation of the proteome shows how cells invest their matter (and energy) in various processes. In particular, proteins play a primal role in driving or channeling the countless metabolic reactions which allow matter and energy transformation. Accordingly, a major portion of proteins is mainly dedicated to anabolism and catabolism with a roughly equivalent investment in these complementary aspects of metabolism (represented in orange/brown in Figure 1). The second largest functional territory of the proteome (depicted in blue in Figure 1) concerns *genetic information processing* and includes proteins which are involved in gene expression (*e.g.* synthesizing proteins from genes).

Proteins are the main actuators of cellular processes, and accordingly, regulation of cellular activity can be achieved by modulating the level of expression of genes into proteins. For instance, the increased expression of a gene encoding an enzyme can change the rate at which a metabolic reaction is performed. Yet coordination itself results from the cellular capacity to store, acquire, transmit and process information. A cell is not only a microscopic factory but can also act as a

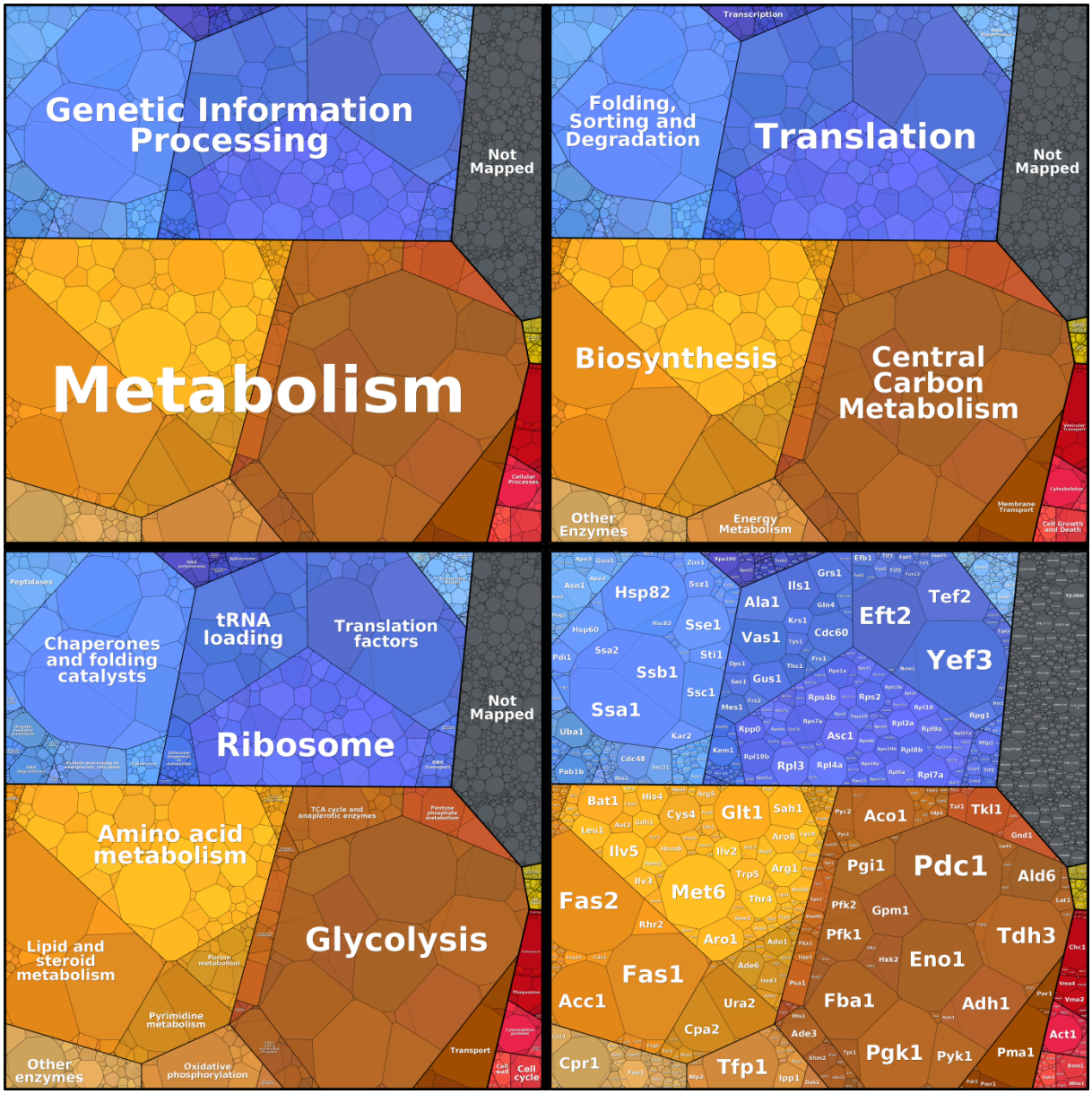



**Figure 1 - Voronoi representation of *S. cerevisiae* proteome. Abundancy of proteins was measured by mass spectrometry (data from (32)) and visualization is described in (73). The area of the Voronoi regions correspond to protein abundance and their position in the diagram to how related the corresponding proteins are (based on KEGG pathway database). Panels show different levels of abstraction from global functional classes to the protein names themselves.**

computer. Many different molecular mechanisms compose the bits and functions which allow cells to deal with information.

Besides the basic role of DNA as a recipe book for gene expression to cook any protein, it is important to grasp the information processing capability that is carried by proteins, RNAs and DNA. In fact, several proteins or other molecules can directly affect the level of expression of specific genes. Such elements are generically called *transcription factors*. Because some genes express proteins which in turn affect the expression of other genes (or of themselves), complex expression patterns can emerge. Therefore, it is the interaction of genes upon each other through transcriptional regulation which allows complex information processing in cells which are deprived of brains or nervous systems. All these interactions form what are called gene regulatory networks (GRN) where genes are nodes and interactions are edges. Figure 2 reproduces a recently estimated (partial) regulatory network in *S. cerevisiae*. Such visualization helps us assessing the level of interaction between genes and functional clusters of genes. It also gives some feeling of the task at hand to rigorously reconstruct and study such a network.

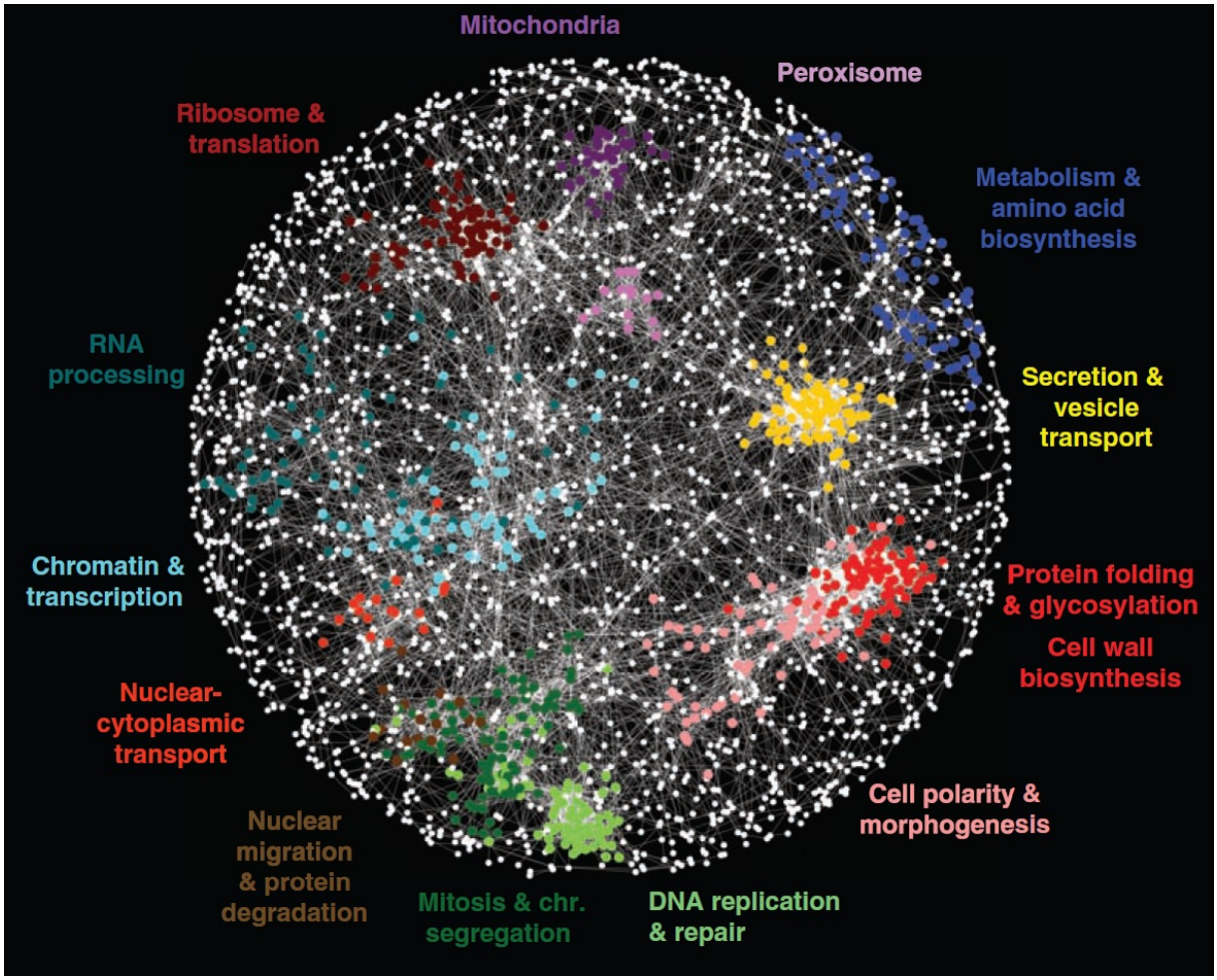



**Figure 2 – Partially reconstructed gene regulatory network of *S. cerevisiae* based on double mutant fitness screening and gene interaction similarity. Figure from (157). White dots represent genes and edges are estimated interactions. Colors code for functions associated to some genes.**

In addition to these *classic* genetic interactions, where gene expression is regulated at the level of transcription, other additional regulatory mechanisms can alter either the production of a protein or the activity of it. In post-transcriptional regulation mechanisms, the production of a given protein is regulated after mRNA has been transcribed (*e.g.* mRNA processing and regulation, translation

regulation etc.). Also, epigenetic mechanisms allow many other layers of regulation (both gene specific and non-specific).

Besides gene expression control, a protein's activity can be further regulated through molecular *signals* or *marks*. Indeed, some proteins can exist in active or inactive states with the transition from a state to another depending on the presence or absence of a specific chemical group. For example, phosphorylation is a common molecular signal which can condition protein's activity to the addition or removal of a phosphate group. In practice, both regulation based on gene expression and activity modifications can act upon each other (*e.g.* with transcription factor needing phosphorylation or genes expressing phosphorylation modifiers such as phosphatases).

Taken together, all these regulation mechanisms allow cells to actively sense and process internal and external stimuli and respond by modifying their protein content or proteins' activity. Importantly, regulation is inherently a dynamic event. A given internal or external information is acquired, processed and eventually leads to a reaction. Because different regulation mechanisms affect cellular processes by different means, their action have different kinetics. For example, whereas regulation undertaken by means of gene expression require minutes to hours to become effective (as producing functional proteins requires many biochemical steps), changing a phosphorylation state can be done in less than a second. Accordingly, cellular information processing is a dynamic process, the study of which requires specific experimental and analytical tools as it will be described later.

***Cellular economics, physical constrains and indirect regulation of cellular activity***

Cells are constantly out of thermodynamic equilibrium. This requires a constant exchange of energy and matter with the environment. Such lack of equilibrium is the driving force of many biochemical reactions occurring in cells. For instance, ATP to ADP conversion would not provide any energy if cells were at equilibrium (3). Because a cell is out of thermodynamic equilibrium, chemical kinetics can be affected not only by the concentrations of the species at play, but also by their immediate molecular context[2]. Yet, although cells are considered as *open* systems in thermodynamic terms, they are nevertheless bounded systems in many respects. Their volumes and mass are finite so is their maximum exchanges rates with their environments. Among all reactants, some molecules appear to be ubiquitous as they are involved in a very large number of reactions. Although usually present in important amounts, their supply is still finite. Accordingly, reactions are effectively competing for such molecules which we can compare to cellular *currencies*.

The most famous currency molecule is probably ATP which acts as a carrier of chemical energy. Because the supply of energy (power) is finite, along with the available pool of energy carrier, a cell needs to balance energy production and consumption dynamically. This imposes particular constrains on cellular activity leading to indirect regulations of cellular activity. For instance, if a cell initiates a highly ATP intensive process, it can affect the rate at which other ATP consuming reactions happen. There are other important chemical *currencies* in cells: redox potential is carried by (NAD/NADH), phosphate, methyl or acetyl groups are also available in limited supply.

[2] This is particularly true for redox reactions which can have spatially dependent reaction propensity even inside the same cellular compartment, unlike acid base reactions where water enforces a constant pH within an organelle.

This effect of cellular economics goes beyond metabolic reactions as it applies to any resource which is shared by molecules and processes. The pool of available ribosomes is a resource shared by many different mRNA, the same goes with RNA polymerase which is shared by genes and transcription factors. Lipid membranes are also a limited supply of real estate for the many membrane proteins. In all these cases, we see the same pattern where elementary processes rely on a pool large enough so it can be considered *constant* for any individual step. Yet the sum of such elementary processes impacting this pool will have important consequences.

The effects of cellular economics have been increasingly recognized and characterized (5–7). Some studies have proposed solutions to mitigate such effects (8, 9) in artificial genetic construct when others considered the implications of cellular economics on natural phenomena such as the shift from efficient to inefficient glycolysis with increasing growth rates (10). Many pools of shared resources are importantly affected by cell divisions and cell growth. Accordingly, the emerging picture of resource allocation in cells reveals also the indirect influence of cellular growth on gene expression (11, 12). In consequence, it appears that GRN both affect and are affected by physical and economical constrains. In a given cellular context, those constrains can be the basis for indirect regulation among processes.

As it was described, active coordination mechanisms can be represented as complex networks of molecular regulatory processes and are dynamic in essence. An inclusive representation of cellular regulation and information processing would require the cellular context to be also taken into account. Defining some form of ***augmented* GRN** would require all active regulatory components, interactions and interconnected layers of regulation to be embedded in a defined cellular context.

In this general introduction we presented from a general perspective some aspects of cellular activity which are of particular interest for the research which will be presented in this thesis. In particular, we depicted partly the raw complexity of cellular activity at the molecular level and highlighted the role of gene expression in coordinating many processes. Gene expression regulation is achieved through active and dynamic information processing molecular mechanisms involving many components and centered on gene regulatory networks. At last, resource allocation and cellular economics impose additional levels of regulation such as passive interplays between simultaneous processes and in particular with cellular proliferation.

# I. Introduction

## 1. Dealing with variability and time scale in gene expression

To a physicist, a cell appears as a complex and dynamical object. As we have seen, a cell has the capacity to perform hundreds of biochemical reactions. At the same time, a cell processes information about its internal state and its environment so as to coordinate and modulate its activity. Although cells divide into genetically identical cells, there are measurable differences between cells overall. Cell-to-cell variability can be easily observed in terms of cellular physiology (cells having different shapes or size for instance) but can also affect information processing and gene expression. In the following section, we present an overview of the known characteristics of variability regarding gene expression.

### a. Twins are not identical

***Stochastic gene expression***

It is known that a certain number of cellular processes are subjected to some inherent randomness which is related to the dynamics of biochemical reactions. In particular, gene expression is *noisy* because in many cases, due to the small number of reactants (a few transcription factors per cell and a single promoter), the typical homogeneous and well stirred assumptions required to derive classical chemical dynamics cannot hold anymore. In fact, when reactants are in such small concentration, a transcription event becomes a fairly unlikely event and the time distribution between successive events is more properly described by a stochastic process than a traditional chemical kinetic differential equation.

Experimental evidence of gene expression stochasticity has been given in *E. coli* and *S. cerevisiae* by using dual reporter constructs (13, 14): two identical promoters driving the expression of fluorescent proteins which are extremely similar in their DNA and amino-acid sequences (so as to avoid systematic bias in the expression of one of the proteins) and are located in similar loci (in order to ensure a similar genetic context[3]). Yet the two proteins of a dual reporter system have a different fluorescent spectrum. We report here results from Raser and O'Shea on dual reporters in yeast: in Figure 3 A. we can see cells showing different mixtures of red (color coded for YFP) and green (color coded for CFP). This allows single-cell quantification of the expression of each part of the dual reporter as reported on the plot of Figure 3 B where each cell is represented by a dot. The color of the dot represents cells from different time points from the start of the induction of the PHO5 promoter. If gene expression was deterministic, the two reporters would have the same expression level and all dots should fall along the *YFP=CFP* line (up to measurement errors). The fact that many cells indeed are away from the bisector means they are more green (in reality cyan) or red (in reality yellow) than what would be expected if both fluorescent proteins had the same concentration. In fact, the variability in protein concentrations can be decomposed in two contributions:

[3] Because of epigenetic effects or transcription machinery effects like RNA Pol II recycling, different repartition of the two reporters along DNA can yield in theory different level of coupling or covariance between the two reporters' expression. For the example in *S. cerevisiae* used here from (14), diploid strains were constructed with two homologous chromosomes bearing each a reporter inserted at the same locus (LEU).

- The **intrinsic** component represents the deviation from the situation YFP=CFP. Intrinsic noise originates from the fact that within the same cell and genetic context, the two promoters have not been transcribed equally due to stochastic transcription events.

- The **extrinsic** component represents the fact that some cells are globally brighter than others, regardless of the proportion of yellow and cyan. This in turns originates from cell-to-cell differences in gene expression which are equally affecting both promoters within a same cell.

In Figure 3 C is represented the fact that both the global level of noise and their decomposition between intrinsic and extrinsic components change with the overall level of expression. Although hardly observable on the figure, the intrinsic noise decrease with the average expression level is consistent with the interpretation of stochastic transcription events. Interestingly, characteristics of gene expression stochasticity in *S. cerevisiae* differ among promoters. It led to the formulation where intrinsic noise may also arises from chromatin modifications at the promoter region (which renders the gene accessible or inaccessible in a stochastic fashion). This is now called the *bursty* type of intrinsic noise compared to the typical noise which seems to originate from the binding kinetics of a transcription factor which is called *Poisson* noise[4].

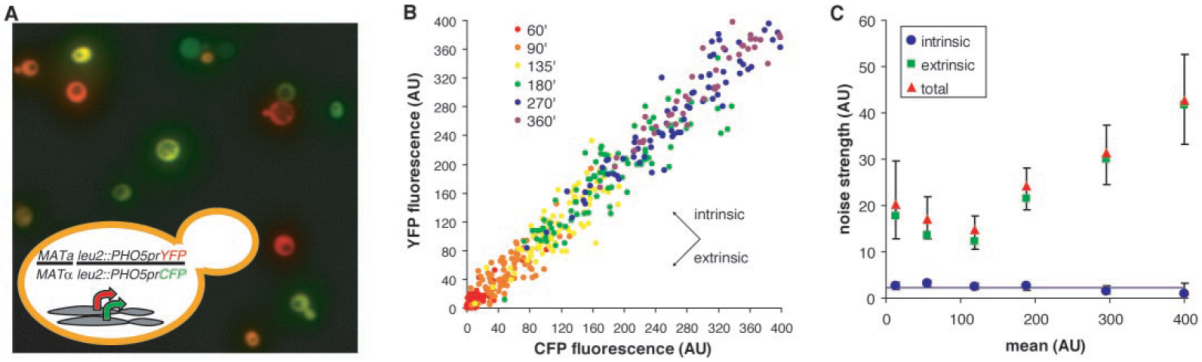



**Figure 3 – Decomposition of gene expression noise into intrinsic and extrinsic compoents. A. Microscopy image of the dual reporter fluorescence and sketch of the construct. B. CFP vs YFP fluorescence intensity measured in different cells at various time of induction. C. Decomposition of the overall variabiity into extrinsic and intrinsic componends at different levels of induction. Figure reproduced from** (14).

At the promoter level, it was mentioned that various kinetics could exists: Highly stochastic fluctuations happen in genes which are low expressed and for which there are few transcription factors leading to Poisson distribution; in both prokaryotes (15) and eukaryotes (16), expression can also come in stochastic bursts and display a large range of possible kinetics; at last, promoters of house-keeping genes and other essential genes are usually *constitutively*[5] expressed and display Poisson kinetics. Although stochastic gene expression is sometimes represented with simple distributions like Poisson, the reality of transcription at the molecular level depends on many factors (*e.g.* chromatin dynamics, transcription factor specific properties, promoter sequence affinity, transcription machinery recycling etc.). Many of such factors can also fluctuate both in a gene-

[4] This is because uncorrelated transcription event occurring at a constant probability follow a Poisson distribution.

[5] It should be noted that so-called constitutive promoters do not usually lead to constant levels of proteins as they tend to depend on growth rate or cell-cycle for instance (159).

specific and unspecific manner, possibly affecting stochastic properties of transcription. This makes it very difficult to cast all endogenous promoters in a single stochastic framework where a comparison would make sense. It also explains important differences between strains and organisms. In yeast, it seems that gene-specific characteristics dominate over genome-wide rules (17). This in turn means that global trend cannot be easily formulated for intrinsic noise, besides loose bounds as depicted in Figure 4 from Sanchez.

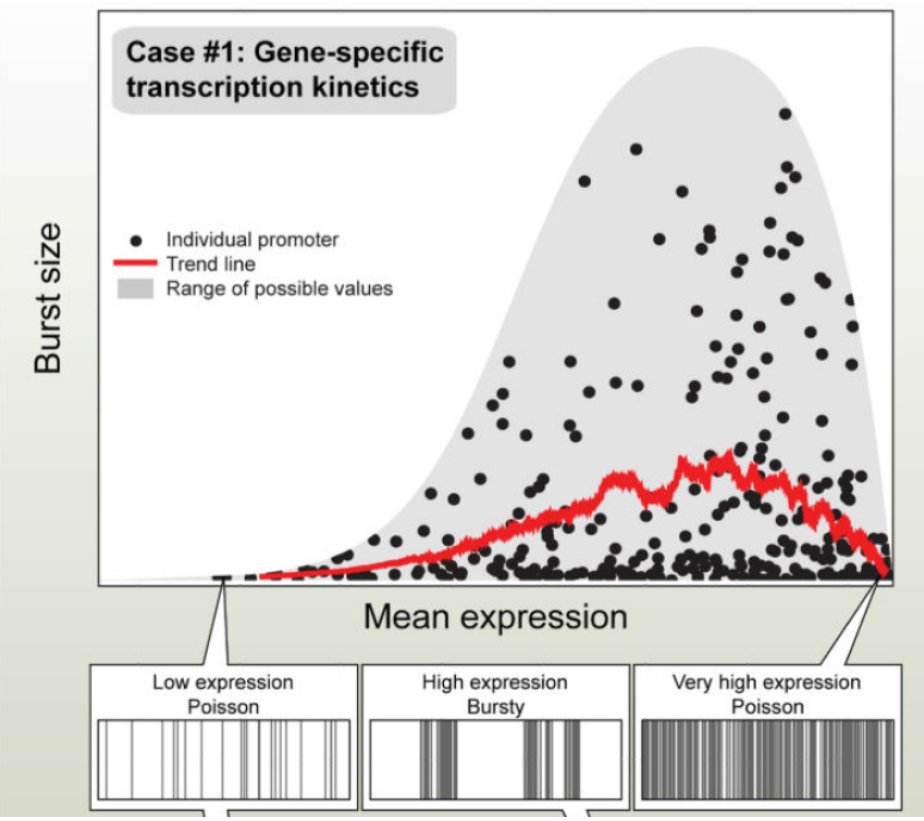



**Figure 4 - Simulation of possible promoter characteristics when noise is gene specific (as it is the case for *S. cerevisiae*). Burst size is the average quantity of mRNA produced during a transcription event and mean expression is the time-averaged level of expression. Figure from (17).**

Since founding experiments on gene expression variability using dual reporters, there has been a lot of subsequent studies which investigated the biological impact and relevance of noise for biological processes. An interesting review of nearly two decades of studies of gene expression noise can be found in (18). Among various situations described in this review are those where fluctuating variability leads to phenotypic *bet hedging* within a clonal population of unicellular organisms. In multicellular organism, stochasticity in gene expression can be used in mice olfactory system development to ensure the necessary development of many different cells using a *Monte Carlo* approach which proves way less expensive than a deterministic control of differentiation into thousands of cell types (18).

Going back to early experiments on gene expression stochasticity like that depicted in Figure 3, we may now consider the other, more traditional cell-to-cell variability which is the extrinsic component. As we see on Figure 3 C, cell-to-cell variability is quantitatively dominant over intrinsic noise. Although the relative levels of both contributions are very different depending on: the level of induction, the promoter observed, the genomic context or the type of organism considered; it is hard to find examples where the extrinsic variability is not higher[6] than intrinsic. Although the biological origins of intrinsic noise are usually known, when it comes to extrinsic noise, the picture is more blurry.

A simple explanation which is sometimes heard is that extrinsic variability is reminiscent of the intrinsic one as it comes from the time integrated intrinsic variability (*i.e.* it is small random event accumulated over time which leads to these pronounced levels of cell-to-cell variability). But such an explanation is at best incomplete and more usually wrong.

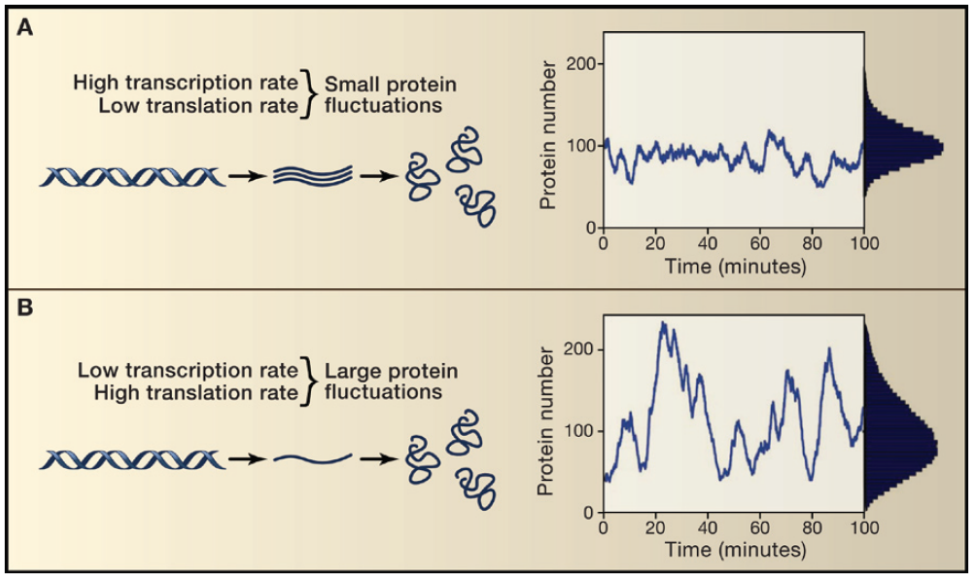



**Figure 5 - Relating single cell stochastic gene expression trajectories and stochastic equilibrium distributions. This figure reproduced from (18) aims at representing the impact of different stochastic gene expression regimes in terms of dynamics as well as equilibrium distributions. Equilibrium distributions should not be confused with the distributions of proteins in a population of cell. For these two different distributions to be equal a fundamental assumption is required: Gene expression needs to be a stationary process and every cell must harbor the exact same process (*i.e.* having the exact same rates of transcription, translation, degradation etc.).**

Some confusion may come from the fact that a stationary stochastic process leads to an equilibrium distribution (see Figure 5 from Raj and van Oudenaarden) which looks similar to histograms of total variability as provided by flow cytometry. From a theoretical point of view, for the equilibrium distribution to be identical to the total variability distribution, gene expression needs to be considered a stationary process and identical in different cells (*i.e.* ergodic). Yet, the cellular context in which gene expression happens differs from a cell to another. Experimentally, if the equilibrium distribution was identical to the total variability distribution, dual reporter experiments would clearly indicate so.

[6] In this respect *S. cerevisiae* is deemed particularly *extrinsic* (18).

The important correlation between dual reporters which is accounted in extrinsic variability both comes from common causes affecting both promoters and from the fact that fluorescent proteins are usually stable, which, as it will be shown in the next section, is fairly common.

As stated in (18), the origins of extrinsic noise are still poorly understood. This comes also from the fact that separating overall variability into intrinsic and extrinsic contributions helps defining the intrinsic component but puts in the other bag many different realities. Figure 6 from Huang contains a tentative breakdown of total variability in a population of cells into several classes. In that context, extrinsic and intrinsic refers to a cell and not to a promoter within a cell as used throughout this paragraph.

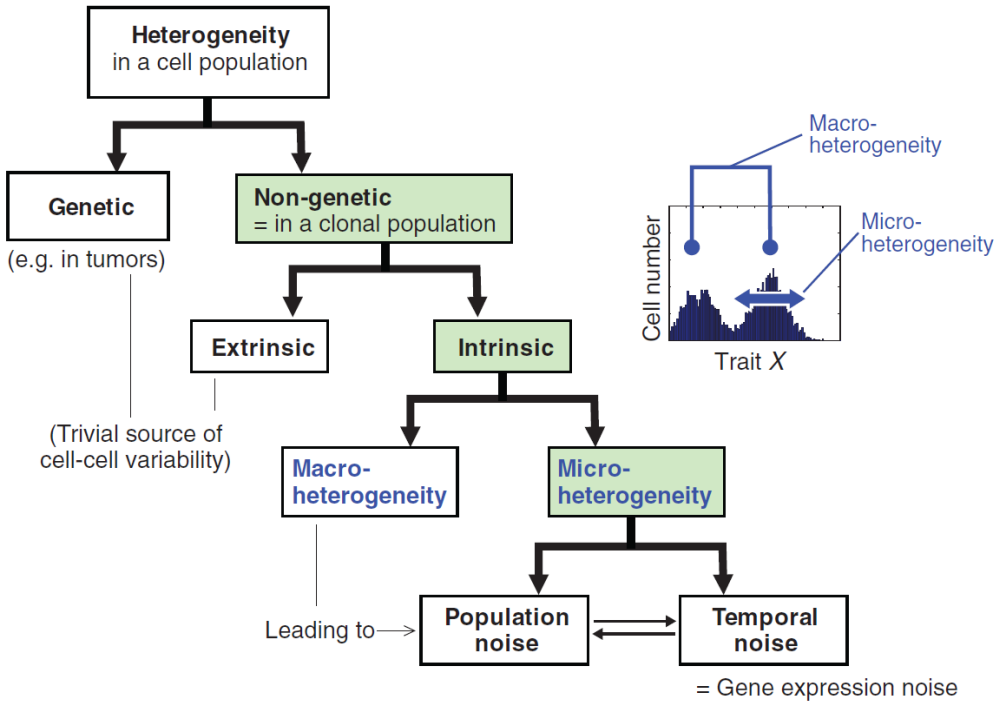



**Figure 6 - A tentative definition of different aspects of total phenotypic variability (or heterogeneity). Figure reproduced from (114). Note that the terminology (heterogeneity, intrinsic, extrinsic etc.) is different from that used in this paragraph.**

In the decomposition of Figure 6, we see that intrinsic gene expression noise is called temporal noise and extrinsic is called population. This underlines the vital aspect of temporality in the context of cellular variability. The following subsection aims at presenting how considerations of time and cellular context are needed to come around a more solid view of extrinsic variability.

### b. What a difference a day makes?

In the previous subsection, gene expression noise was introduced and the stochastic nature of gene expression was presented. Studying variability obviously requires obtaining single cell measurement. A popular and efficient method to do so is using flow cytometry. Nevertheless, single-cell measurements as obtained by flow cytometry lack crucial temporal information on variability dynamics.

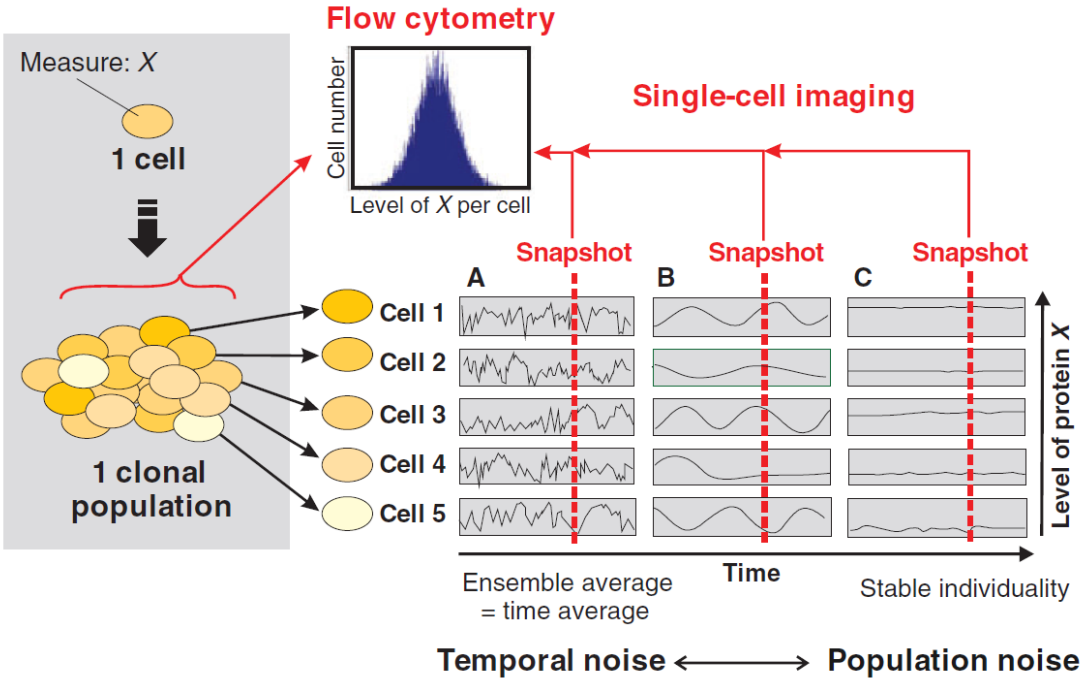



**Figure 7 - Sketch representing the impact of lacking temporal information to distinguish between different types of variability. Figure reproduced from** (114)

As we see in Figure 7 from Huang, flow-cytometry only provides a snapshot of a population and therefore cannot alone distinguish between fast-changing properties and nearly static ones. When it comes to its biological interpretation (*i.e.* looking for causes or consequences), the temporality of variability is essential. In the following we will try to estimate what time scales are effectively at play from gene transcription to proteins and GRN.

If we consider the simple sketch of Figure 7 and consider the overall time line depicted in A, B and C is one cell division, it can be expected from what was mentioned in the previous section that promoters can have activation patterns falling in all categories. As we saw, different transcription rates are possible and determine the kinetics of mRNA production. Yet, mRNA overall kinetics also depend importantly on degradation rates (19). Despite many technical issues in dissecting production and degradation from mRNA levels experimentally, some genome scale idea of mRNA half-life *in vivo* in *S. cerevisiae* is given in (20) and reproduced in Figure 8 below.

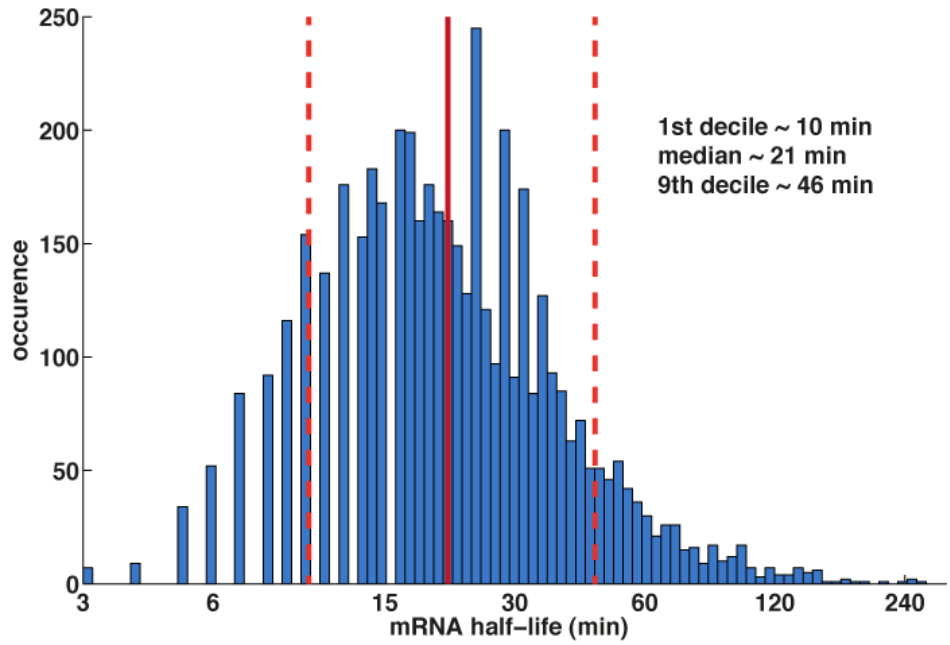



**Figure 8 – Genome wide distribution of mRNA half-life in *S. cerevisiae*. Data from** (20)

As we can observe, most mRNAs have a half-life in tens of minutes. This means that mRNA will have a time evolution that can either follow closely promoter activation for the most unstable transcripts and fall in category A of Figure 7 or lead to a fairly smoothed time profile regardless of their promoter dynamics which would be in category B or C. Therefore, when it comes to mRNA levels, many different temporal situations can arise. Depending on the total number of mRNA itself, both deterministic and stochastic representations may apply.

Looking now at possible variability in protein levels, we can see in Figure 9 that most proteins are present on average in fairly large numbers. This means that when it comes to proteins, we should be more confident about using deterministic chemical representations which arise from the law of large numbers.

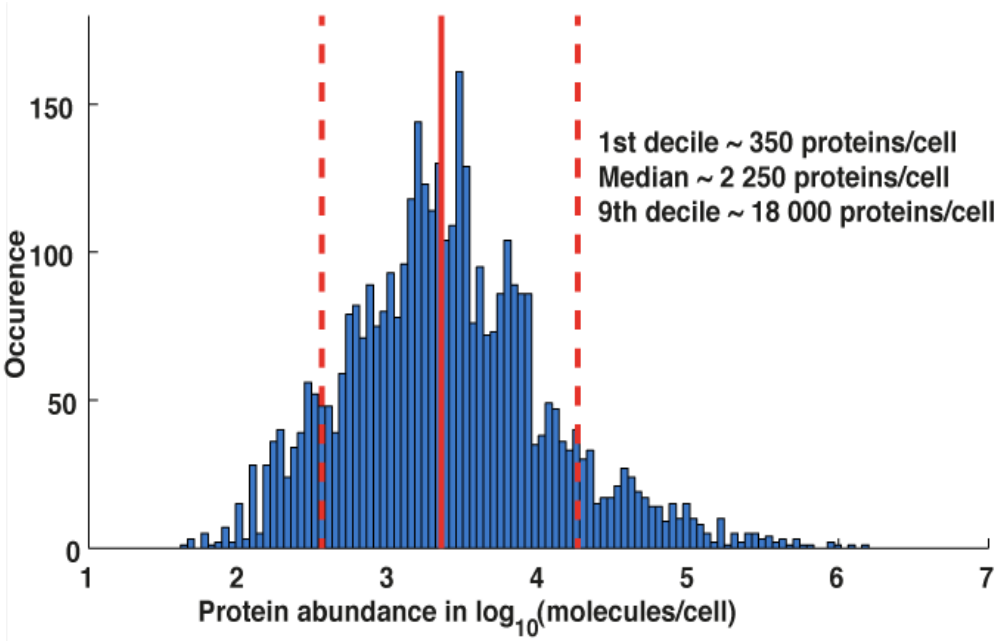



**Figure 9 – Genome-wide distribution of proteins abundance in *S. cerevisiae*. Data from BNID 101845**

It is trivial to state that as in the case for mRNA, protein abundance results from the balance between production of new proteins and decay through degradation and dilution. But considering proteins are present in large numbers allows an easier deterministic representation. The nature of decay processes, despite some protein specific regulations is well described by exponential first order kinetics which means that the rate of protein decay (in concentration units) can be represented by $-\alpha[P_i]$ where $\alpha$ is the total decay rate and $[P_i]$ is the cellular concentration for some protein $P_i$.

On the other hand, the production of a given protein can be either: independent of the protein concentration (as for constitutively expressed proteins or proteins whose expression depends on other factors than the given protein level itself, for instance the position in cell-cycle); or dependent of the protein concentration through direct (auto-regulation)[7] or indirect feedbacks in the GRN.

In the first case, proteins kinetics will be largely determined by the decay rate alone. In fact, regardless of the production rate, for an increase in expression the typical time to reach half the maximal value is $log(2)/\alpha$. For a decrease in expression, dynamics will have exactly the same typical time. In the second case, there is no direct answer since the accurate time scale depends on the nature and precise parameters of the feedback. Still, we can note than much faster time scales are possible with feedback which may both increase or considerably reduce variability (see section 3.4 in (21)).

Recent experiments (22) managed to perform a less harmful measurement of proteins degradation rates *in vivo* than previous large scales studies (23). Their results, depicted in Figure 10

[7] Auto-regulation is a famous motif in which a direct feedback exists between a gene and its transcription. Yet, out of the ~180 gene encoding transcription factors existing in *S. cerevisiae* (160) we can estimate than 10% include some auto-regulation in their interactions, in sharp contrast with *E. coli* where this proportion is estimated to range between 52% and 74% (161). Overall, autoregulation is therefore fairly rare in budding yeast.

confirm and somehow reinforce precedent reports that most of *S. cerevisiae* proteins are highly stable.

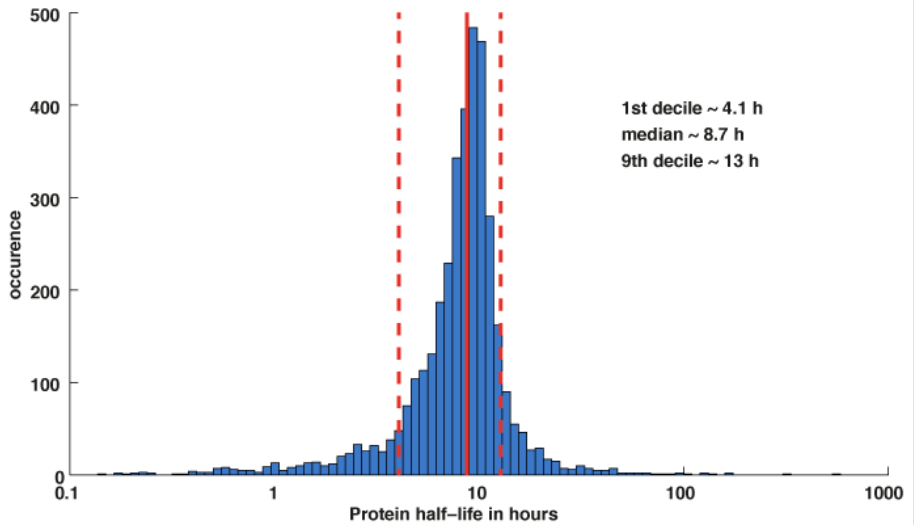



**Figure 10 – Measured protein half-life (in absence of dilution) for approximately half the genome of S. cerevisiae. Data from** (22)

The order of magnitude of protein stability is arguably the main reason why in most cases, fluctuations in protein levels are mainly driven by dilution. Most proteins are therefore expected to fall into category B of Figure 7 when upstream components are noisy and in category C for proteins with stable upstream kinetics (which includes many essential genes (24)). These results on protein degradation are in line with an economical consideration of variability and protein synthesis. Protein production is an expensive process and therefore, whenever possible, protein stability ensures that synthesis expenses are amortized. Most protein decay being driven by dilution, we see that a significant portion of proteins are inherited and transmitted along a cell lineage. In that respect, the initial pool of proteins which a cell receives at its birth constitutes somehow a working capital that will take roughly one cycle to be doubled. Nevertheless, some particular proteins will have very fast dynamics which requires more production and degradation and comes at a significant cost. From this perspective, we see that rapid turnover required for fast fluctuation in protein levels is expensive and must therefore bring some advantage which would explain their selection and maintenance.

From this overview, we can recall that variability comes in many flavors. Cells have mechanisms which can both increase or decrease its level. Furthermore variability cannot be reduced to a source of instability nor to a *one fits all* solution to generate phenotypic differences for evolution to select from. Variability in its global meaning is a very broad subject of research. Because of all its aspects, precise and quantitative analysis is useful to progressively disentangle contributing factors and assess their importance and synergies. Studying variability requires specific experimental tools which allow measurements at the single cell level. In the next section, we will present a short review of available technologies for single-cell data collection. This will allow us to draw the perimeter of

currently accessible single-cell information as well as to point out expected developments in the near future. After this summary, we will discuss the global research methodology which was employed in this study along with the precise biological system and research objectives we employed.

## 2. Measurements at the single cell level

Many challenges in quantitative biology require single-cell data, in particular to account for cell-to-cell variability. From the early days of molecular biology to the present times, technological improvement has been enormous and many tools are now able to reveal biology at the single-cell level and even below with single molecule measurements.

***Measuring mRNA***

Single-cell transcriptomics is now becoming a reality. It is now possible to get accurate readouts of mRNA abundance from single cells at the genome scale with larger and larger coverage, which is often called scRNA-seq[8]. Different technologies (next generation sequencing or pre amplification with subsequent use of microarrays) have been used at the single-cell level already (25). The first studies using these techniques used it on mammalian cells (even at a subcellular level in the case of neurons (25)) but was not applied to budding yeast yet. Interestingly, obtaining genome-wide transcriptional information at the single cell level triggers the integration of important analysis paradigms coming from research in cellular heterogeneity to traditional *omics* frameworks (26, 27). This in turns may enlarge the audience interested in the study of variability.

FISH (Fluorescent In Situ Hybridization) is a technique which allows to couple fluorescent dyes to single-stranded DNA that in turn can bind to a specific mRNA. This method, coupled with super resolution[9] fluorescent microscopy, can resolve single mRNAs in budding yeast (28). Yet, this method requires cells to be fixed and washed and can therefore at best provide snapshots of cell-to-cell variability at different points in time by using different experiments (29). In addition, although barcoding[10] might help a bit, the number of mRNA that can simultaneously be quantified is limited by the number of distinct fluorescent dyes (in practice 3 at a time without barcoding). Although scRNA-seq could be expected to make FISH obsolete when it comes to estimating mRNA abundance, FISH will mostly still be very useful when it comes to acquiring spatial information and in particular at the sub-cellular level.

***Measuring proteins and metabolites***

For a long time, development of antibody tags (antibodies coupled to either staining or fluorescent molecules) allowed to visualize in fixed samples the repartition of specific molecules (mostly proteins) and to quantify their abundance. These labelling methods allow *in situ* measurements which are by essence single cell (and even sub-cellular). These techniques suffer from the same issue of the limited number of elements which can be simultaneously observed because of the limits in resolving tag-specific spectra within the visible/near visible light spectrum[11].

[8] Single-cell RNA sequencing.

[9] Super resolution microscopy can resolve details below the diffraction limit of traditional microscopy by using advanced image reconstruction algorithms and 3D microscopy data as obtained by performing a Z-stack.

[10] The idea of barcoding in this context is to attach to probes several fluorophores with precise ratios. Therefore, it is a particular mixture of colors rather than one color which allows the identification. This in turns require single molecule resolution imaging techniques.

[11] It is to be noted that given typical protein abundances, single molecule cannot be resolved and therefore, the barcoding extension which may boost a bit FISH cannot be applied. But other extensions are possible for instance sequentially bleaching and staining or washing samples.

Traditional *destructive* techniques to identify molecule and quantify abundances in relatively large samples (containing many cells) include various types of mass spectrometry (MS) used in combination with other separation methods based on electrophoresis gels or chromatography techniques like HPLC[12](30). These techniques are usually *label free* since they do not rely on addition of specific tags. Without going into much detail, mass spectrometry based methods allow the identification of proteins according to the *m/z* ratio[13] of various peptides composing these proteins (31). Yet, when it comes to precise quantification and increasing sensitivity, the introduction of some form of labelling is useful with methods like SILAC[14] which relies on feeding cells with isotopically marked amino acids to measure complete proteomes in a single experiment (32).

Using MS for *in situ* measurement (and therefore with single-cell or sub-cellular resolution) is made possible by vaporizing a minute amount of (fixed) sample with a laser within a microfluidic system which collects the corresponding vapor and convey it to the detector. This approach has been applied with label free MS techniques (33) but could only resolve properly 35 metabolites which were abundant enough and had well defined spectra. Alternatively, single cells can be isolated prior to MS analysis. This was done in *S. cerevisiae* metabolic studies for instance in (34). A promising research direction aims at increasing sensitivity along with quantification using metal labels. Such methods somehow bring the antibody tag technique into the realm of MS. Instead of coupling antibodies to fluorophores or dyes, it is possible to couple them to molecules (usually metals) that possess a distinctive MS signature[15]. Limitations from the visible spectrum being therefore amended, these techniques allow already the simultaneous measurement of tens of labels *in situ* (35, 36) and are expected soon to quantify more than 100 labels simultaneously.

***Dynamic measurements of variability***

Flow cytometry is a widespread technique allowing fluorescent measurements at the single-cell level for large populations of cells. Although flow cytometry is not destructive, single cell history is lost after one measurement as all cells are replaced in a same vial. This method provides several snapshots of a population, with single-cell resolution but without any information on single-cell dynamics. In this respect, it is very similar to destructive methods performed on parallel experiments sampled at different time points. Because variability exists both between cells and in time, using flow cytometry data only to study variability requires specific assumptions on variability dynamics to interpret the data. In other words, data alone cannot identify temporal characteristics of variability and therefore the plausible sources of variability. FACS[16]allows the sorting of single cells in several different flasks based on their level of fluorescence. This provides some form of temporal information as subpopulations can be subsequently measured again. Because traditional systems based on flow cytometry require manual liquid handling, it can limit temporal resolution (both in terms of sampling and duration). Yet, novel automated platforms allow to improve significantly this aspect along with making it possible to run several experiments in parallel (37).

[12] High Pressure Liquid Chromatography, now called also High Precision Liquid Chromatography.
[13] In MS, molecules are ionized prior to analysis and therefore acquire a charge *z* along with their molecular mass *m*.
[14] Stable isotope labeling by amino acids in cell culture
[15] These methods include DOTA (162) or MeCAT (163).
[16] Fluorescent-activated cell sorting

Because dynamic aspects are of crucial importance in understanding the functioning of gene regulatory networks, direct fluorescence microscopy is essentially the only option to obtain single-cell time-lapse quantitative data. Gaining quantitative information on cellular processes usually will require adding molecular probes in the cells. In this aspect, fluorescence tags are the most common tools. Illumination for fluorescence is higher than in bright-field, but can usually be maintained low enough as not to cause too much phototoxicity. A more detailed discussion of fluorescence imaging issues is given in subsection II.1.b. The obvious limitation of microscopy methods is that these are pretty low throughput. This can be partially compensated by using microfluidics systems, strain libraries and automation which allowed for example to record single-cell dynamics and cellular localization for ~3000 different proteins in different conditions (38) and stresses (39) for screening purposes.

Because the dynamic resolution of fluorescence labeled protein has several limitations (mainly fluorophores maturation times and degradation rates), other systems have been proposed which rely on bioluminescence to measure gene expression. The main issue being much weaker signals in bioluminescence compared to fluorescence. Yet, as both CDD sensors and bioluminescent systems are improved, these tools might become more widespread as an alternative to fluorescence. For instance, the most famous bioluminescent system which uses Luciferase was recently optimized for yeast (40) showing more rapid dynamics than fluorescent proteins upon gene expression. Nevertheless, only one label can be used at a time so far.

Several techniques such as Spinach or MS2 allow *in vivo* dynamic measurements of mRNAs. These require the inclusion of specific sequences in the mRNA to be observed which will produce specific secondary structures. These structures in turn will allow dyes provided in the medium or expressed fluorophores to bind to the mRNA. Although promising, these techniques which have been used in *S. cerevisiae* (41–44) are still challenging in practice when it comes to precise and dynamic mRNA quantification. Because mRNAs are small, precise counts require super-resolution type of imaging (*i.e.* acquiring many frames per time point) and since signal-to-noise ratio from single mRNAs is low, imaging these systems require high illumination which can have phototoxic effects. This can limit in practice the use of these techniques for long term experiments. At last, these mRNA tagging systems can affect some cellular processes (45). Yet, some shortcomings of early techniques have been overcame by other systems such as IMAGEtag (46) which use FRET[17] based fluorescence in order to improve signal to noise ratio. Therefore, although *in vivo*, time-lapse mRNA measurements at the single cell resolution is not yet a very mature experimental technique, it can be expected to become more common in the next decade and will be extremely valuable for the study of gene expression (47).

[17] Förster Resonance Energy Transfer (or Fluorescence Energy Transfer) is the mechanism by which the emission of a first fluorophore can excite another fluorophore if both are spatially very close to each other. In this context, this technique allows only fluorescent tags which are bound to mRNA to be visible which improves greatly the signal-to-noise ratio.

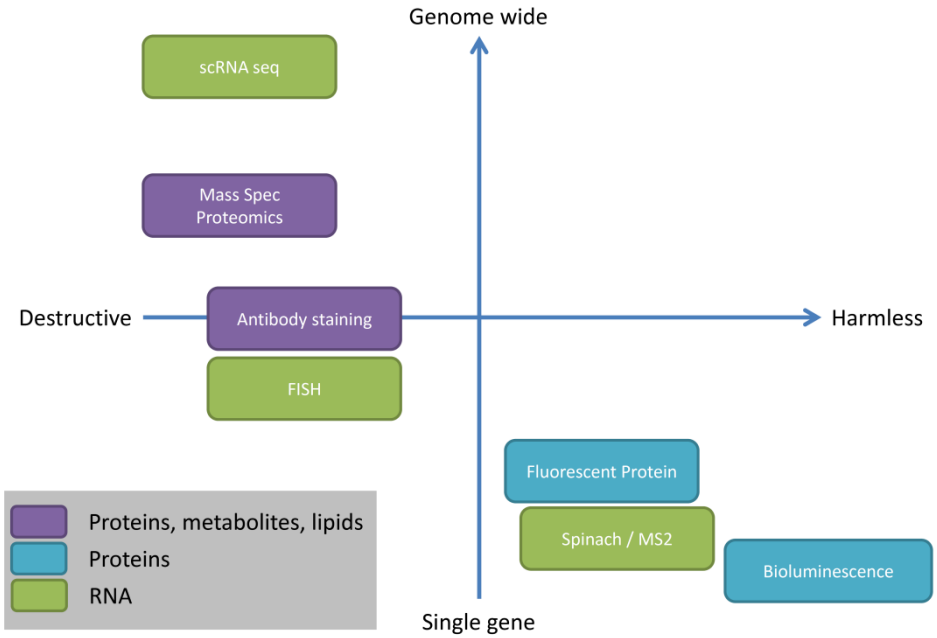



**Figure 11 - Synthetic comparison of several single cell experimental techniques. Colors encode the measured elements.**

***Time-lapse microscopy for measuring single-cell dynamics***

Our brief review of current experimental techniques allowing single-cell measurements shows the typical trade-off (represented in Figure 11) between having a wide coverage (genome wide methods) and being harmless enough to be conducted in time *in vivo*. To put it simply, one has to choose between poor dynamic information with a broad measurement scope (*e.g.* whole proteomes or transcriptomes) and precise dynamical insight on a very limited subset of elements. Although all methods presented here provide single-cell data, the use of any specific technique will condition importantly the type of questions which can be investigated.

In this study, we are interested in cell-to-cell variability and various dynamical aspects at the single cell-level. This requires acquiring longitudinal single-cell data so as to study cellular dynamics (*i.e.* measuring the same cell in time). Time-lapse fluorescence microscopy is therefore a natural choice and has the direct advantages of a mature technology (cost, expertise, optimization etc.). Yet, it goes with its inherent limitations: very few different proteins can be visualized simultaneously.

Note that imaging mRNA in addition to proteins would have surely been very informative. Nevertheless it is still experimentally very challenging and can hardly be done for long term experiments as it is necessary here. A study conducted in collaboration with my lab aims at constructing such mRNA + Protein fluorescent reporters in *S. cerevisiae*.

## 3. A synthetic and systems biology approach

### a. Cells as systems

Proposed in the 60' and more precisely defined in the 70', the concept of *system* is now omnipresent in science. Although applicable to nearly everything, this concept was historically structured for a large part in relation to biology (48). While the knowledge of detailed biological mechanisms expanded, it became clear that a purely reductionist approach might be insufficient to understand many features of Biology. As more constituent and basic processes of living entities were identified the focus shifted increasingly towards understanding properties of their mutual interactions (49). This paved the way to a new discipline: Systems Biology, which defines itself in opposition with reductionism and seeks a more holistic understanding of Biology, mainly through mathematical modeling and a system's theory approach (49).

As we know, biology features such complexity that engaging its study globally and precisely in a frontal manner is hopeless. What system biology advocates is using *abstraction* rather than *isolation* to render the analysis of complex phenomena tractable yet meaningful. Finding the proper level of abstraction is in fact much more difficult than simply isolating a few components and interactions, but when successful, it provides a much more general understanding because not only it can also explain or predict phenomena, but it also answers the question: "*What matters in this process?"*. Knowing what matters is vital as it helps foreseeing in which contexts the studied system will behave differently and why. Making a parallel with physics, we can observe that thanks to known abstractions, we know for instance that to determine the trajectory of an object, the most important things are its mass, initial velocity and the external forces. Its temperature or its color is completely irrelevant to this problem but as we were taught directly the proper abstraction we often forget what it took to find out among all things what is the relevant information here. Systematic thinking can also work together with reductionism, introducing a Russian nesting doll description of a phenomenon. Continuing our comparison with physics, we know that if we also need to predict how an object will spin, we do not need the full description of its mass repartition but only its inertia matrix which is a compact description of its mass repartition symmetries. So not only in this physical example we already know what the proper abstraction level is, but we also know how to refine or simplify it according to the precise question at hand.

Answering to the question "*what matters?"* is still usually very difficult in biology and this is also why quantification is so important: it is a fundamental tool in separating main drivers from exotic refinements. When confronted with more than a handful of numbers, we usually cannot distinguish any pattern anymore. This is why systems biology relies most of the time on mathematical representations. Using mathematical models of biological processes serves multiple purposes. In my opinion, a first advantage is the inherent precision in the system description it requires. Luckily, this does not mean that everything needs to be completely understood, but proposing a mathematical description of a problem clearly indicates what is known and what is more fuzzy. It also forces assumptions to be more explicitly stated than what is usually done with words.

Another advantage is that mathematical models benefit directly from already powerful abstractions[18] which may help finding what are the equivalents of *mass*, *Reynolds number*, or *elasticity* in Biology.

***Typical challenges in systems biology***

Having the objective of finding "*what matters?"* in a biological system, we can enumerate a given set of properties which are expected to play a role in an abstract representation of biological processes: the map or structure of possible interactions (along with their nature) between elementary components; the dynamics of elements and the dynamics of interactions which allow propagations; the robustness to internal or external perturbation. The study of such properties in the case of biological systems leads to several fundamental questions:

Concerning the **structure of interactions**, how can we infer maps of interactions which are heavily interconnected and include redundant components? Linear chains of causality are an exception and usually, many interlaced cycles of cause and action are at play in biological phenomena. In particular, many elements have several functions by themselves or may passively affect other functions. What properties of such networks are necessary and sufficient to capture a given type of behavior?

Concerning the **dynamics of biological processes**, how can we distinguish between genuinely constant features and deceivingly stable ones which are maintained by homeostatic processes? Also, to what extent processes having different typical time scales but common elements are independent?

Concerning **robustness**, which are the relevant perturbations for a given biological process (transitory or constant? Structural or functional? Unitary or multiple?). Also, which aspects of its unperturbed behavior a biological system should be robust in? Should we treat differently perturbations which are biologically plausible (*i.e.* which could have been present in nature thus imposing a selective pressure) from those which are completely artificial?

As an illustration of how such properties can help characterize a system, and more precisely "*what matters?*" in it, we report a study from Muzzey *et al* (50) in Figure 12. Authors investigated the capacity of *S. cerevisiae* osmotic stress response to "perfectly adapt" (*i.e.* to deactivate adaptation mechanisms as soon as the cell is adapted). Perfect adaptation is a type of dynamic robustness property which requires part of a system to act as an integrator. Abstracting the biological adaptation process into four sub-processes and characterizing several system level properties allowed the enumeration of all possible combinations of integrators number and position and the subsequent identification of the only coherent possibility.

[18] On a more personal prospective note, I believe that new mathematical frameworks designed for biology still have to be invented. An old attempt at forging such a framework can be found in the work of Gilbert Chauvet.

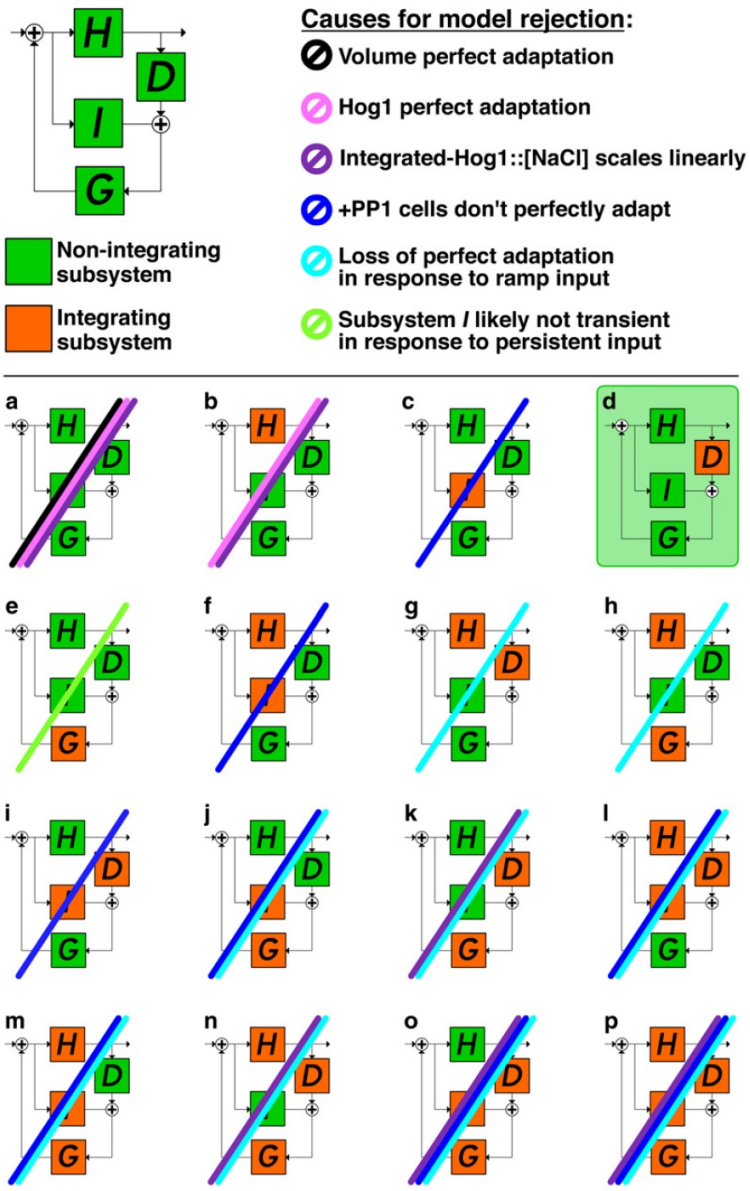



**Figure 12 - Example of a system level analysis of a biological process. In** (50), **from which this figure has been reproduced, the authors aimed at identifying a structural characteristic of budding yeast's adaptation to osmotic stress (the number and position of integrating subsystems). Using specific experiments that revealed several system level properties (related to robustness, dynamics and structure), it was possible to identify the only coherent topology (d) among all candidates (a-p).**

### *Drawing frontiers around cells*

Because most of the time we still lack the answer to the question "*what matters?*", a study of the aforementioned properties of biological systems is inherently difficult to design. Indeed, the results of any given investigation will depend on the proper definition of the system and of the scope of analysis. Although the cell seems an obvious system to study, its proper systematic definition is not trivial.

The first question is whether we speak of ***a cell*** as a real biological entity or if we consider ***the cell*** as a general object of which particular cells are instances. The distinction may seem merely philosophical but has important consequences in practice. Indeed, how should we consider and describe the various types of variability (stable or time changing cell-to-cell variability, extrinsic or intrinsic)? How long can we consider a cell still to be the same given that accumulated random changes within it or external alterations of its micro-environment can make it different? To what extent two cells are comparable?

What are the consequences of variability for the representation of a population of cells? Under which circumstances a population can be simply abstracted as a collection of identical object? Conversely, how can we take into account cellular variability when studying population of cells? Is there an abstract representation capturing "*what matters?*" without fully accounting for all composing elements?

In this work, we investigated dynamic processes in single cells with a focus on the representation of cellular variability. For any cellular process, the cell as a whole defines a context which can influence the process' behavior and outcome. Variability between cells therefore entails a variability of context. In our study, we wonder what influences the cellular context and how we can represent the ensemble of context present in a population of cells.

### b. Experimenting within a cell: Synthetic biology and microfluidics

Parallel to Systems Biology, Synthetic Biology takes an engineering approach where the focus is more on acting upon biological systems (mostly at the genetic level) than on observing their natural behavior. Somehow, synthetic biology looks for conditions and tools that would escape the "everything impacts on everything" aspect of biology. It aims at creating "orthogonal" systems which should enjoy the homeostatic conditions of cells while pursuing their artificial functions in an uncorrelated manner with the rest of their biological environment. This approach has proven fecund, although sometimes its limitations where overlooked (51). In any case, it both managed to accomplish this orthogonality until some point and to provide very useful tools to act on biological systems. Because synthetic biology is interested in *designing* and *controlling* biological systems, it also helps answering questions central to systems biology (49). Conversely, systems biology provides insights which help designing more efficiently new biological constructs (52).

In order to improve our understanding of cellular information processing, alterations of the structure and interventions on the behavior and dynamics of GRN are unparalleled sources of information. This can be done in several ways: We can act directly upon systems at the genetic level. This in turns proves to be a very versatile way of modifying a system's structure in a designed manner. A basic genetic operation is gene deletion which simply removes a gene from the cellular repertoire. Analysis of the consequences of gene deletions is the primary source of annotation of gene's functions. Gene deletions are useful, but they are often either without any noticeable consequence or lethal. Lethality as such is informative but will often not help in understanding the precise mechanism and role of an essential gene. Instead of deletions, modifications can also be performed which will partially alters the molecular function of a protein. The most ubiquitous modification is fluorescent tagging which allows a direct readout of gene expression. Besides the gene itself, it is possible to modify the regulatory sequence upstream a gene to control its expression

externally (or to put it under a different endogenous control mechanism) Such external control of expression is known as an *inducible system* or *inducible promoter* and usually requires integrating other genes (expressing exogenous transcription factors for instance).

Inducible systems are a cornerstone as they open the door for conditional expression (the choice of when a gene is expressed). Moreover, many inducible systems have also a range of gradual expression which makes it is possible to tune the level of gene expression and therefore resolve quantitative functions in gene expression. At last, within certain bounds, it is possible with inducible systems to exert precise dynamic perturbation at the molecular level *in vivo*.

Given the complexity and dynamic nature of GRN, each level of intervention (deletion, conditional expression, gradual expression, dynamic expression) allows researchers to access a finer and finer level of information. Deletion and conditional expression can already provide a great amount of knowledge about gene's functions and some interactions. Gradual expression provides the first true quantitative data which allows building quantitative model. Yet, static gradual control only provides information about equilibrium in GRN. At last, dynamic expression, would it be perfect, allows in principle to truly *hack into* GRN and recover any sort of information concerning its present working. In fact, as genes naturally interact through dynamic expression, being able to control it give researchers a way to modify all relevant aspects (its level, frequency, fluctuation profile etc.). In the eyes of a control theoretician, having dynamical input (together with outputs) is the basis for a methodical reverse-engineering of information processing occurring in GRN. For example in (53)

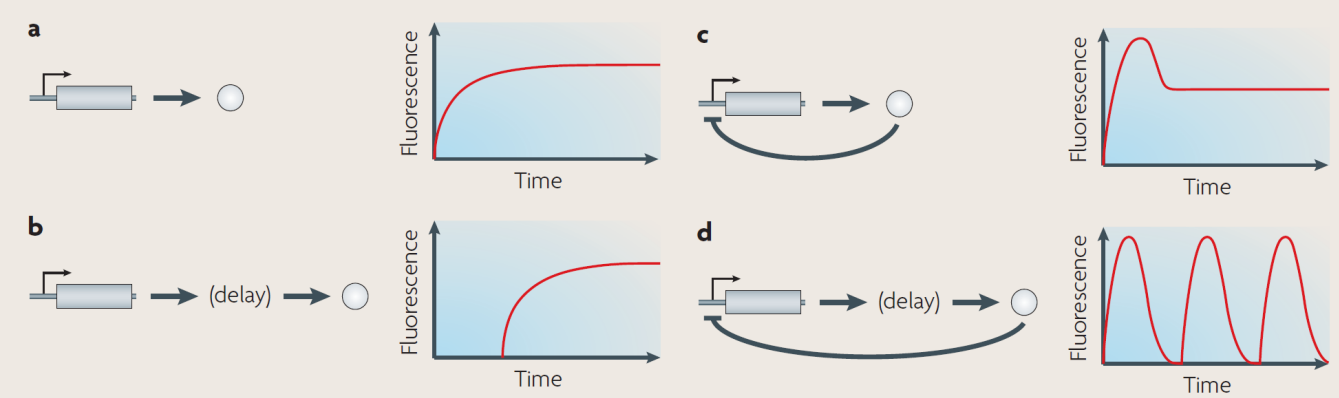



**Figure 13 - Schematic representation of the use of inducible systems along with temporal information to identify motifs of genetic regulation. Adapted from (53)**

Bennett and Hasty review how dynamics of gene expression can be captured using time lapse measurement and inducible systems. In Figure 13 we reproduce a sketch showing how various regulatory structures will lead to distinct dynamics upon induction. More advanced examples of the use of dynamic measurements and stimulations for systems biology are provided in (54–57).

**Reverse engineering of cellular information processing by GRN is one of the broad and long-term objectives of my research team.** As such, it really is a fusion of systems biology and synthetic biology for a more systematic understanding of cellular information processing. This thesis participates in this long term endeavor both technically and theoretically.

From a technical perspective, the tools that are needed for such reverse-engineering are only partially available and much further development is needed for investigating complex biological

questions. We already mentioned that capturing single-cell dynamics requires time-lapse single-cell data as provided by fluorescence microscopy. In addition, we consider here variability in dynamical processes. This means that we compare cells regarding *features* of their dynamics. In order to have informative single-cell time course data from which we can assess such *features*, we employed time varying stimulations. This was done experimentally using custom microfluidics and custom hardware which are presented in chapter II among other crucial developments in image analysis. In our final perspectives, we will also mention some additional technical development which was carried on but not used yet during this project.

From a theoretical point of view, existing methodologies used in systems biology usually originate from traditional reverse-engineering and were designed for specific types of systems. Although applicable to some extent to biological systems as well, living matter displays some specific features for which pre-existing frameworks were not designed. As it was discussed, focusing on the single-cell level and on variability affecting the cellular context leads to many questions about our representation of biological systems as simple as cells and isogenic populations of cells. To investigate this matter, we used a well-documented biological system on which the lab had already some expertise: the response of *S. cerevisiae* to osmotic stress. In the rest of the introduction, we will describe this system and refine the broad questions we raised so far into more specific ones which will be addressed in chapters III and IV.

## 4. *S. cerevisiae* response to osmotic stress

Having exposed the global research objectives driving our study, we will now present the biological system on which we performed our experimentations: the cellular response to osmotic stress. First we will present what osmotic stress is and give an overview of how yeast cells adapt to it. Then, we will motivate the use of such system and detail some of its characteristics which are relevant to our study. At last, we will present a short literature review of various analytic approaches and mathematical models of this system.

### a. An overview of the HOG response

***The physics of osmotic balance in cells***

Osmolarity is related to the properties of water as a solvent for dipolar molecules or ions when various compartments are separated by semi-permeable membranes[19] (such as most of cellular membranes). Although incorrect (58), a simple explanation is that water tends to dissolve all the available molecules evenly across semi-permeable membranes. Osmolarity quantifies the propensity of a given solution to be further dissolved by water. Therefore, the osmolarity of a solution is related to the concentration of soluble chemical species. Let's consider a system composed of two compartments separated by a semi-permeable membrane. If one of these compartments has initially a higher osmolarity, water will flow from the lower osmolarity compartment to the higher osmolarity one until solutes in both compartments are dissolved more evenly. This phenomenon is known as osmosis. A consequence of osmosis is that it can lead to differences in pressure between the two compartments, which is termed *osmotic pressure*.

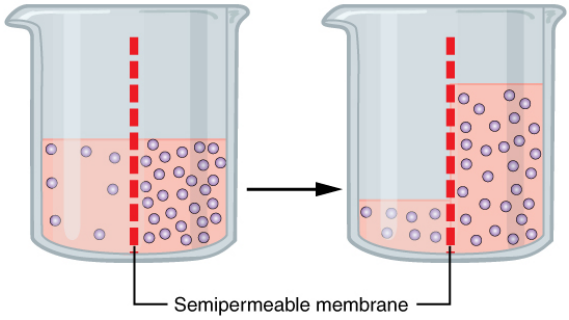



**Figure 14 - Sketch of an osmosis experiment in which an initial disparity in solute concentration across a semi-permeable membrane induces a flow of water. Osmotic pressure is defined as the pressure to apply to the right part of the beaker in order to restore the initial state. Image by OpenStax College, via Wikimedia Commons. License CC 3.0**

Although this simple explanation gives some qualitative idea of osmosis, it is more rigorous to represent osmolarity by considering the chemical potential related to water. Following such thermodynamic representation it yields that the osmotic pressure ($\pi$) applied between a pure

[19] A semi permeable membrane is permeable to some chemical species (typically water molecules) and impermeable to other (typically the solutes). Phospholipid bilayer membranes harbor usually aquaporins which are transmembrane proteins allowing influx and outflux of water.

solution of solvent and a solution where a solute is in concentration $c_i$ is approximated to the first order (*i.e.* for small solute concentrations) by Van't Hoff law:

$$\pi = R.T.c_i \quad (1)$$

where *R* is the perfect gas constant and *T* the temperature in K (59).

Under normal osmotic conditions (termed *iso-osmotic* or *isotonic*[20]), yeast cells maintain a given osmotic equilibrium with their environment where the cellular content has a slightly higher solute concentration than the environment. Since the cell wall and the cellular membrane are semi-permeable, this in turns produces an osmotic pressure which is balanced by cell wall tension and is called *turgor pressure*. Under typical growth conditions, turgor pressure in S. cerevisiae was estimated to be around 0.5 bar (BNID: 104997) reaching 2 bar in stationary phase (BNID: 104998). When the osmolarity of the extracellular environment changes, this balance between cellular and extracellular osmolarity is altered and has important consequences for the cells.

If the environment osmolarity decreases (it is termed *hypo-osmotic* or *hypotonic*), water will flow in the cell and increase turgor pressure. This can lead to cells swelling and even bursting although yeast, like plants, have a robust cell wall which makes this situation much less probable than it is the case for other cell types lacking a cell wall.

On the other hand, if the environment osmolarity increases (it is termed *hyper-osmotic* or *hypertonic*) water will flow out of the cell. If mild[21], this in turn will only reduce turgor pressure. If external osmolarity is higher, the cell volume will decrease and in extreme osmolarity (we consider here increases of extracellular solute higher than 2M to be extreme), plasmolysis may happen (where the cell membrane detaches from the cell wall).

In natural conditions (*i.e.* on fruits), it has been estimated than *S. cerevisiae* is exposed to external osmolarity of 0.1 to 1.5M (60), with external osmolyte being mainly sucrose and hexose. Osmotic stress is usually obtained in laboratory using either salts like NaCl and KCl or sorbitol.

In any cell, responding to changing osmotic conditions is essential to maintain homeostasis. In this respect, we speak of *osmotic stress* to emphasize the risk it represents for cells. Because important changes in osmolarity will change water activity, it will affect most of the chemical reactions occurring normally and, would it be unanswered, would prove very detrimental or lethal[22]. Having explained the basics of the physics besides osmotic changes, we will now briefly present how baker's yeast reacts to osmotic stress.

[20] In this thesis, we often use the term tonicity and its derivatives in place of osmolarity. Although this misuse is very common, it should be noted that both terms are not formally equivalent.

[21] Here a mild osmolarity increase is defined as one not leading to cell volume change. From Van't Hoff relation and considering a normal turgor pressure of 0.5 bar, a mild osmotic condition can be defined as lower than a 20 mM increase in extracellular solute concentration. For a turgor pressure of 2 bar mild osmotic stress would correspond to less than 80 mM increase in extracellular solute concentration.

[22] This explains the antique conservation techniques of drying food or salting it along with conservation of marmalade.

***Adaptation to hypertonic environment***

The adaptation of *S. cerevisiae* to increased levels of osmolarity has been extensively studied. Rather than giving an extensive review of what is known about this canonical stress response, the goal of the present description is twofold: to illustrate how global the impact of osmotic stress on cellular physiology is and to give a minimal description of the principal mechanisms at play for the proper understanding of the remainder of the work presented here. Readers interested in a more precise and exhaustive review might consider the following references (61–63).

The main response to hyperosmotic stress is mediated by the High Osmolarity Glycerol (HOG) Pathway. The HOG pathway coordinates several important acclimation/adaptations mechanisms acting at different time scales. In order to restore size and more importantly water activity, cells need to force water to enter the cytoplasm again. To do so, they will increase further their internal osmolarity by accumulating glycerol, a biocompatible osmolyte which was measured to counter balance up to 95% of the external osmolarity (64). The production of glycerol from common metabolic intermediaries in glycolysis occurs in two steps which are catalyzed by two pairs of paralog enzymes: Gpd1/Gpd2 and Gpp1/Gpp2. The sequence of events following an osmotic up-shift is depicted in Figure 15 from Miermont *et al.* and described below:

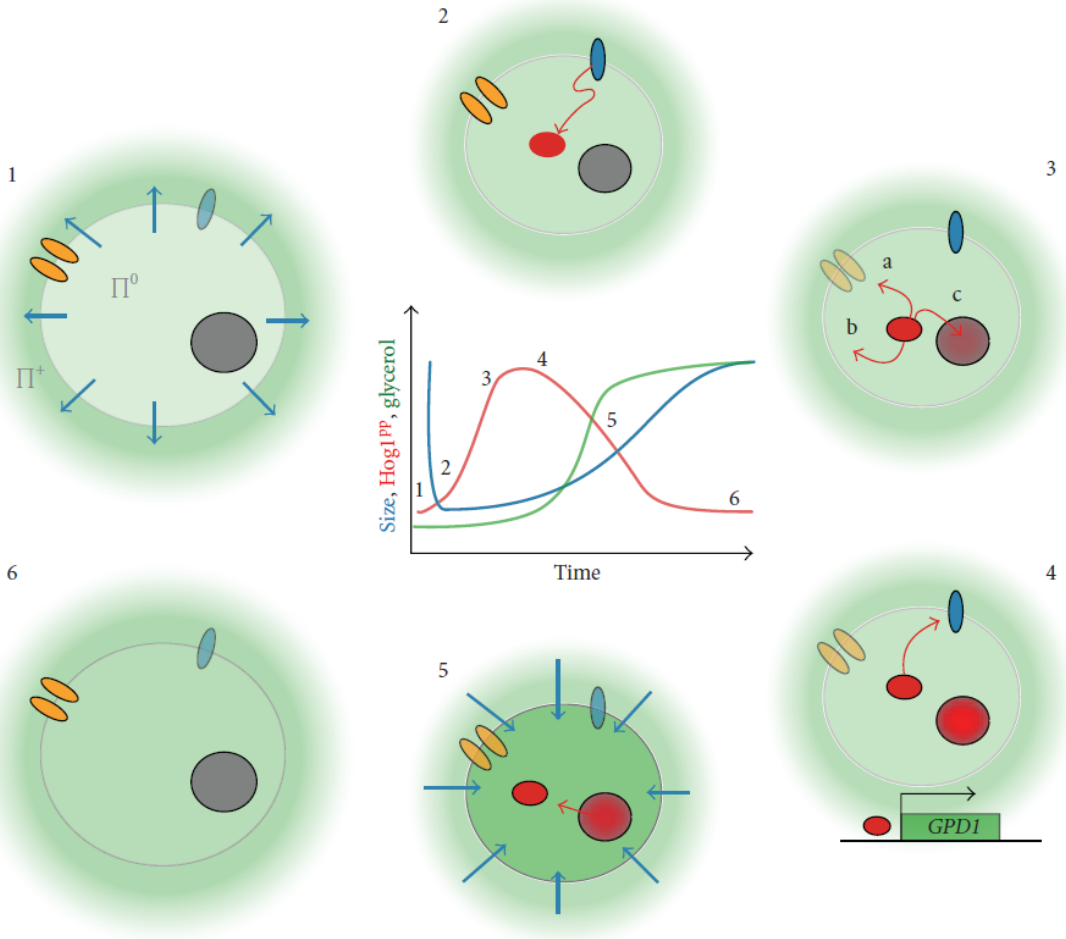



**Figure 15 - Sketch of the different steps involved in the HOG pathway response to a hyperosmotic stress. Reproduced from** (63)**. Phase 1:** Cells mechanically shrink by losing water. **Phase 2:** Osmosensors at the membrane are activated by the osmotic imbalance which initiates the HOG cascade. **Phase 3:** HOG pathway's activation culminates in the phosphorylation of Hog1. Phosphorylated Hog1 in the cytosol induces: (a) the closure of the main glycerol membrane export channel, Fps1; (b) activation of several enzymes involved in glycerol synthesis. (c) Phosphorylated Hog1 accumulates into the nucleus which alters the transcription of more than 600 genes including GPD1. **Phase 4**: Proteins of osmo-responsive genes (such as Gpd1) are produced and the HOG pathway is deactivated progressively. **Phase 5**: The build-up of intracellular glycerol makes water flow back in the cell which progressively recovers its initial size. The HOG pathway is completely deactivated**. Phase 6:** The cell is adapted, turgor pressure is restored and osmo-responsive genes are transcribed in a near basal amount.

**Phase 1**: Within the first second of a hyper osmotic shock, cells mechanically shrink by losing water. The volume reduction depends on the severity of the applied shock and will range between 20 to 40% reduction of the cellular volume given the range of concentrations used in our study.

**Phase 2**: Within less than a minute, osmosensors at the membrane are activated by the osmotic imbalance which initiates the HOG cascade.

**Phase 3** (a-b): Within 1 min, the sensing from the osmosensors is transduced by the HOG pathway which culminates in the phosphorylation of Hog1, a MAPK[23] which is the central molecular actor of the HOG pathway and has both cytosolic and nuclear actions. Phosphorylated Hog1 in the cytosol induces: (a) the closure of the main glycerol membrane export channel, Fps1, so glycerol won't leak out; (b) activation of several enzymes (including Gpd1, Gpp2 or Pfk2) involved in glycerol synthesis which increases their enzymatic activities.

**Phase 3** (c): Within 3 minutes, a significant portion of phosphorylated Hog1 accumulates into the nucleus where, along with other co-factors it will alter the transcription of more than 600 genes, which represents about 10% of yeast's genome[24]. For example, transcription of GPD1 is upregulated.

**Phase 4**: Within tens of minutes to hours, depending on the severity of stress, proteins of osmo-responsive genes are produced, including Gpd1, Gpp1 or Gpp2, previously mentioned and Stl1, a $H^+$/glycerol symporter which can actively pump glycerol from the extracellular environment and is commonly used as a reporter of hyperosmotic related gene expression. As glycerol is produced and thanks to hog1-dependent feedback mechanisms we only begin to understand, the HOG pathway is deactivated progressively. In fact, a system level analysis of a mutant lacking the Sho1 branch of the HOG pathway argued that this feedback includes a single integrator (which may consist of integrators in parallel) downstream of Hog1 but upstream of glycerol production (50). Yet, as it was postulated in (61) and recently at least partially confirmed, the feedback seems to be implemented at the level of the osmosensors Sln1p (65) and Sho1 which surprisingly interacts with Fsp1 (66).

**Phase 5**: The build-up of intracellular glycerol makes water flow back in the cell which progressively recovers its initial size. This leads to the complete deactivation of the HOG pathway with the de-phosphorylation of Hog1 which returns mainly to the cytoplasm. The fact that the HOG pathway is deactivated precisely when osmotic balance is restored is termed *perfect adaptation.*

**Phase 6**: The cell is adapted, turgor pressure is restored and osmo-responsive genes are transcribed in a near basal amount.

***Adaptation to hypotonic environment***

Compared to the adaptation to hypertonic environments which has been described, the response of *S. cerevisiae* to hypo-osmotic stress is far less characterized[25]. Following an immediate

[23] Mitogen-Activated Protein Kinase. Hog1 is highly conserved, with its mammalian homologue being p38.
[24] It is important to mention that a major portion of osmo-responsive genes are indeed differentially expressed for several or all types of cellular stress.

osmotic down-shift, we saw that cells swell and that turgor pressure increases as Baker's yeast very strong cell wall (20 to 30% of dry mass (67)) endures the increase in osmotic pressure. Concerning the mechanical response to a hypo-osmotic shock from normal condition and in a more pronounced manner during a osmotic down-shift from previously high osmolarity, it is essential to underline the role of Fps1 as a glycerol *valve* (61) which allows previously accumulated glycerol to leak out of the cytoplasm, therefore reducing rapidly the osmolarity of the cell.

After this immediate mechanical response, several stress mechanisms are directly and indirectly activated. Cell membrane sensors activate the cell wall integrity pathway (CWI) also called Protein Kinase C (PKC) pathway (68, 69) which orchestrates the reinforcement of the cell wall. Hypo-osmotic stress is also known to produce a transient increase in cytosolic $Ca^{2+}$ (70). An ensemble of stress responsive genes whose expression is affected in several stress conditions is called the Environmental Stress Response (ESR) and is mainly controlled by the Msn2 and Msn4 transcription factors (71). Yet, there is empirical and sometimes direct evidence of a direct regulation of some genes of the ESR under osmotic stress which could possibly be mediated by the Skn7 transcription factor (61). Interestingly, genome-wide transcription analysis following hypo-osmotic stress revealed that the gene expression change pattern is in a large part opposite to that occurring during a hyper-osmotic stress as we can see in Figure 16 from Gasch *et al*. In this figure, we can also notice that

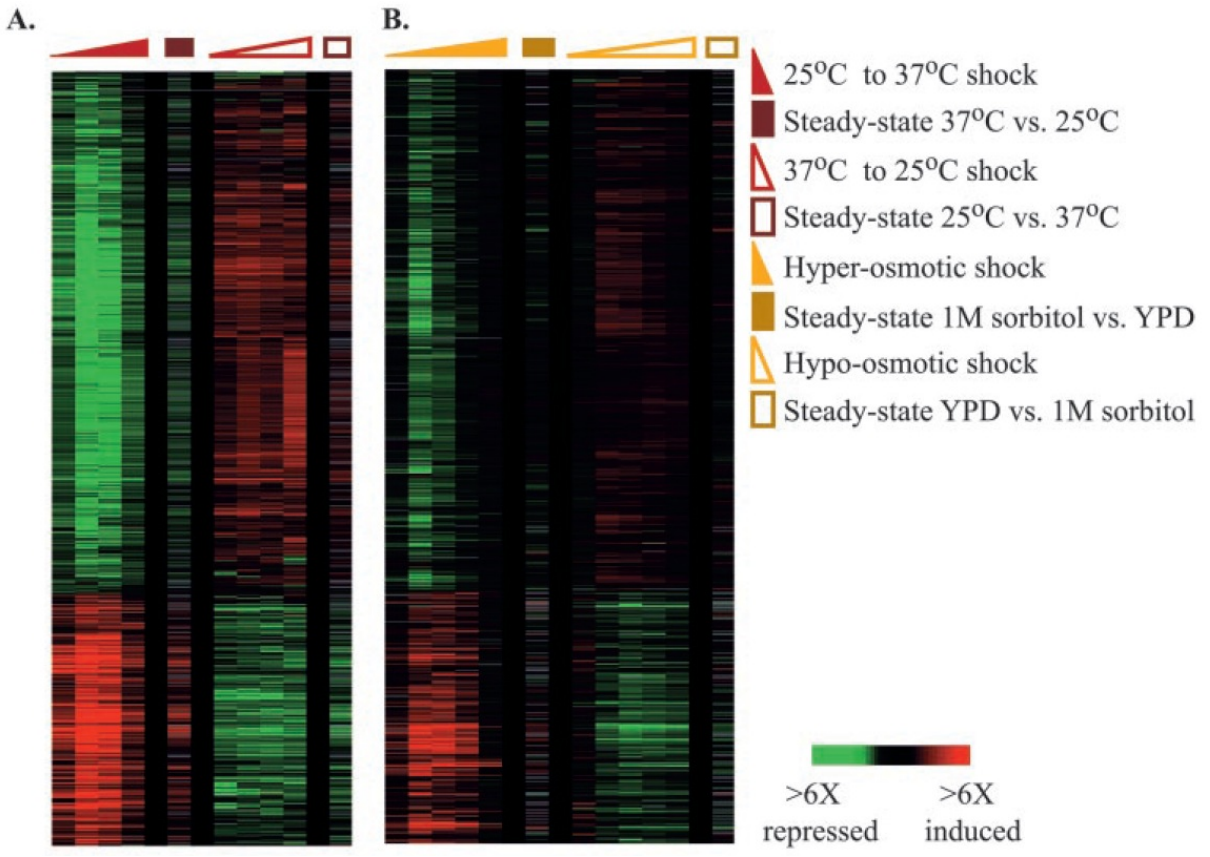



**Figure 16 - Gene expression change over 1h following stress for genes composing the (general) Environmental Stress Response (ESR) which are transcriptionally affected by most stress types. Figure reproduced from (71) and based on micro-array transcriptome analysis.**

[25] The reason why hypo-osmotic stress is less known may include the lesser interest for the bio-production industry compared to hyper-osmotic conditions, the milder effect on cellular activity since yeast cells can endure very high hydrostatic pressure (in the tens or hundreds of MPa (164)) and the fact that the response to hypotonic condition seems to rely mostly on the combined activation of other stress response mechanisms.

hyperosmotic stress leads to a more transient transcription (corresponding to the period when Hog1 is nuclear) than it is for a hypo osmotic stress which dynamic is more similar to the heat shock response.

### b. Yeast response to osmotic stress as a model cellular process

***Osmotic stress transcriptional response is transient and global***

We previously mentioned that hyperosmotic stress affects directly the transcription of 10% of *S. cerevisiae* genes, most of them being commonly induced by many different stress conditions (71). But what such a number really represent at the scale of the overall gene regulation network is not straightforward. As a first comparison, we know that 12% of genes are estimated to be expressed in a cell-cycle dependent manner (72). Focusing on Hog1, we can ask which portion of yeast's GRN is directly connected to this multifunction (pleiotropic) protein.

In Figure 17 we propose a visualization of the interactions of Hog1 with other proteins. This representation uses a synthetic view of yeast's proteome under normal conditions as a basis on which we overlay information about Hog1 interactions. The basis representation, from (73) and already shown in the general introduction (Figure 1) represents proteins as Voronoi regions. The surface of a region is proportional to the abundance of such protein. The relative position of regions (proteins) is related to how related and interacting proteins are. From this representation of the overall proteome under normal conditions, we color proteins which interact[26] with Hog1, genetically and physically (Hog1 itself is barely visible and is located in the yellow region on the right).

It appears that Hog1 interacts through at least one gene (or protein) with all the major cellular processes depicted here. Also, an important number of associated genes are in the unmapped grey territory whose genes and proteins are those for which no clear function or physiological effect is known.

[26] Note that the interaction database used in this section, (165), compiles many types of experimental data. Having different experiments and estimation procedures means that the level of information or the confidence in detected interaction among proteins is variable.

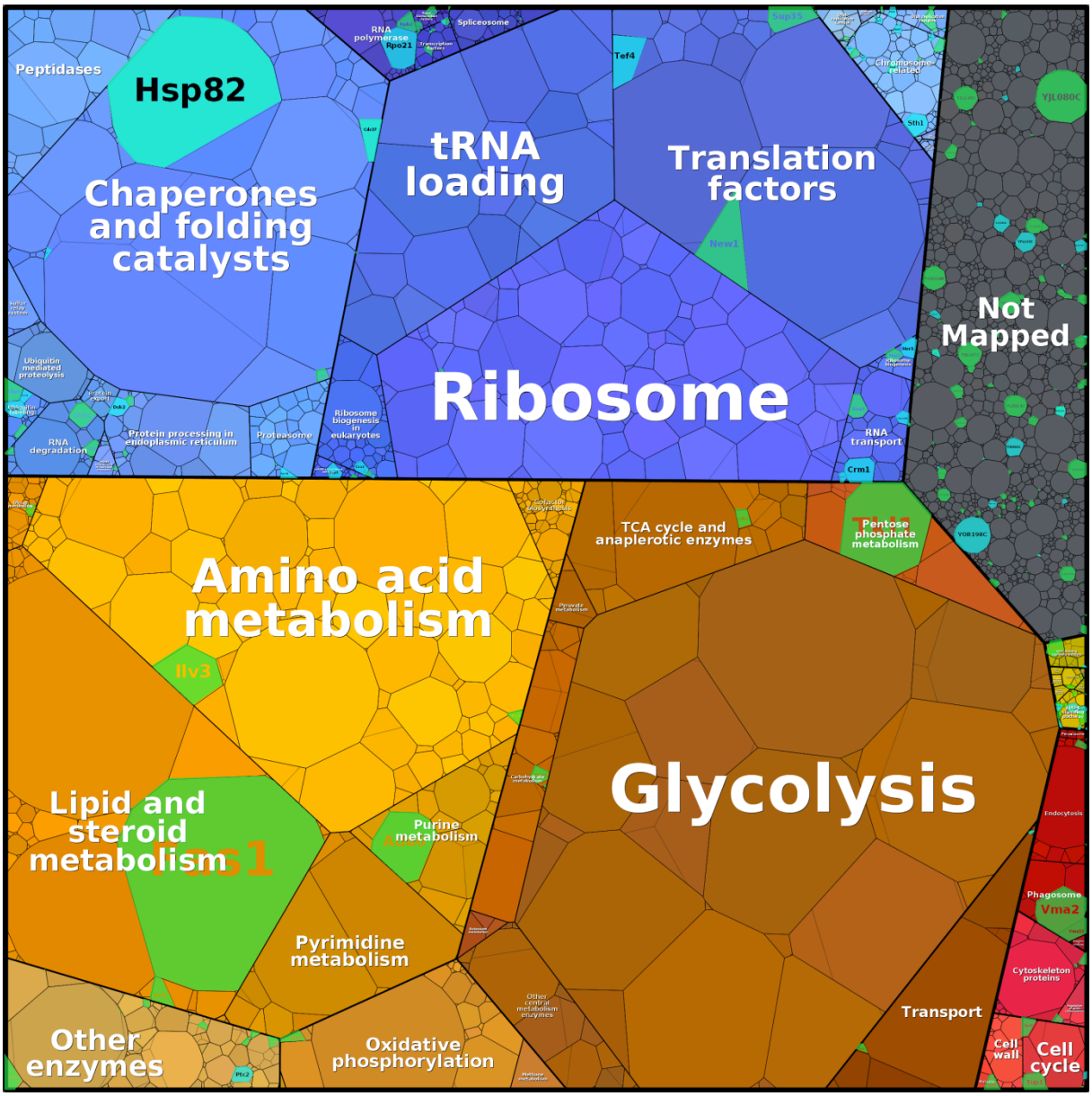



**Figure 17 - Voronoi representation of S. cerevisiae proteome in normal growing condition, with superimposed in green proteins genetically interacting with Hog1 and in light blue proteins physically interacting with it. Protein abundance are given by mass spectrometry in (32) Voronoi visualization comes from (73), interaction database from (165). The area of the Voronoi regions correspond to the normal protein abundance and their position in the diagram to how related the corresponding proteins are.**

Yet, in Figure 17 we highlighted interactions which do not necessarily take place at the same time. In fact, protein-protein interactions (in light blue) are much faster than genetic ones (in green). In order to have a coherent view of the influence of Hog1 at the transcription level, it would be necessary to consider not only genes whose transcription ifs affected by Hog1 (among reported genetic interaction of Hog1), but also those whose transcription is modulated by the proteins Hog1 physically interacts with. In Figure 18, these indirect transcriptional effects have been added. What appears is that through both physical and genetic immediate interactions, Hog1 is connected to most of the overall gene regulatory network[27]. Needless to say, considering genes one degree further in

[27] More than half the genes can be affected which represent roughly 70% of the proteome in terms of abundance.

terms of interaction from Hog1 would lead to virtually all of the genes represented here being accessible.

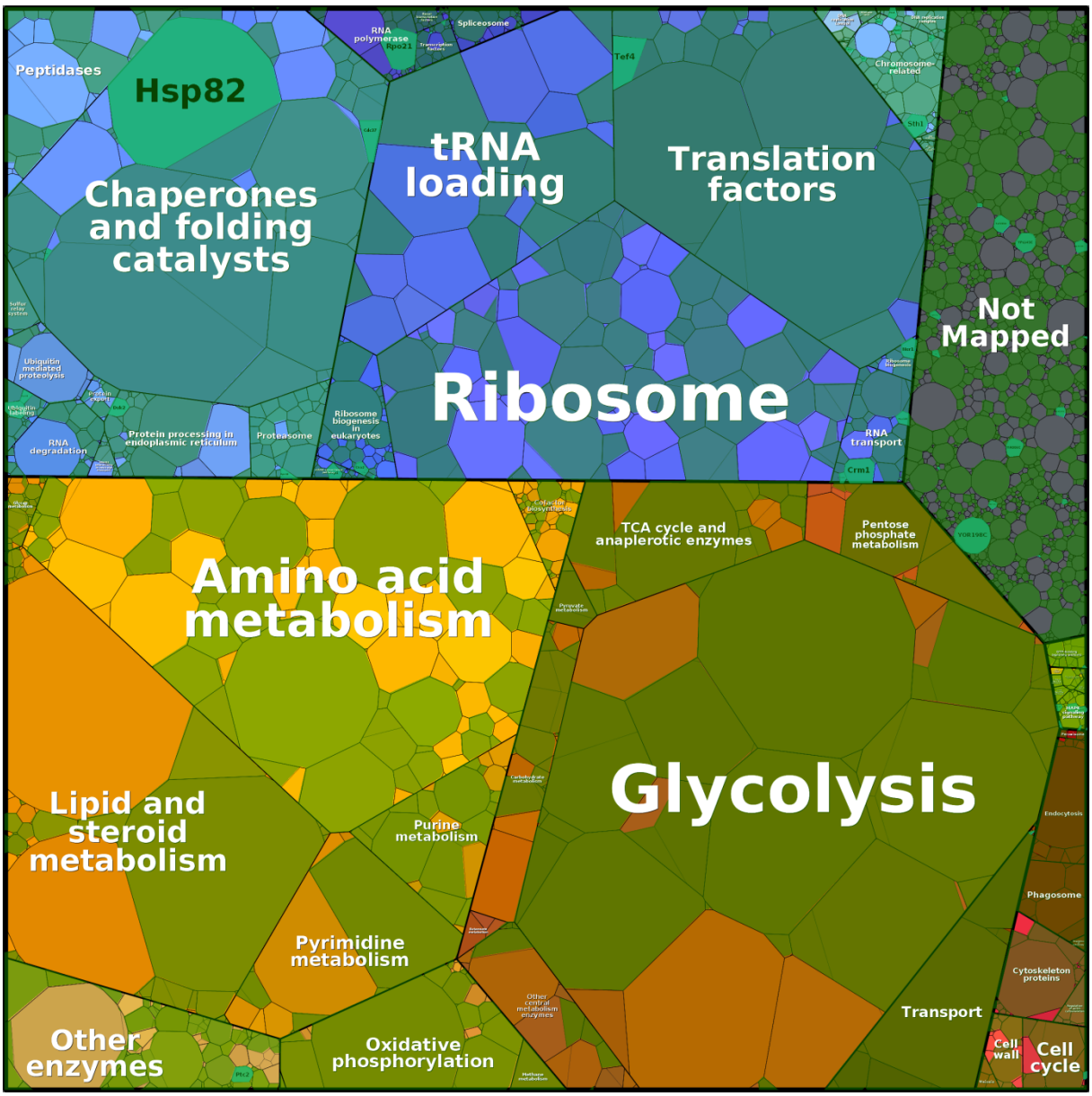



**Figure 18 - Voronoi representation of S. cerevisiae proteome in normal growing condition, are superimposed in green: proteins genetically interacting with Hog1; proteins physically interacting with hog1; genetic interaction of the proteins which physically interact with Hog1. Mass spectrometry data from** (32) **Voronoi visualization from** (73)**, interaction database from** (165) **The area of the Voronoi regions correspond to normal protein abundance and their position in the diagram to how related the corresponding proteins are.**

It is important not to get confused about this representation as what is depicted here is more the *potential action* of Hog1 on the overall gene regulatory network through known interactions, than a representation of the actual impact following an osmotic stress. It illustrates well the fact that HOG1 is a pleiotropic gene, a *hub* in the overall GRN. In annex 3, we provide a similar visualization representing a time course of genome-wide transcriptome changes upon hyperosmotic stress. It reveals that the actual transcriptional impact of hyperosmotic stress is globally closer to what is represented in Figure 18 than to the restrictive view given in Figure 17.

As it was evoked for the adaptation to hypo-osmotic conditions and visualized in Figure 16, many genes which are affected by hyperosmotic stress are also affected by other stress conditions such as heat shock. It is commonly admitted that *S. cerevisiae* has a general stress response, the ESR, which is activated in nearly all stress condition (71) in addition to stress specific responses such as that of the HOG pathway for hyperosmotic stress[28]. The ESR is modulated by many different pathways (*e.g.* the TOR pathway, the PKA pathway, etc.) and although lacking a clear picture, is related among other mechanisms (74) to the activity of Msn2/Msn4 transcription factors and to STRE[29] promoter elements (61, 75). It is also important to state that parallel to the overlap in the transcriptional response by various signaling pathways (which composes the ESR), there is evidence of several cross-talks between different signaling pathways themselves. For instance, the HOG pathway and the pheromone response pathway (76).

Upon osmotic stress, Hog1 leads a global, yet transient modification of gene expression in cells. As it was evoked in the previous subsection, yeast's response to a hyper-osmotic stress follows a precise temporal pattern which depends mostly on the severity of the stress. Figure 19 gives a collection of time profiles giving a more precise temporal vision of several aspects of osmotic stress response. In Figure 19 A from Muzzey *et al* we can see that the level of external osmolarity applied

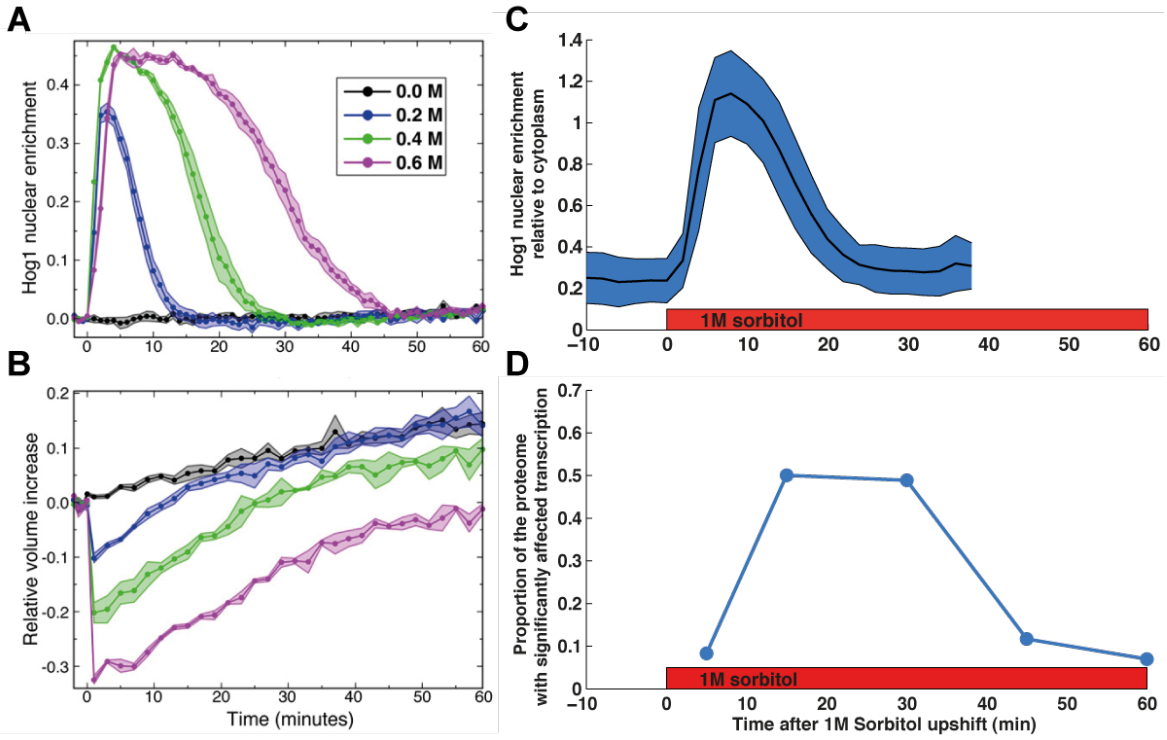



**Figure 19 - Timing of the response to hyperosmotic stress. In A and B are represented the impact of the osmotic stress strength (in NaCl concentrations) on Hog1 nuclear localization and cellular volume dynamics. A and B are reproduced from (50) and shaded areas represent the standard error on the mean for several experiments. In C, we show the dynamics of Hog1 nuclear localization for a 1M sorbitol stress (comparable to a 0.5M NaCl stress) with black line showing the population average whereas the shaded area is the standard deviation between single cells. In D, we report the proportion of yeast proteome which is transcriptionally impacted following a 1M sorbitol osmotic stress. We used mRNA time course data from (71) to extract genes whose transcription differed from more than 20%. The proportion of genes with differentiated expression was weighted by the fraction of total protein accounted by each gene using proteome abundance data from (32).**

increased resistance to other type of stress (75).

[29] STRE stands for Stress Response Elements which are specific DNA sequences often found in promoters of stress response genes.

during an osmotic stress affects mostly the duration of Hog1 activation and nuclear localization rather than its level. When it comes to cellular volume (Figure 19 B), we observe an apparent linear increase of the volume loss upon osmotic stress with the applied osmolarity. In addition, to some extent, it appears that Hog1 deactivation and subsequent nuclear export is approximatively concomitant with volume recovery. For a standard osmotic upshift (Figure 19 C, our own data) induced by 1M sorbitol, we see that Hog1 is active and nuclear for about 20 to 30min, affecting transcription. Yet, if we look globally at transcriptional data as in Figure 19 D, we can see that mRNA levels will still be higher than normal ~10 min after Hog1 which is expected for regularly stable (the median mRNAs half-life in *S. cerevisiae* is 20 min, BNID 100205) transcripts.

When considering the impact of osmotic stress on gene expression, the transcriptional control is not the only aspect to be taken into account. For instance osmotic stress impacts also mRNA stability *in vivo* (77) both in an unspecific manner (by globally destabilizing mRNAs) and in a specific fashion (by stabilizing most of mRNA related to osmotically induced genes). Other post-transcriptional modifications include an impact on protein translation (again both unspecific and specific) though activation of the Rck2 kinase (78). In addition, the remodeling of the epigenetic chromatin landscape has also been recently related to the action of the HOG Pathway (79, 80).

***Osmotic stress and cellular physiology***

By definition, osmotic stress is a physical phenomenon. It has consequences both in terms of chemistry because of a reduced water activity and on mechanical properties of the cell by means of cell wall tension and turgor pressure. Turgor pressure is known to affect cell wall properties and is central to remodeling the shape of walled organisms which is essential in budding or mating (81). Osmotic stress may even induce nutrient starvation by affecting membrane active transporter (61). It is worthy to note that the impact of severe osmotic stress goes beyond a simple increase of what happens in mild stress because it leads to macromolecular crowding (82). This affects the normal diffusion of molecules in a size dependent manner and therefore affects nearly all chemical reaction rates. It is also probably the reason why at very high osmolarity, the kinetics of the HOG pathway are slowed down, a phenomenon which was previously unexplained (61).

Given all the aforementioned effects of osmotic stress, and considering the potential impact pictured in Figure 18 on proteins involved in metabolism and cell-cycle, it comes with no surprise that osmotic stress effectively affects cellular proliferation activity. In fact, osmotic stress pauses proliferation both by stopping the cell-cycle and through modifications of metabolic activity until the cell is adapted and proliferation resumes (61). This will be discussed in more detail in chapter IV, where we present experimental quantifications of the impact of osmotic stress on growth and cell cycle which are fundamentals elements of a cell's physiology.

From our cursory review we can state that most challenges encountered in quantitative biology are present in the study of osmotic stress. For example it is a highly dynamic process which includes several phases and sub-processes with distinct time scales. Some elements carry many distinct functions depending on the molecular context[30] or time scale[31] and our knowledge of the

[30] For instance Hog1 display many different roles between its phosphorylated, non-phosphorylated, nuclear or non-nuclear forms. Also, spatial proximity of Sln1 and Fsp1 on the membrane modify the activity of the former (66).

structure of interactions among key molecular players is still incomplete. Also, we can expect the cellular physiology to be both affected by stress response and to condition the cellular adaptation to it. This later challenge is at the core of this study. For all these reasons, the response to osmotic stress can be considered as a *model* process. This means that we can expect methodological improvements in the comprehension of this system to be at least partially translated to many other biological systems.

When studying such a process from a systems biology perspective, the use of quantitative analysis and mathematical modelling in particular is essential and the next paragraph aims at reviewing existing modelling approaches of the many biological phenomena triggered by hyperosmotic stress.

[31] Again, Hog1 acts differently at different time scales: Its action in the cytosol as a kinase has a typical time scale of protein-protein interactions (<s to s). Its nuclear localization, where Hog1 is passively transported (although retained actively) has a typical time scale of minutes. At last, its action as a transcription factor as a typical time scale in tens of minutes to hours.

### c. Modelling the cellular response to osmotic stress

As explained, yeast's response to osmotic stress displays such complexity and interplay that it can be considered as a canonical example of cellular process. There is an abundant literature concerning osmotic stress and Hog1 and many different experimental and analytical methods have been used on this system. In this subsection we try to summarize the different mathematical modelling approaches aiming at representing the HOG pathway rather than providing a thorough review. Only dynamic and quantitative models will be considered here although other interesting approaches (*e.g.* logical/boolean models) have been applied to the HOG pathway. Finally, we mention that more detailed discussion of some models particularly relevant to our work will appear later in this thesis.

#### *Detailed mechanistic models*

A straightforward modelling approach consists in translating into mathematical models all available knowledge on a system. Such explicit models are usually formulated as a system of ODEs. Building this kind of models is not very difficult from a conceptual point of view since they do not abstract any details. Yet, the precise definition of interactions is usually tricky along with assigning values for the many parameters that come with detailed descriptions. Moreover, given existing experimental limitations, some parameters can be structurally or practically non-identifiable[32]. Nevertheless, detailed models are usually able to benefit from comparison with a very broad range of experiments (including mutants or dynamical measurements of any modelled species) and when properly constructed and parametrized can be highly informative. For example, detailed models can be used to test different biological assumptions for which no obvious experiment can provide a direct answer. This approach was used in (83) to study the interplay between the two signaling branches of the HOG pathway. It should be noted that detailed models which include all component at the same level can also be helpful in finding proper abstractions (*i.e.* answering *what matters?* for a given coupling aspect) by using well defined mathematical aggregations methods of several precise components into an *abstract* one (*e.g.* adimensionalization or time scale separation). In the case of osmotic stress some detailed models include complex couplings such as the interplay of: physical mechanisms, MAPK signaling, Hog1 localization, osmo-induced gene expression, and metabolism (84); or Hog1 cytosolic response, glycerol metabolism, cellular energy utilization, growth and turgor pressure (85). Currently, there is no detailed model which integrates cell-to-cell variability and this leads to important issues discussed in chapter III and our article in annex 6.

#### *Engineers and physicists models*

In line with the idea of reverse-engineering of biological systems, several studies have looked at osmotic stress from an input-output perspective. These studies are interested mainly in the phenomenological aspect of the HOG pathway, and particularly in the feedback mechanism which allows *perfect adaptation*[33] (50). Experimental data is usually in the form of temporal profile and part or all of the signaling cascade can be studied in terms of its transfer function (which displays properties of a low-pass filter) (54, 55). These models are usually deterministic and minimal in their

[32] See (166) for more information on non-identifiability of ODE models and the supplementary information of the article in annex 6 for an example of non-identifiability analysis in the context of our work.

[33] Perfect adaptation can be defined as the property of the HOG pathway to shut down as the osmotic balance is recovered without over-shooting (*i.e.* over reacting) or missing its target (*i.e.* under-reacting).

description of cellular processes so as to capture only the system level properties required. For instance, in (86) a minimal model of gene expression proved to be sufficient to control gene expression in real time with a model predictive control approach.

***Stochastic models***

As it was presented in section 1, biochemical reactions and more specifically transcription of mRNA can occur in a random manner. To quantitatively model such behavior, stochastic models are used. These are generally constructed from chemical reaction stoichiometry only using the chemical master equation (CME)[34]. These models being probabilistic in nature, many independent realizations are needed to be measured in order to infer their parameters. Although dual reporters or others similar experimental systems should be preferred, most of the time these models are inferred from flow cytometry data. Given the important computation burden of exact simulations (typically performed with the SSA[35] algorithm which uses a Monte Carlo approach), the number of modelled species is limited and parameter fitting is very long. Nevertheless, from the CME can be derived ODEs for all statistical moments (mean, variance, skewness, etc.). Using particular assumptions like moment closure[36] allows therefore parameters of the CME to be fitted to experimentally measured moments at much lower computational cost. Stochastic models are useful to represent the impact of random fluctuations in gene expression on the overall behavior of a biological system, both at the single-cell and population levels. In the case of the HOG pathway, such model (87) along with moment closure and an extrinsic parameter[37] was used to explain reported bi-modal expression of pSTL1 at intermediate osmolarity (88). Another interesting example is provided in (89) which aimed at characterizing the precise nature[38] of pSTL1 stochastic expression. Several of these models will be precisely described in chapter III.

***Integrative and hybrid models***

Other notable models of the response to osmotic stress display both explicit, detailed mechanisms and engineer type of black box input-output modules. This allows models to be adapted to the precise scope of the research question at hand and to accommodate with an uneven knowledge of all interconnected processes. A good example is given in (90) which focuses on the interplay between cell-cycle (S to G1 transition) and osmotic stress. In (91), a simpler mix of detailed and black-box modelling was calibrated to time-varying data and used to investigate *in-silico* the response of the HOG pathway to various fluctuating stress patterns. Finally, population models such as Mixed Effects models which will be detailed in chapter III allow combining a semi-detailed model of pSTL1 response to hyperosmotic stress with cell-to-cell variability (92) and to compare such representation of cellular heterogeneity with that obtained by stochastic formulations from (87).

[34] A rigorous derivation of the CME from mechanical considerations is given in (167).

[35] Gillespie's Stochastic Simulation Algorithm (SSA) is a type of Monte Carlo method generating statistically exact trajectories for a stochastic equation (168, 169).

[36] Moment closure is a set of assumption regarding the modelled distribution which is necessary in most cases since the equation for $n^{th}$ moment depends on the $n+1^{th}$ moment, see (170) for examples of derivations.

[37] This extrinsic parameter was suggested to be related to chromatin remodeling complex abundance. Mathematically, it allows, on top of the typical stochastic gene expression, to account for a cell-specific parameter value within a typical stochastic model.

[38] See the next section for a description of the two main stochastic regimes in gene expression: Poisson and Bursty.

## 5. Introduction Conclusion and Outline

Along this introduction, we presented several open questions concerning the fact that biological processes do happen within single cells which although similar are not identical. In isogenic cell population, the physiological state of each cell is different and most cellular activities are affected by it. For a given process, the subset of the cells physiology which can impact it constitutes therefore a context and a same process occurring in different cells can therefore display some extrinsic variability.

Here we focus on cellular information processing and on gene expression in particular. This biological process is dynamic and emerges from the many interactions between genes, proteins and other elements. In sections 2 and 3 we described important methodological and experimental considerations for this study such as the use of single-cell longitudinal information along with dynamic stimulations.

In section 4 we presented the biological process we used in our investigation: the response to osmotic stress. In particular we presented the large and transient change in gene expression which is triggered by the HOG pathway and Hog1. In the following chapters, we present our investigations of two aspects of the relationship between process and context in the osmotic stress response. More precisely we ask:

- **To what extent the transcriptional response coordinated by Hog1 is affected by its cellular context, leading to cell-to-cell variability?**
- **How the stress response itself can in turn affect proliferation (which is a major driver of cellular physiology)?**

In chapter II we will discuss technical developments allowing long term experiments in controlled dynamic environments such as the design and use of microfluidic chips and image analysis methods to recover single cell information on gene expression and physiology from microscopy data.

In chapter III we will present our investigation of cell-to-cell variability in gene expression under repeated stress

In chapter IV we will present our quantitative analysis of the impact of repeated osmotic stress on cell-cycle and cellular growth.

We will then propose general perspectives for our research topics and mention additional developments undertaken during this thesis which will contribute to future projects. At last we will present our general conclusions and final discussion of the presented investigations.

# II. Long term dynamic experiments and single-cell data

In this chapter, we will describe the methods that were developed specifically for this project. Most experiments were carried within the general pipeline depicted below.

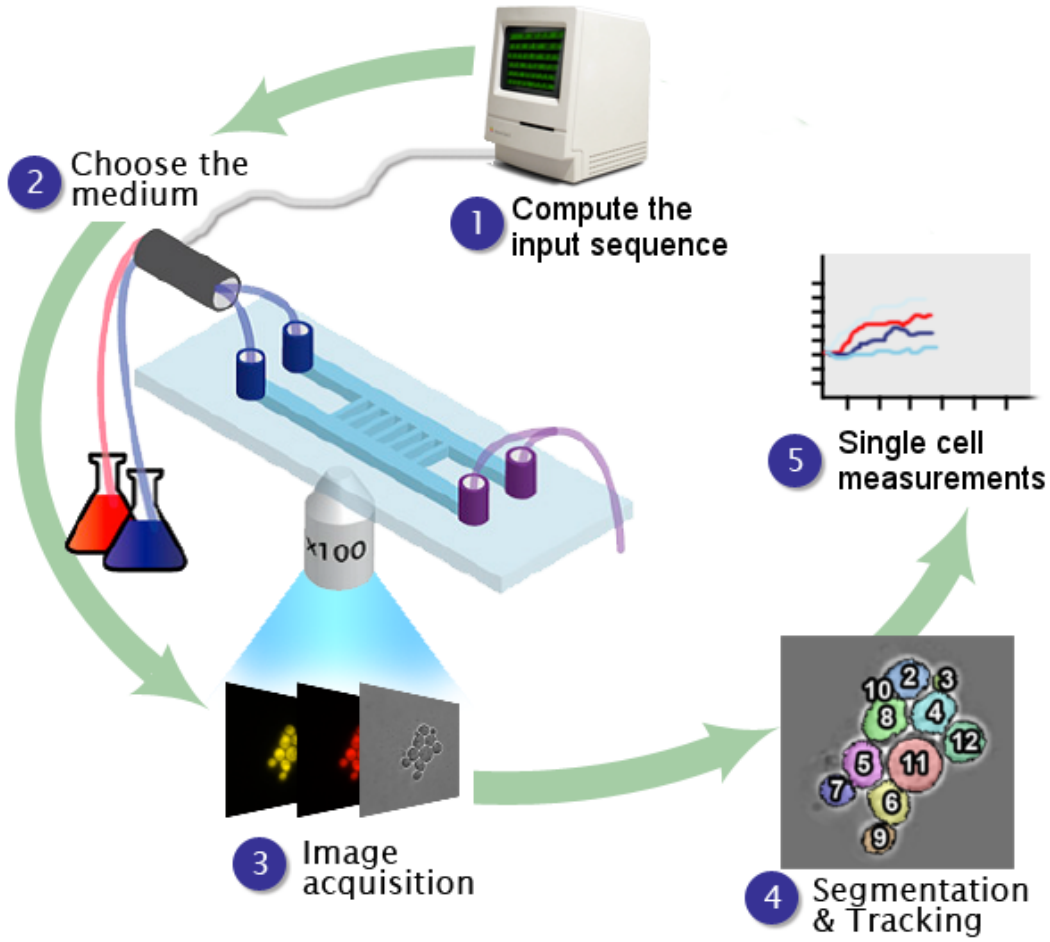



Figure 20 – Schematic representation of the main experimental pipeline used in this study.

# 1. Single-cell measurements in precisely changing environments using microfluidics and microscopy

## a. The Truman show: the use of microfluidics

Using microfabrication techniques like soft lithography[39], it is possible to construct systems adapted to handling very small quantities of liquids (in the order of µL to fL). Fabrication at this scale is used for many different purposes (93–95) and can include many components as it is reflected in the term *lab on chip* which is often employed to describe complex microfluidic systems. In this work we are interested in the possibility of creating chips which fulfill the following objectives:

- Providing a controlled homogeneous environment for cells. This can be ensured by constantly renewing the liquid medium, thereby removing waste and ensuring a constant availability of nutriments. Other physical properties like temperature or pressure or mechanical cues (like shear stress coming from liquid flowing on cells) can also be controlled (53).

- Controlling the chemical composition of the medium dynamically so as to impose time-varying concentration of chemicals for instance.

- Improving observation conditions *e.g.* by forcing cells to grow as a monolayer and therefore prevent cells from overlapping (which makes it impossible to distinguish them).

- Allowing several experimental replicas to be performed simultaneously or allowing several experiments to run in parallel.

A typical chip is composed of several elements:

- Imaging chambers where cells will be observed. These chambers usually have a low height in order to force cells to grow as a monolayer. For haploid *S. cerevisiae* cells, chambers of 3 to 5 µm are commonly used.

- One or more flow channels where culture medium can flow. These channels are usually higher than imaging chambers and include connectors regions which will be connected to tubes outside of the chip.

Some additional common elements include: loading channels which are used to inject cells inside the imaging chambers; mixers (96) which have a geometry helping fluid mixing[40]; geometric elements such as pillars, nozzles, traps etc. Complex chips can also include control layers which are

[39] It is interesting to note that, as it has been tested in our laboratory, recent improvements in stereolithographic 3D printing already allow *minifluidic* devices to be constructed much more easily (and with much cheaper equipment) compared to soft lithography. It can be expected that most microfluidic devices will be constructed with this technology in the future.

[40] It is important to bear in mind that at the scale of typical microfluidic chips, fluid motion is quite different from what we are used to. More precisely, the Reynolds number is pretty low which means that flows will be very laminar. This in turns means that mixing rapidly two liquids is difficult and usually requires geometries creating turbulence to reduce mixing time.

composed of channels containing pressurized liquids that allow for example to control fluxes in flow channels using valves or peristatic pumping on chip.

***Microfluidic chips used in this project***

**Figure 21 - Picture of one microfluidic chips used in this work (parallel-H-Shaped). Left: picture of a wafer covered in hardened PDMS and which acts as a mold. Right: Once cut out from the wafer, a chip is punched, plasma-cleaned and bonded to a coverslip.**

A microfluidic chip itself is usually made using a mold (*a.k.a.* a master or wafer Figure 21 Left) on which molded resins are hardened, cut out and bound to a coverslip (Figure 21 Right). While some chips require several layers of channels and therefore several molded parts that are assembled together, in our case a single mold making both culture chambers and flow channels is used. Using single-layer chips simplifies significantly the fabrication process and reduces the proportion of faulty chips.

We design and make our own masks and wafers (protocols for master mold design fabrication and chip fabrication are given in annex 4). This is not only cost efficient, it also allows us to fine-tune designs and parameters such as chambers height in order to obtain a microfluidic chip which really fits our needs. For instance, depending on culture conditions (high or low growth rate), I found chips with chambers of different heights[41] to be most efficient in trapping cells in chambers.

[41] For big fast growing cells, chambers of 3.5 to 3.7µm high are good whereas smaller cells growing slowly will stay more easily in a 3.1µm high chamber.

For the work presented here, two designs were used: a H-shaped chip (Figure 22) and a parallel H-shaped chip (Figure 21 and Figure 23). I made wafers of H-shaped chip with 3.7 or 3.1µm high chambers and flow channels having a height of 50 to 80 µm..

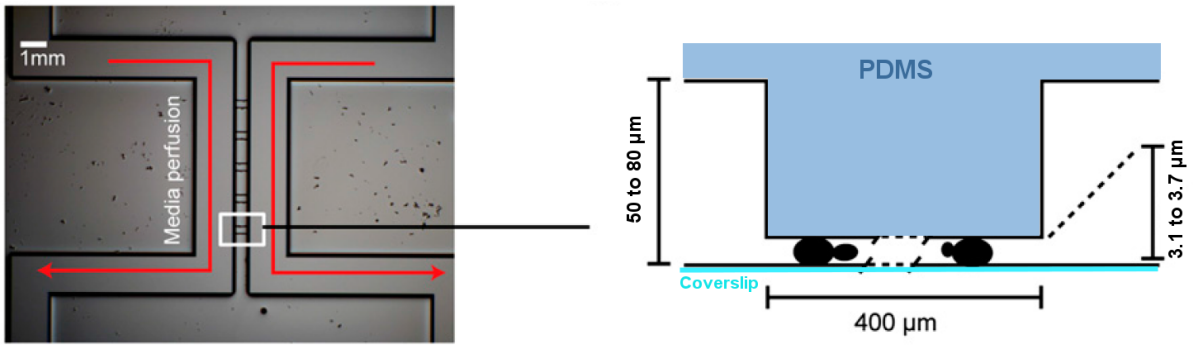



**Figure 22 - Left panel: Microscopy image of a H-shaped chip. Red arrows indicate medium flow in flow channels. This chip has 5 culture chambers, one identified by the white box. Right panel: sketch representation of a cross section of the white rectangle. Figure adapted from** (86)

We can see in Figure 22 that culture chambers are connected to two large flow channels. With this design, the medium inside the chambers is renewed by diffusion only[42]. This design makes it possible to renew the liquid content of a chamber within ~2 min (see Figure S1B of (86) or computations in Annex 5) while keeping cells in position. In addition, the H-shape conveniently allows using the flow channels for cell loading: if necessary, it is possible to inject cells from one flow channel extremity while blocking the exit of that same flow channel. This in turns forces a flow in the culture chambers where cells get trapped.

Using the H-shaped device, it is possible to perform relatively long experiments, obtaining 5 fairly independent[43] replicas (one per chamber) at a time. This was convenient for experiments used in chapter III. Nevertheless, the study of cell growth presented in chapter IV required longer experiments which were hardly doable with the H-shaped chip. Indeed, starting with a relatively low number of cells (30 in field of view overall), and under normal growth conditions, image field and more importantly culture chambers are completely filled in typically 6 to 10 hours. This makes it difficult to conduct much longer experiments because at high density, image analysis becomes very challenging, mechanical strain due to continuing growth in a crowded chamber may have physiological effects and homogeneous nutriment availability across the chamber is not ensured anymore (see Annex 5).

To allow longer experiments to be performed, I designed a modified version of the H-shaped chip represented in Figure 23. This chip has larger chambers (400x400 µm) which will delay the complete filling of chambers (yet as the field of view is determined by the microscope camera and

[42] Residual low-velocity flow in chambers can still happen from time to time.

[43] It could be argued that cells in the most upstream chamber could alter the composition of the medium which will reach chambers downstream. Yet, given that we impose typically a flow of 120µL.min$^{-1}$ for each flow channel, we see that the overall volume of one chamber is flown in the flow channels every 70 µs. Said differently, anything produced by cells in a chamber over 1 min will be on average diluted nearly a million time by the flow in flow channels.

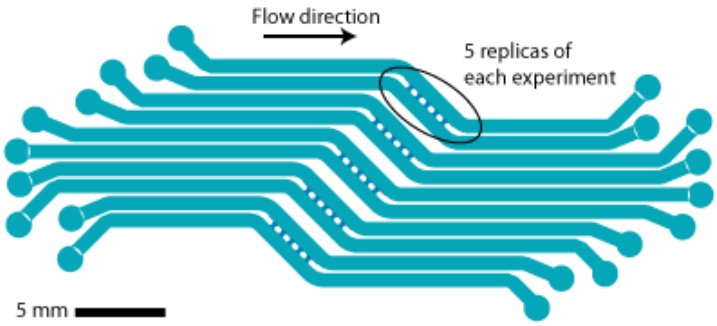



Figure 23 - Sketch of the parallel H-shaped design

optics, it fills up at the same rate, the chip delays only possible physiological effects and do not changes the image analysis challenge of crowded field of view). In addition, this chip allows 5 different experiments to be performed in parallel (with again 5 replicas for each). This increases the experimental throughput and places us at the limit imposed by the microscope: stage motion, automatic focus and fluorescence filter change (for one color) limits imaging to 25 fields of view every 3 min.

Another important aspect is that in this new design, chambers are not only larger; they are squares[44] instead of rectangles in the H-shaped device. This improves the nutriment homogeneity at high densities as shown in annex 5. This design features some less critical improvements such as: round connecting areas which reduce the possibility of tearing PDMS[45] during punching; larger angles in flow channels which remove the possible formation of small bubbles which often occurred in the straight angles of the basic H-shaped chip; smaller flow channels (800x50 µm cross-section) which allow to impose the same fluid speed without consuming as much medium compared to the basic H-shaped chip[46].

[44] In fact, I designed different masks allowing changing the chambers form. Square is called type C and type A has the form of an hourglass.
[45] PDMS is the abbreviated name of the typical polymer used to make microfluidic chips. See protocol in annex 4 for more details.
[46] This obviously reduces a bit the chamber volume to flow dilution discussed in footnote 43 for the H-shaped chip. Nevertheless, with a dilution ratio of 110 000, replicas can still arguably considered as independent.

***Digital control of cellular environments with valves***

During an experiment, we flow fresh medium constantly in flow chambers in order to ensure homogeneous conditions. This is performed by peristatic pumps (Figure 24 D) (Ismatec IPC) which impose a constant flow rate to the flow channels (we use 120 and 80 µL.min$^{-1}$ for each flow channel for H-shaped and parallel H-shaped chips respectively). This is done by *pulling* fluid out of the media bottles (Figure 24 A) through the chip (Figure 24 C) and then to the trash (Figure 24 E) rather than by *pushing* it. Such a pulling arrangement ensure no media will leak on the inverted microscope would there be a leak on the chip.

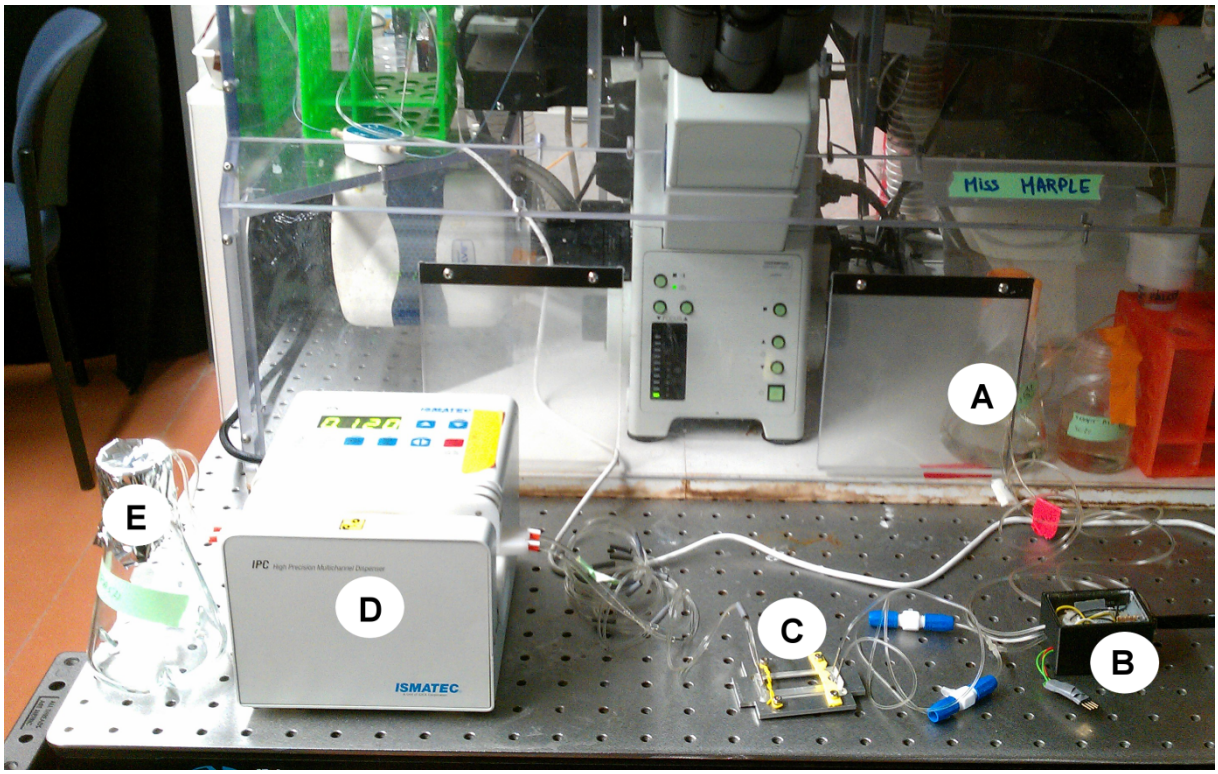



**Figure 24 - Image of our microfluidic experimental setting (here with the H-shaped device). A. Fresh medium bottles (we can switch between two media) are kept in the enclosure equipped with a 30°C thermostat. B. A custom made digitally controlled valve can switch inlets between input bottles and connects to the microfluidic chip inlets. C. The H-shaped microfluidic chip mounted on custom chip holder. D. A peristaltic pump pulls liquid out from the bottles in A and flows it towards the trash E.**

In order to change the media dynamically, a valve (Figure 24 B, custom enclosure, valve from The Lee Company) upstream of the chip which has two inlets and one outlet. This valve is controlled by a custom-made, Arduino-based controller which can be programmed in Matlab. This design is easily extendable as in the case of the parallel H-shape chip (which requires 5 valves to run 5 independent experiments) a single Arduino and slight modifications to the driver were enough to parallelize the control without stacking simple controllers.

It is informative to note that a protocol to create a very similar experimental platform (yet using a more expensive and less versatile control board and having a microfluidic setup where cells are growing in the flow channels) was deemed worthy of the cover of *Nature Protocols* in 2015 (97).

### b. Fluorescent probes to peep into cellular activity

Fluorescent proteins (FPs) are definitely the tool of choice when it comes to monitoring gene expression dynamically and *in-vivo* at the single cell level. As any measurement technique, it has pros and cons. In the pros we have that fluorescent reporters are easy to integrate in the genome. These can be inserted in place of a gene or fused to it (at either the N or C terminal[47]). They allow quantifications, although relative (*i.e.* not absolute molecule values but relative changes in values) and do not require any additional chemical or construct[48]. One big advantage of PFs is that they are very popular and therefore a large palette of FP is available and many are way better characterized than most other measurement techniques (98–100). Their use in fact goes beyond traditional imaging (*e.g.* when performing FRAP[49] or considering some FPs can act as molecular timers (101)). At the same time, using FPs leads to several issues, which we will discuss here.

When using FPs as gene expression reporters, it is important to consider that these proteins have their own kinetics. In fact, in chapter III we use an yeast optimized yellow FP (yECitrine (102)). This FP is expressed under the control of an endogenous promoter, pSTL1 and therefore is expected to be transcribed pretty much in the same manner than the original gene (STL1) which was replaced by the FP. This is because transcription is believed to depend mainly on the sequence upstream (*i.e.* promoter region) which in our case was left untouched. With the same kind of argument, the mRNA produced with our FP has the same 5' UTR[50] than endogenous STL1 and therefore, should be translated also in the same way. Nevertheless, since the insertion of an FP in the genome requires a selection marker (in our case the auxotroph marker HIS5, see strain yPH91 in strain table in Annex 1) which is placed after the FP stop codon, the 3' UTR of our exogenous mRNA is different than the wildtype STL1 3'UTR. This can affect the mRNA decay rate (103) which in turns changes mRNA levels and therefore, protein levels. Moreover, although some extensive work has been done to quantify mRNA decay under osmotic stress for nearly all of *S. cerevisiae* genome (77), such genome wide studies never include fluorescent proteins in their scope (which would be very useful in practice). The same remark applies to protein degradation rate although in that case, several approaches to modify natural degradation rates of FPs (which are usually very stable proteins) have been proposed (104, 105). For all these reasons, when we looked for literature values for several parameters of our model of gene expression in chapter III, we took care not to consider our reporter to behave like the gene it replaces (see supplementary information in the article in annex 6).

One important factor to be taken into account when conducting quantitative fluorescent measurements has to do with photobleaching: the fact that fluorescent proteins lose their fluorescence when excited too much. This means in practice that the more fluorescent proteins are imaged, the lower the signal. The precise reasons for photobleaching are poorly understood (106). From that, and considering that bleaching is a random event with constant probability in time (excitation time) the most common model of photobleaching is a Poisson transition to bleach state which leads to an exponential decay of fluorescence intensity upon illumination. In order to

[47] In some specific cases, FP are also inserted within a protein, or in an exon.
[48] Unlike luciferase reporters for example.
[49] Fluorescence Recovery After Photobleaching is a well-known technique that uses FPs to measure diffusion *in vivo*.
[50] UnTranslated Region. Part of mRNA which is transcribed but not translated and appears before the gene (start codon) when reading DNA in the 5' to 3' direction. After the stop codon there is a 3' UTR.

compensate or at least account for bleaching, it is necessary to measure the rate of fluorescence decay due to bleaching. The classic experiment consists in placing cells expressing the FP in similar conditions as the experiment[51] and leave excitation light shining on them while recording the diminution in fluorescence (106).

Here, motivated by the fact that fluorescent proteins can also be subjected to photoblinking (a process similar to bleaching but transitory whereas bleaching is irreversible) (107), we conducted a photobleaching experiment where excitation illumination was intermittent[52] with the same exposure time as for experiments. The average fluorescence over 10 cells is reported in Figure 25. Linear fit on log transformed, background free fluorescence (which is more appropriated since an exponential fit in normal fluorescence give much more weight to the highest points) allowed an estimation of a bleaching rate of **0.0032** per frame[53]. In other words fluorescence intensity decreases by 0,32% for each illumination frame of 200 ms at 50% of illumination power. As it appears in Figure 25, bleaching

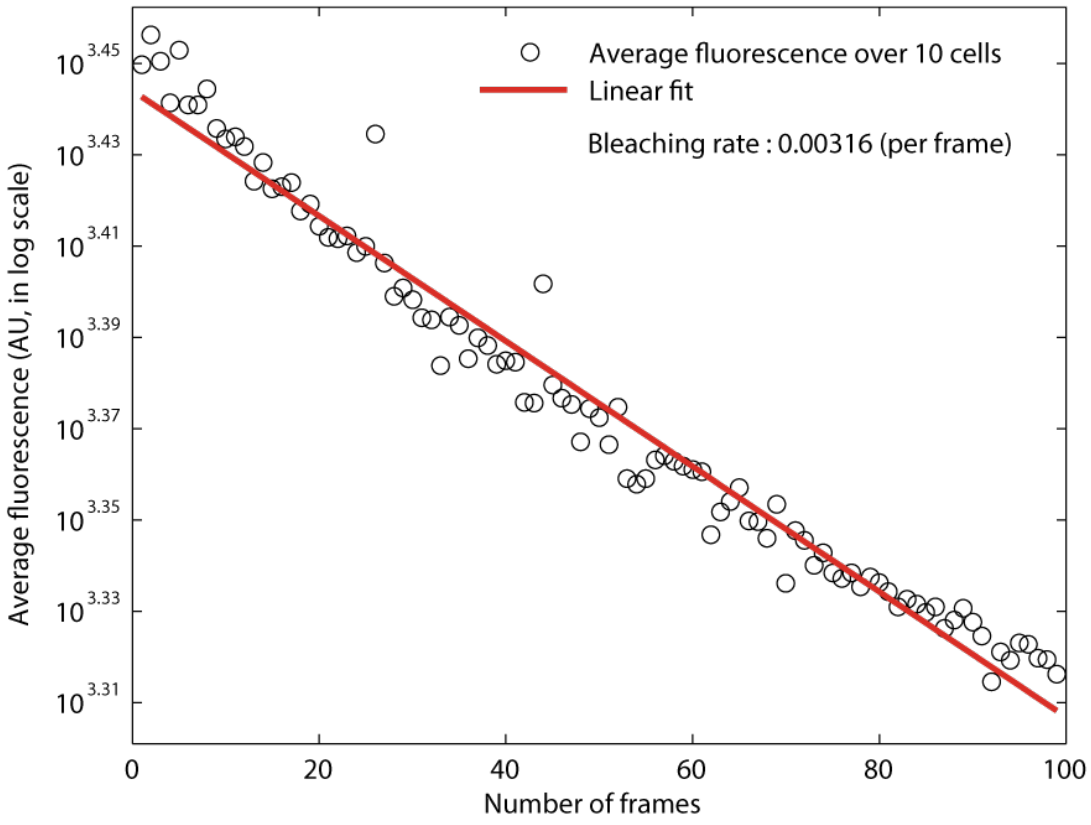



**Figure 25 - Bleaching rate estimation. The average fluorescence of 10 cells decays as an increasing number of fluorescent images are taken. Bleaching experiment was carried following a regular experiment (100913), therefore, having cells from the yPH91 strain with typical YFP fluorescence levels (~4500 AU). Imaging conditions are the same as for experiments of chapter III except here one image was taken every 5 seconds.**

is only imperfectly represented by an exponential decay. This mild deviation from a pure exponential was found to be more pronounced at very low fluorescence level in our system (~300 AU).

[51] In fact, beaching rate can be specific to a given strain in particular experimental conditions.
[52] yECitrine, excited for 200ms every 5s at 50% intensity on an X-Cite 120PC lamp, under a 100X objective PlanApo 1.4 NA, Olympus
[53] This means that with the sampling of 6 min rate used in chapter III, fluorescence decay from photobleaching has a rate of $5.3.10^{-4}$ which is ten times less than protein dilution and is therefore negligible.

A last consideration on the use of FP is phototoxicity. Although detrimental effects for cells can already appear in bright field, the intensity of fluorescence excitation illumination is much higher. Moreover, in order to fluoresce, FPs jump to excited states where they are likely to react and form ROS[54] which can have various harmful effects on cells. Because so many factors related to microscopy can produce toxic effects, it is hard to properly quantify its effects and not much literature is available concerning phototoxicity in *S. cerevisiae*[55]. From a practical point of view, a simple rule is to minimize overall exposition (in terms of sampling rate, exposure time, light intensity and color with short wavelength being more energetic) as much as it does not hurt the Signal-to-Noise ratio or under-samples in time the phenomenon of interest. In this aspect, it is usual to use different sampling rates for bright field images and fluorescent ones. This allows having a high enough bright field sampling rate (so as to obtain good single cell tracking) and sufficient sampling for fluorescence which fluctuates more slowly while minimizing phototoxic effects. At last, measuring the growth rate in absence of perturbation other than imaging itself is a standard control.

[54] Reactive Oxygen Species

[55] A quantitative investigation of the matter in *S. cerevisiae* has been done recently by a research group but publication is not yet available at the time of redaction.

## 2. Image Analysis

In this section, we will discuss image analysis which is a crucial step consisting in automatically extracting relevant and quantitative information from microscopy images.

### a. Segmentation and Tracking using Cell*

When we look at a microscopy images we usually have no problem in distinguishing single cells one from another, and this even when images are of rather poor quality. Working at the single cell level usually requires obtaining a fairly good amount of data in order to derive meaningful statistical values. In a typical experiment, we record from 5 to 20 fields of view for several hours which rapidly amount to thousands of images featuring hundreds of cells. Performing image analysis by hand is therefore a lost cause for most of the experiments we are interested in. When it comes to having algorithm performing that same task of distinguishing cells, things become much more difficult. Basically, the goal is, from bright field images (Figure 26 A) to obtain for each image the pixels which corresponds to each cell (**segmentation**, Figure 26 B) and for two images taken at different time, the knowledge of which cell is which (**tracking** Figure 26 C).

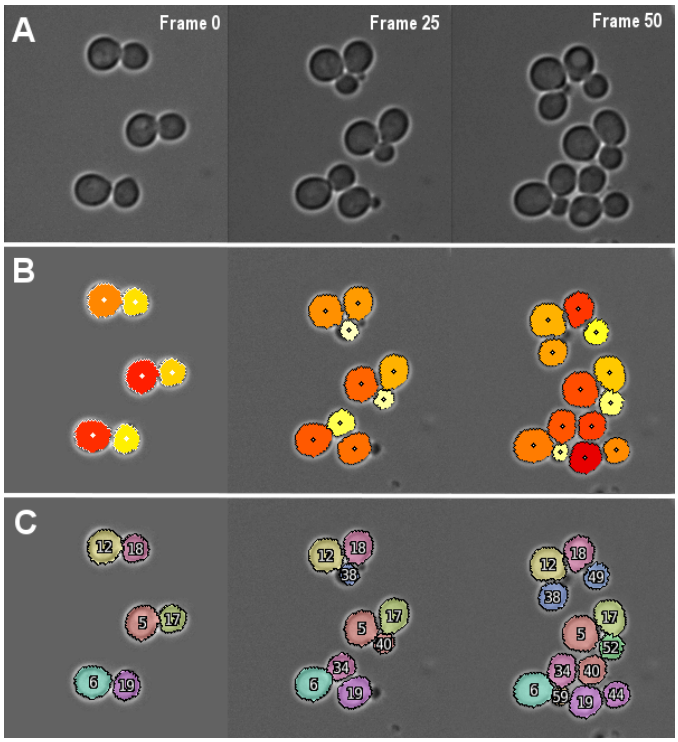



Figure 26 - A. Bright field time-lapse images. B. Segmented images. C. Tracked & segmented images

Importantly, we want to obtain dynamical data for single cells over relatively long periods (several hours). We call *trace* all the segmented pixels in time corresponding to a single cell. It is important to see that a single error in segmentation or tracking in one frame will affect the whole trace for the concerned cell. This leads to what we can call the exponential decay of precision which is depicted in Figure 27. What this figure represents is the fact that having a per-frame accuracy of 95%, which might seems fair at first, is insufficient for long term experiments. In fact, after only 50

frames, less than 10% of traces would still be correct. Because of this accumulation of errors, automated segmentation and tracking (S&T) needs to aim for very high levels of per-frame precision.

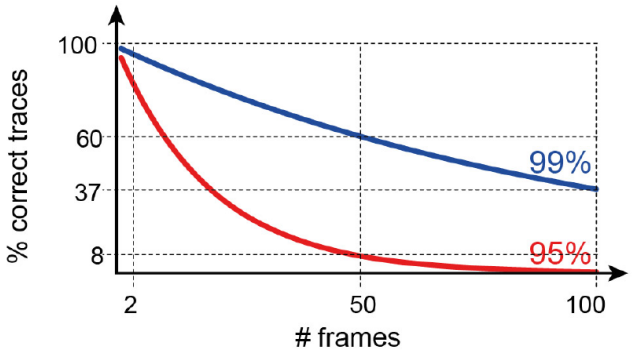



**Figure 27 - Segmentation and tracking errors per frame accumulate over long sequences. Here is plotted the accumulated error (in terms of single cell traces) for two systematic error rates per frame. Figure from** (108)**.**

Realizing that the level of accuracy of available methods was insufficient for long-term S&T, we helped collaborators in designing a new tool: CellStar (Cell*) which is dedicated to long term S&T of *S. cerevisiae* microscopy images. Whereas most budding yeast segmentation tools are based on the generalized Hough transform, in Cell* segmentation is based on active rays, a technique inspired from active contours. Tracking is performed through on a multi-criteria optimization which allows taking into account typical tracking hardships encountered when imaging yeast in microfluidic chambers (see (108)). This allows this tool chain to have a better precision than any other tool it was compared to. In addition, since at some point any per frame precision will eventually lead to a trace being lost in time, Cell* also includes a user interface to correct traces rapidly. The paper describing this new tool is still in preparation (108). In this collaboration, our contributions were the following:

- We produced a set of time-lapse microscopy images displaying several common hardships for S&T (*e.g.* small movement of the microscope stage in between frames, osmotic stress or *jumps* of patches of cells). This allowed the construction by collaborators of an open source benchmark tool allowing quantitative comparison of several available S&T algorithms for budding yeast[56].

- We provided regular and extensive feedback on performance, bugs, user experience in order to produce a relevant tool for the community. In this respect, many interns and other experimentalists were also provided with different alpha or beta versions of the tool to gather larger feedback. Overall, we tested it on tens of movies, both in batch with scripting upon the tool and with the interactive user interface to provide feedback both for advanced use and for beginner use.

- We created part of the documentation which will come with the official release of Cell*. This is in the form of a tutorial, allowing a broad audience of users, some of

[56] Algorithm benchmark results, methodology and open source tool, see http://yeast-image-toolkit.biosim.eu/ (accessed on 15th September 2015)

which may not be very at ease with Matlab or Octave (the two compatible software environments to use the tool) to get stated with the tool.

An important aspect of Cell* has to do with the usually very difficult task of setting a S&T algorithm internal parameters. To the practitioner, such meta-parameters are most of the time abstract and would require time-consuming trial and error to be adjusted for new imaging conditions. In Cell*, a machine learning approach has been implemented which allows a user to provide the algorithm with a few hand curated images which are used to automatically tune Cell* meta-parameters. Interestingly, besides this perks for Cell* robustness and versatility, it appears that with time one finds how to tune microscopes so as to produce images which are the most informative for the algorithm. In the end machine learning and human learning converge to produce higher quality data.

As it will become apparent in the rest of this chapter, S&T is a crucial step in image analysis since many measurements derive from it to some extent. An elementary derived measure is that of single-cell fluorescence: by imaging cells not only in bright-filed but also in fluorescence, we use S&T results to compute single cell fluorescent levels (usually the averaged[57] fluorescence) over time. Co-localization of proteins tagged with different FP is also straightforward[58] when images have been segmented.

### b. Measures of cellular identity

As we investigate single cell variability, we ask the question of what constitutes a single cell's identity. It has been claimed that variability could be related to several aspects of cells physiology and could be influenced by the microenvironment (109). Many putative features which can relate to variability are accessible to quantification from microscopy images. In this study, we quantified for instance local cellular density, cells age, size as well as cell's perception of osmotic stress[59].

Genealogical origin is also expected to contribute importantly to variability because of all the features which are transmitted from a mother cell to its daughter. This includes not only a genome but also epigenetic traits and a given cytoplasmic composition which together carry the *state* of the augmented gene regulatory network of the mother (if we see this as a large dynamic system).

Cell* development plan includes an automated way to extract cell lineages from bright field images only but this option is not yet functional. In addition, such lineage reconstruction from bright field relies on some assumptions (*e.g.* all cells dividing regularly) which can limit its use. For instance, when studying the impact of repeated osmotic stress on cell division, cells do not divide regularly indeed.

[57] Here we refer to the fluorescence intensity averaged over all pixels attributed to a given cell at a given time frame.

[58] Co-localization is usually defined as a ratio of pixels showing both fluorescent colors to pixels only having one. In the case of nuclear localization (of Hog1 for instance) it can be measured as the ratio of the summed fluorescence for Hog1 over pixels corresponding to the nucleus (as identified by another FP) divided by the total summed fluorescence of Hog1 in the cell.

[59] For a description of the quantification methods employed for these features, see the supplementary information of the paper draft given in annex 6

Using yeast strains containing a nuclear tag[60] and with a frequent enough image acquisition it is possible to observe mitosis and in particular anaphase[61] which allows to reconstruct faithfully mother-daughter pairs. Time lapse imaging of such cells is represented in Figure 28.

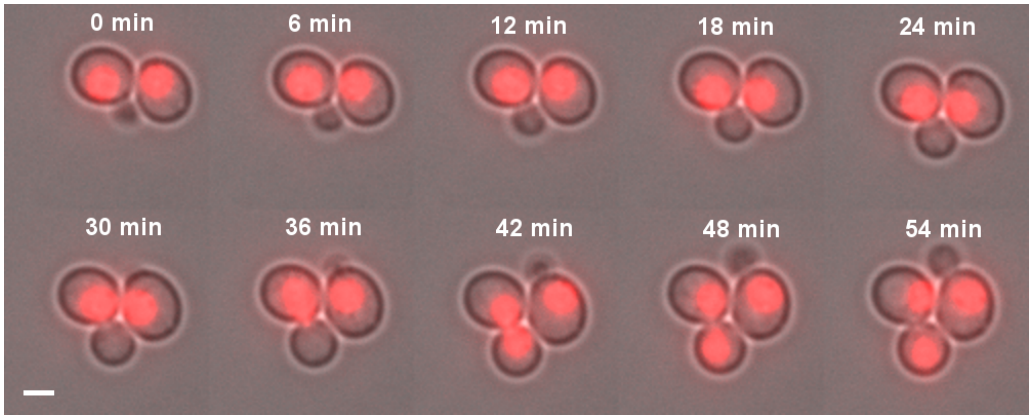



**Figure 28 - Montage of time-lapse showing mitosis for cells having a fluorescent nuclear tag (HTB2-mCherry). Microscopy images at 100X (scale bar is 5µm), overlay of bright-field (grey) and RFP fluorescence (red). Strain used: yPH15 growing in 2% glucose SC medium, experiment 140214.**

We implemented an automatic lineage reconstruction algorithm from microscopy image similar to those of Figure 28. This algorithm performs the segmentation of nuclei (Figure 29 A) using a gradient based approach (Figure 29 C) and matches it with S&T data from Cell* (Figure 29 B). Then,

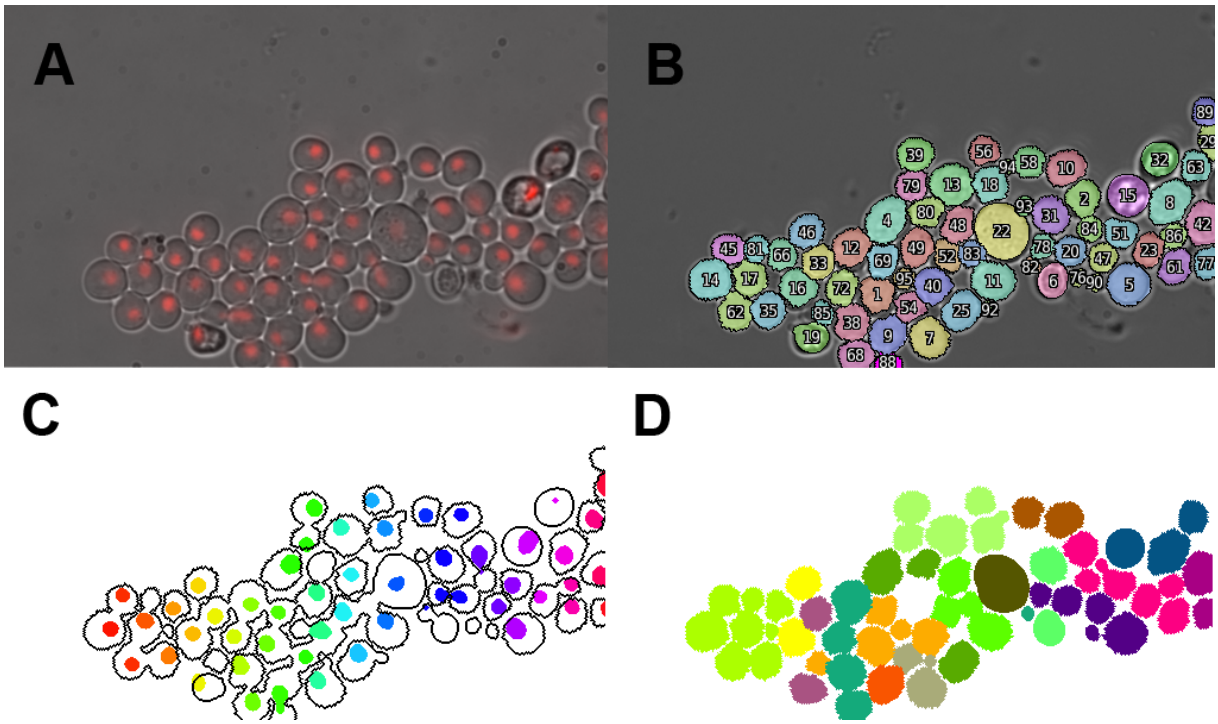



**Figure 29 - Illustration of automated lineage reconstruction from nuclear markers. A. Composite image of bright field at 100X and nuclear marker HTB-mCherry fluorescence. B. Segmentation and Tracking result. C. Segmented nuclei. D. Results of lineage reconstruction. Cells coming from the same initial cell are coded with the same color.**

[60] A nuclear tag is a fluorescent protein which is restricted to the nucleus. This is done usually by fusing this protein to a nuclear endogenous protein. To have a good signal, such a nuclear protein should be expressed in large amounts. A typical fusion choice made here is to use histones proteins, *e.g.* Htb2.

[61] Anaphase is the spatial separation of chromosomes between cells. As this event lasts approximatively 10 min, sampling time should be set accordingly.

mother-daughter relationships are extracted by detecting nuclear separations and genealogy is reconstructed (Figure 29 D). This algorithm requires the sampling rate to be high enough so as to ensure both mother and daughter nuclei will be fused in at least one frame. Yet, this condition is not always fulfilled as nuclei can often split very rapidly.

The performance of this automatic lineage reconstruction was improved by including more features than nuclear merging (11 overall[62]) which are computed systematically for each daughter and potential mother (defined by cells being close enough to the new born cell when it receives a nucleus). In its current form where feature selection and weighting has been done by hand, it already has a true positive rate of 94% and a false positive one of 1,7% (based on 375 potential mother daughter pairs among which 82 correct ones). It can even resolve situations like that shown in Figure 30 where two cells bud simultaneously *crossing* each other. Once stable, the code will be available at: http://github.com/Lab513/YeastImAnalysisToolchain/LineageTool

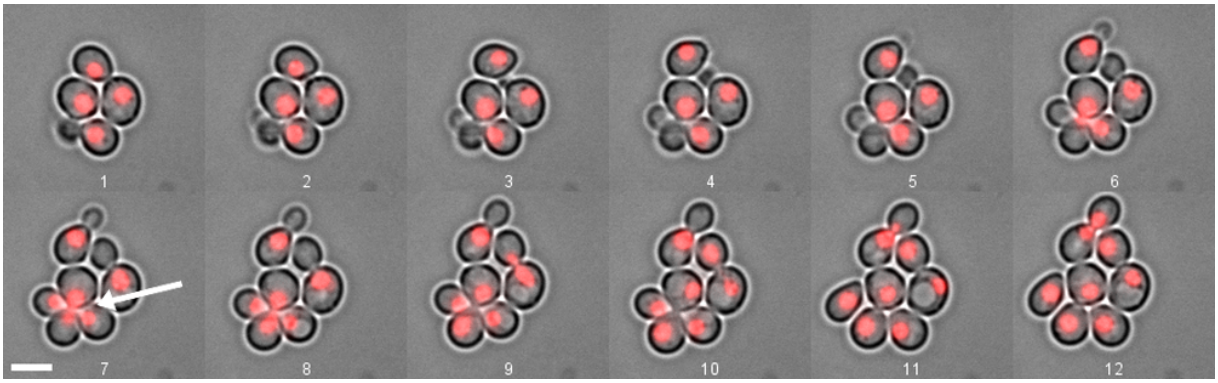



**Figure 30 - Example of lineage reconstruction difficulty as indicated by the white arrow. Numbered frames are separated by 6 min. Strain used: yPH15 growing in 2% glucose SC medium, experiment 140214. Scale bar is 5µm.**

[62] Features include metrics of nuclei shapes, nuclei position relative to cells, nuclei movements. Distance between cells, fluorescence in between cell centroids and variations of metrics thereof.

# 3. Measuring growth in populations and single cells

## a. Going beyond the field of view: An Eulerian measure of population growth

Generally population growth could only be measured in the initial part of our experiments, before the imaging window was full of cells. This is because the typical way to measure growth is by following the number of cells present in time. When the imaging window is full, cells flow in and out and it is no longer possible to measure the number of cells. This imposes a trade-off: estimating finely growth rate of a population in time requires having many cells so budding occur frequently. But the more cells, the faster the imaging region gets filled.

To circumvent this imitation, I designed and implemented a technique allowing the estimation of population growth rate from microscopy images filled with cells (but segmented and tracked). It relies on tracking the flux of cells present in the image. Bulk measurements are performed chamber by chamber. Out of the 512x512 pixels, we follow the number of cells present, entering and leaving a centered "window" of 482x482 pixels as depicted in Figure 31.

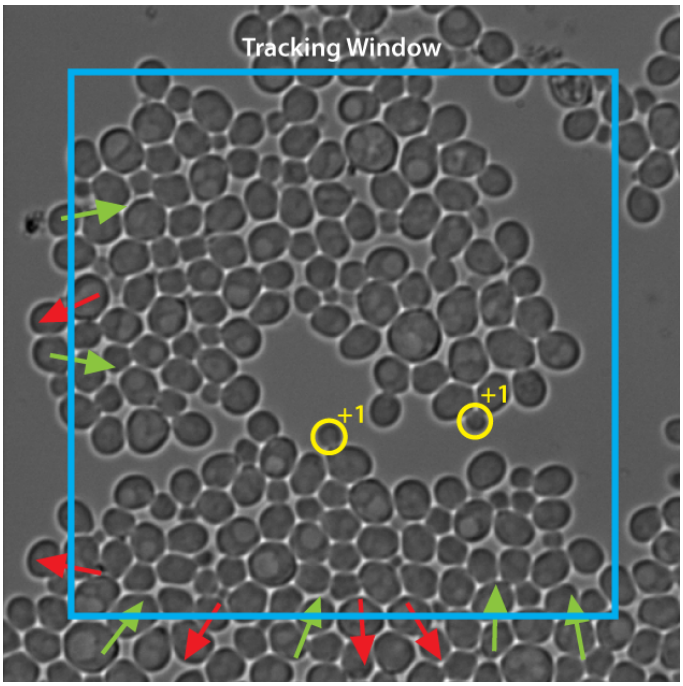



**Figure 31 - Principle of the window growth estimation method. We track cells coming in and out of a reference window (blue square and green and red arrows) along with cells appearing in the window (yellow circles).**

For each frame, we have the relation:

$$N_w(t+1) - N_w(t) = N_{new}(t) + N_{in}(t) - N_{out}(t) \quad (2)$$

where:

- $N_w(t)$ is the number of cells present in the window at frame t
- $N_{in}(t)$ is the number of cells that entered the window between t and t+1
- $N_{out}(t)$ is the number of cells that left the window between t and t+1
- $N_{new}(t)$ is the number of cells that were born in the window between t and t+1

By simply counting the number of cells observed during an experiment (black curve in Figure 32 A), we cannot estimate the global division rate as soon as cells leave the imaging field (Figure 32 A, dark blue curve). Retrieving $N_{new}(t)$ allows computing a first estimation of the population growth by simply considering $N_w(0) + \sum_{s=1}^{t} N_{new}(s)$ (Figure 32 A, green curve). This estimator accounts for cells born in the window and which have left. Yet it only includes cells which are born within the field of view. As a consequence, it does not take into account the contribution to population growth of a cell which has left the observation window. Accordingly, it cannot be compared to the traditional model of population growth (where the division rate is defined) in which population growth is

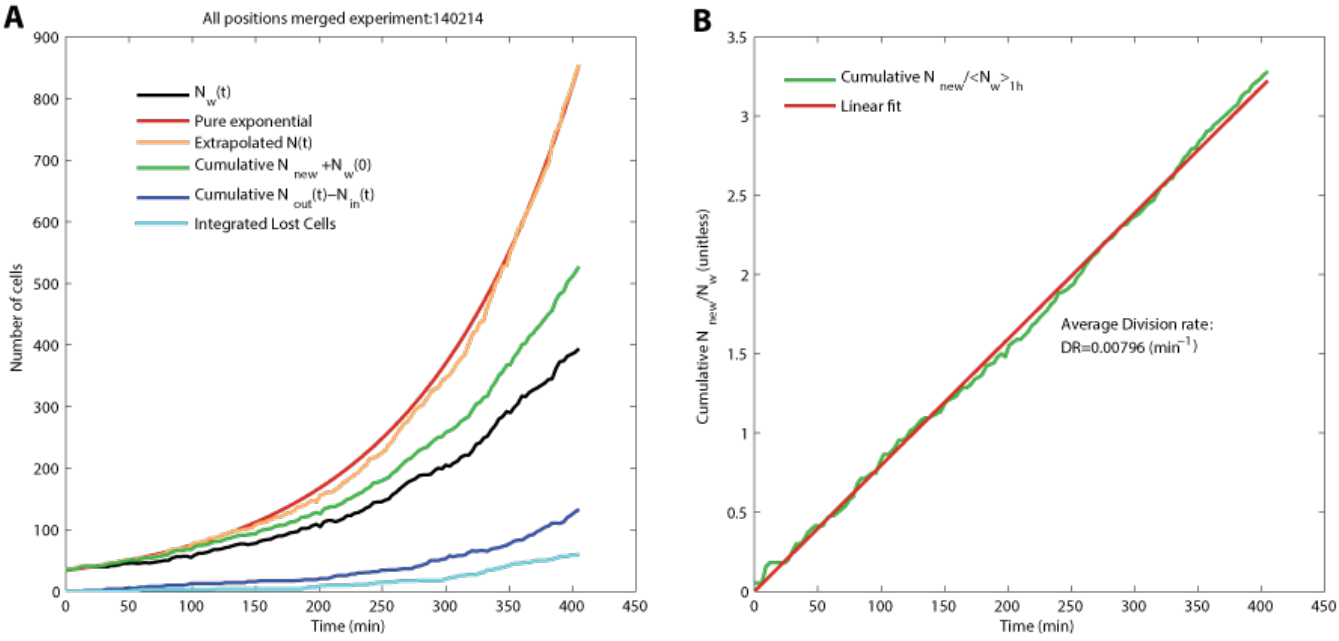



**Figure 32 - Results of the window growth estimation on experiment 140214 (pooled data from 5 positions). In this experiment, the strain yPH15 grows in 2% glucose SC medium in absence of stress.**

unconstrained.

From this relation, we estimate $\frac{N_{new}(t)}{N_w(t)}$ which is the relative number of newborns in the window between t and t+1. Although fluctuations in $N_w(t)$ are mild, and the results are pretty similar, we deemed more pertinent[63] to use as denominator the mean number of cells in the window during the last hour (or during one average division time), *i.e.* $< N_w(s) >_{s=t-1h\ldots t}$.

To obtain an equivalent population division rate we estimate the instantaneous division rate as $\frac{N_{new}(t)}{<N_w(t)>.\Delta t}$ (with $\Delta t$ being the time interval between two frames). This instantaneous division rate can then be extrapolated which yields the orange curve in Figure 32 A.

A linear fit on the cumulative sum of $\frac{N_{new}(t)}{<N_w(t)>}$ (Figure 32 B, green curve) allows measuring directly the average division rate in the window (given by the slope of the red linear fit in Figure 32

[63] If we consider that some cells might enter or leave the window at any time, using $N_w(t)$ may lead to estimation errors coming from either accounting incoming cells or excluding cells which have just left in the "cell production pool".

B). Simulating a pure exponential population growth at the average rate yields the red curve in Figure 32 A.

In practice, this computation is done chamber by chamber as a sanity check and we use data pooled over all the chambers to compute the overall division rate for a given experiment (as it is the case in Figure 32).

We can note that the current method is still sensitive to fast flow (fast being when cells come in and out of the window in less than an average division time). This could be improved marginally with finer tracking of single cell's residency in the window but would become more sensitive to eventual segmentation and tracking errors.

At last, although the example presented here concerns an experiment where cells were growing in the absence of stress, the proposed approach has been also applied to experiments with osmotic stress.

### b. Measuring growth at the single cell level

It is possible to extract division rates at the population level from microscopy images. Yet, division rate is also variable within a population. When nuclear markers are available and lineages are reconstructed, single cell division times are naturally accessible. Nevertheless among other practical reasons, given the limited number of fluorescent marker which can be included at the same time in a strain, measuring single cell division rates without nuclear markers is useful.

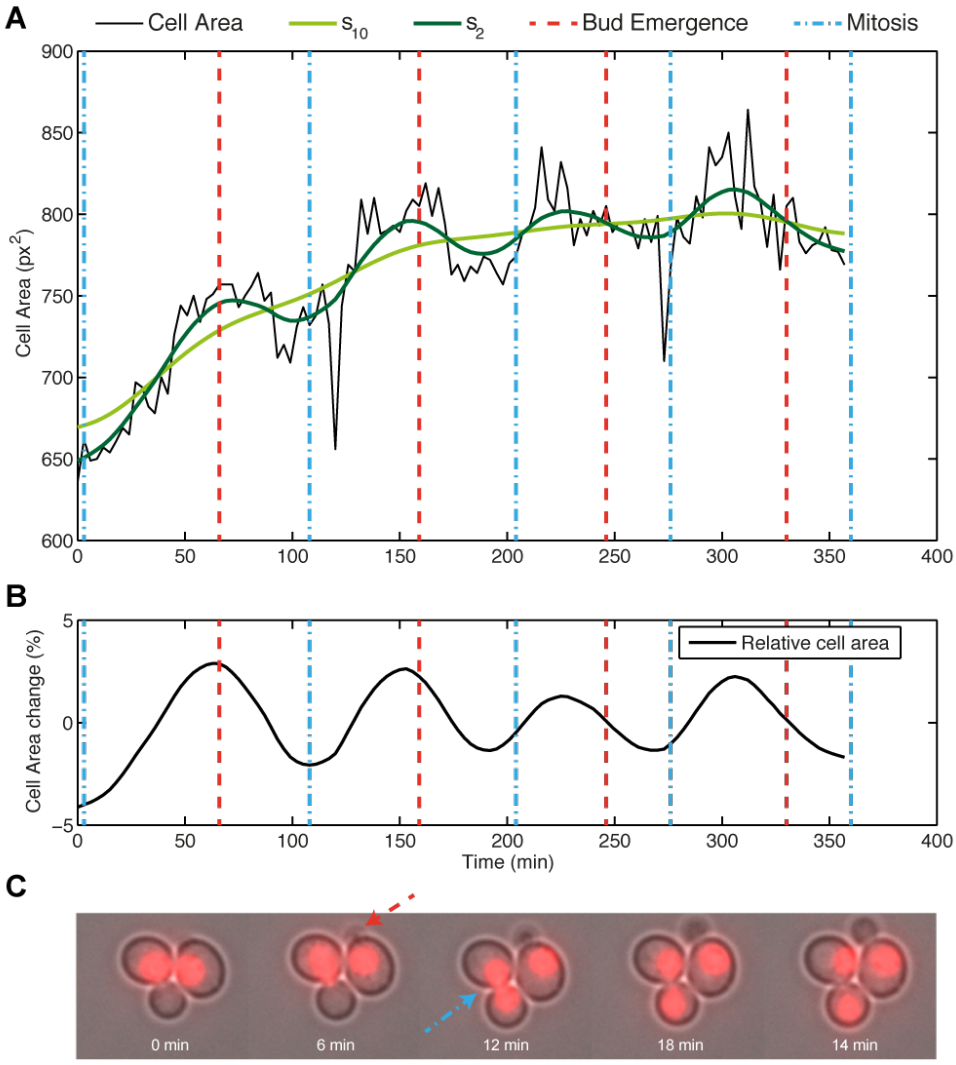



**Figure 33 - Fine oscillations in cell size are related to cell cycle. A. Subtle oscillations in cell size are measured through segmented images. The black curve is raw data, dark green is a mildly smoothed version and the light green curve is an even smoother version. B. The black curve represents the relative variation of the smoothed cell size (dark green curve in A) over the global cell size trend (light green curve in A. C. We Extracted by hand bud emergence events (red arrow) and nuclear separation events (blue arrow). In A. and B. these events are represented by red and blue vertical dashed lines respectively. Example of cell #1 from 140214 (no osmotic stress)**

When segmentation is accurate enough, we can observe subtle oscillations in cell size (see black and dark green curves in Figure 33 A) which we found to be related to bud emergence[64] and

[64] This corresponds to the time at which a new bud is visible and marks the G1/S transition, see red arrow in Figure 33 C.

mitosis[65] (red and blue dotted vertical lines in Figure 33). Note that the example given here comes from an experiment without osmotic shocks. Yet, the presented method can be used also when short stresses are applied, provided data points corresponding to stress are removed.

In order to use this size effect to infer single cell division rate, we need to filter out both fast variations coming from pure imaging and segmentation noise, and slow variation of cell size. To remove fast variations while keeping variation due to cell cycle, we applied a simple iterative averaging window smoothing[66] twice which yields a signal, $s_2(t)$ (dark green curve Figure 33 A). To obtain a more global trend in cell size we apply the same window smoothing 10 times, which yields $s_{10}(t)$ (light green curve Figure 33 A). We then consider the fluctuation of $s_2(t)$ relative to $s_{10}(t)$, *i.e.* $\frac{s_2(t)-s_{10}(t)}{s_{10}(t)}$ (Figure 33 B). In order to extract frequency information from this relative size change, we perform a Fourier transform to compute the power spectrum and retrieve the average division frequency[67].

This approach has been also used on experiments in which repeated osmotic shocks were applied (typically experiments form chapter III). Despite the fact that osmotic shocks lead to brutal drops in cell size, this method was still able to infer single cell division rate. A graphical example is given in supplementary materials and methods of the article in annex 6. In a data set featuring osmotic shocks, this approach was manually validated on fifty cells yielding an average error (compared to manual bud appearance based doubling rate[68]) on the mean doubling rate of 12%. Note also that the particular use of $s_2(t)$ and $s_{10}(t)$ here is not critical, using $s_1(t)$ or $s_{11}(t)$ would give similar results. These were simply determined by hand on a few examples. The filtering part of this method could probably be improved, for example by using directly Gaussian filters with different windows.

[65] We report times where nuclei divides, corresponding to G2/M transition observed with a nuclear marker. See blue arrow in Figure 33 C.

[66] Here we use a (11 frames *i.e.*33 min) centered window with left and right padding. Therefore $s_N(t)$ is the average of $s_{N-1}(t-5)$, $s_{N-1}(t-4)$, …, $s_{N-1}(t+5)$. Note that the iterative application of a simple smoothing window converges toward the application of a Gaussian filter.

[67] It would be possible to compute the average division time by simply taking maxima or minima of the oscillation but this might be more sensitive to sampling time respective to the proposed method. To compute the global division rate we compute the power averaged frequency for frequencies having a period between 60 and 400 min, conservatively including possible doubling times for yeast.

[68] The precision of estimating division rate from bud emergence being itself limited by visibility of buds and imaging frequency.

## 4. Conclusions on: long term dynamic experiments and single cell data

In this chapter, key experimental elements used in our study have been presented. In the introduction chapter, a short review of measurement techniques applicable to the single-cell level in yeast was proposed. This motivated the choice of microscopy as a main source of measurements for its capability of long-term acquisition of single-cell longitudinal data. The use of custom microfluidic chips allows improved imaging and homogeneous culture conditions along with a precise temporal control of cellular environment, thanks to custom made hardware and software.

The crucial aspect of image analysis to retrieve information from long-term microscopy experiments was stressed out. The initial step in image analysis is segmentation and tracking for which we used Cell*, a new tool with superior performance. Our contribution to its development was mentioned and several original image analysis methods which can be subsequently performed were presented. These notably included the inference of genealogy in a cellular population using nuclear markers as well as measures of growth in population and in single cells.

In the following chapters III and IV, we will see how these single-cell measurements can help to assess quantitatively cellular variability in the response to osmotic stress and provide some characterization of the cellular context affecting osmo-induced gene expression and cellular proliferation in particular.

# III. Individuality in the transcriptional response to osmotic stress

In this chapter we report our investigations of cellular non-genetic individuality in gene expression. The work presented here is for a large part reproduced in an article currently under review in *PLoS Computational Biology* and available in Annex 6. Here we are interested in lasting differences between cells regarding their gene expression features. By features, we mean not only the level of expression but rather parameters of gene expression dynamics at the single cell level. Such parameters usually represent several biochemical processes within global *rates* (*e.g.* transcription rate, translation rate, maturation rate etc.). These features are useful in describing gene expression as a dynamic process so when we ask *"When are genes expressed?"* and *"What matters in gene expression?"* and not only *"How much genes are expressed?"*. More precisely, we wonder to what extent we can capture cell-to-cell variability with single-cell models of gene expression where each cell has its specific behavior. An important question is also to what extent it is possible to represent a population while taking into account cell-to-cell variability.

## 1. Modelling dynamics of gene expression at the single cell level

### a. pSTL1 as a reporter of HOG transcriptional response

Here, we are interested in the dynamics of gene expression at the single cell-level in response to osmotic stress. We will focus more precisely on the stable differences between cells regarding gene expression dynamics. To measure gene expression induced by Hog1, we use a strain where the STL1 gene (which normally encodes for a glycerol/$H^+$ symporter) was replaced by a yellow fluorescent reporter.

***STL1 is expressed specifically in hyper osmotic conditions***

As it was presented in I.4.a, upon hyperosmotic stress, the MAPK Hog1 is phosphorylated and quickly translocates into the nucleus where it alters the expression level of hundreds of genes. This major remodeling of gene expression serves several purposes: Several genes are directly involved in the response to osmotic stress specifically. Also, a large subset of osmo-induced genes provides a rather general protection to stressful conditions (they are part of the ESR[69]) and code for chaperones, or enzymes producing protecting molecules like trehalose. At last, a global repression of many genes involved in protein synthesis, along with a global destabilization of mRNAs frees up resources (like RNA Pol II or ribosomes) for the synthesis of osmo-induced genes.

About 80% of osmo-responsive genes are Hog1 dependent (62). Although all the molecular details of how all these gene are affected by Hog1 is not known[70], it is clear that Hog1 affects many gene in association with several co-factors such as the transcription activators Hot1, Smp1, Msn1, Msn2, Msn4 and the transcription repressor Sko1 (see Figure 34). Each of these transcription factors

[69] The Environmental Stress Response is a set of genes activated in many different types of stress.
[70] It should be reminded here that Hog1 is known to affect chromatin remodeling factors such as the SAGA or SWI/SNF complexes.

(TF) has specific targets, for instance, genes of the ESR are targeted by Msn2/Msn4 and Msn1 is related to DNA replication stress. Hot1 and Sko1 have less targets (~70) but these are much more specific to osmotic stress (110). In fact, although the targets of these various TF acting with Hog1 can sometimes overlap, these regulations will sometimes be effective in different conditions. For example, STL1 which is induced by Hot1 in exponential growth seems to be activated by Smp1 during the stationary phase (111). Also, some targets of Msn2/Msn4 and Hot1/Sko1 overlap but when an hyperosmotic stress is due to extracellular glucose, Msn2 and Msn4 are not activated (110).

***STL1 is a standard reporter of the HOG pathway transcriptional response***

STL1 belongs to the category of genes which are expressed specifically in hyper-osmotic stress conditions. Its activation is mediated by Hot1 which also activates the expression of GPD1 (110).

Production of glycerol (which is the principal biocompatible osmolyte) is the main mechanism for adaptation to hyperosmotic stress (*i.e.* recovery in size and water activity). Two pairs of paralog enzymes, Gpd1/Gpd2 and Gpp1/Gpp2 are essential for the synthesis of glycerol from common metabolites and the genes GPD1, GPP1 and GPP2 are up-regulated under osmotic stress (112). It was reported that Gpp1/Gpp2 is not rate limiting in glycerol synthesis (113) and therefore, the expression of GPD1/GPD2 has a more direct impact on cell's adaptation capability. GPD1 plays a major role under aerobic conditions while GPD2 is required and produced in absence of oxygen (113). In addition unlike GPD2, GPD1 is induced by Hog1.

Gpd1 plays a major role in specific adaptation to hyper osmolarity. Yet, although it is possible to produce functional fusions of Gpd1 with fluorescent proteins, such reporter imposes several practical limitations. As glycerol production is necessary in normal conditions, Gpd1 is expressed already at significant levels prior to any osmotic stress. Therefore, its expression is enhanced but not conditioned to hyperosmotic stress which makes it an unspecific reporter of the HOG pathway. In addition, when it is not used, this enzyme accumulates in peroxisomes which makes it difficult to quantify its level properly. On the other hand, STL1 is among the most strongly induced genes upon osmotic stress. Also, in contrast with the central role of Gpd1 in osmotic stress response, STL1 encodes a membrane transporter which can actively import glycerol from the external medium. Therefore deleting STL1 has very limited consequences on cells survival to osmotic stress and in our case, ensures some increased independence between cells regarding adaptation (since cells cannot import glycerol which could have been produced by their neighbors).

For all these reasons, STL1 is one of the most popular reporters of the HOG pathway's transcriptional activity and accordingly, experimental results models and parameters about STL1 are readily available in the literature.

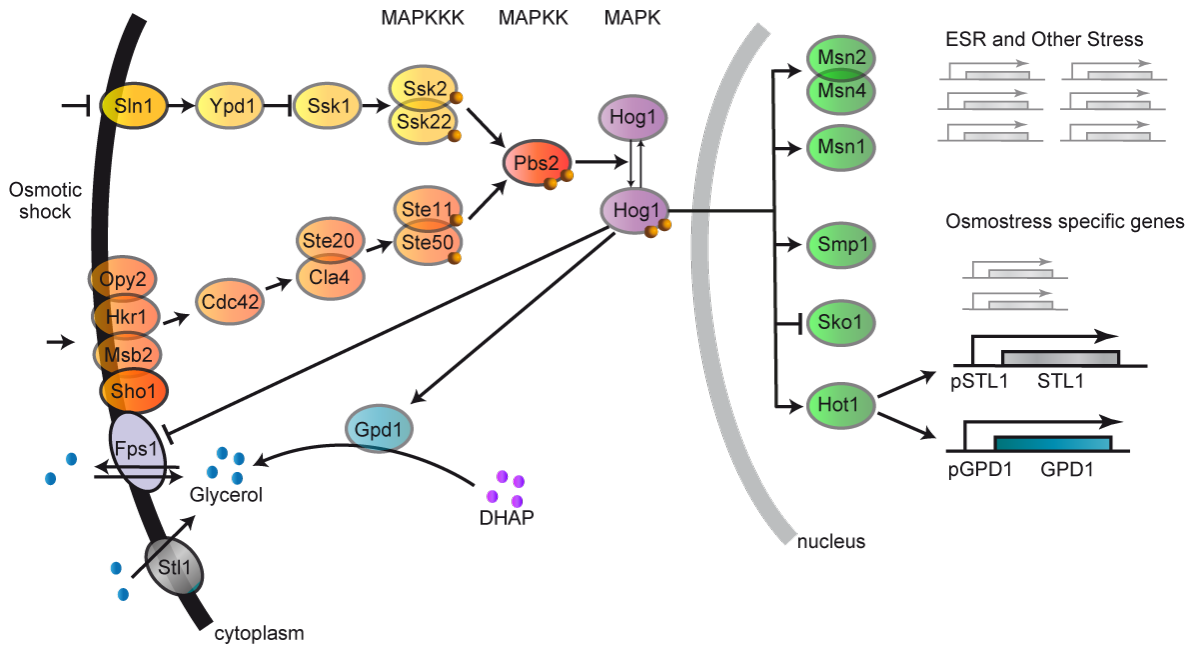



**Figure 34 - Representation of the HOG Pathway including osmosensors, signaling cascade, major transcription factors working in association with Hog1 and the main feedback mechanism due to glycerol production. Although it is not considered to be a major feedback mechanism, the possible import of extracellular glycerol by the Stl1 transporter is also represented. At last, the action of phosphatases on phosphorylated Hog1 is not represented while it is essential for perfect adaptation.**

### *Pulses of osmolarity avoid the effects of negative feedback coming from adaptation*

It is clear that the transcriptional response is not mandatory for cells to accumulate glycerol as glycerol producing enzymes are constitutively expressed and their activity is increased under hyperosmotic conditions regardless of transcription (112). The transcriptional response ensures a protection from possible harmful effects of osmotic stress and helps adapting to very severe osmolarity changes. Also, it can prepare the cell for subsequent shocks. When applying medium stress (1M sorbitol), cells will adapt within 15 to 30 min. Once adapted, Hog1 is dephosphorylated and leaves the nucleus which ends its transcriptional effect. In this respect, the cytosolic activity of the HOG pathway acts as a negative feedback for its transcriptional activity. When interested in transcriptional dynamics, this negative feedback through adaptation makes it difficult to quantify properly gene expression features. Indeed, the actual output would reflect not only gene expression but also adaptation through glycerol production.

In order to disentangle transcriptional features with adaptation, we use the same trick as in (86). By applying 8 min pulses of hyper osmotic medium (1M sorbitol) we stay in the region where adaptation has not progressed enough to modify Hog1 nuclear localization. This is illustrated in Figure 35 where we show Hog1 nuclear localization in response to a 1M sorbitol upshift[71]. Performing

[71] Data shown in Figure 35 corresponds to experiment 250915. We used a yPH15 strain having a nuclear marker (HTB2-mCherry) and a fusion of Hog1-GFP in a microfluidic device. We flow SC medium with 2% glucose and at time 0 the same medium with 1M sorbitol in addition. ~30 cells were used and imaged every 2 min to compute the curves. The relative enrichment corresponds to the relative difference between nuclear average GFP fluorescence and cytoplasmic average GFP fluorescence.

an 8 min shock avoids adaptation which is visible in sustained stress and which leads to Hog1 nuclear export (See dark blue bar in Figure 35). In addition, we impose pulses to be separated by at least 30 min[72] in order to make sure the HOG pathway has been completely deactivated before a new stimulation is made[73]. In fact, the typical period with which Hog1 can shuttle in and out from the nucleus while faithfully following external osmolarity and displaying complete deactivation is larger than 16 min (55). By stimulating the HOG pathway shortly and repeatedly we can experimentally enhance the signal corresponding to the transcriptional response while neglecting adaptation through glycerol production.

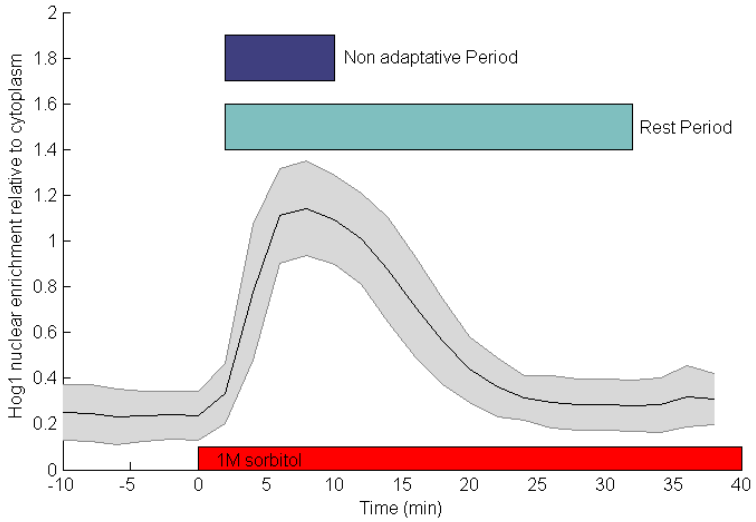



**Figure 35 - Average and standard deviation of Hog1 nuclear enrichment following an osmotic upshift applied at t=0 (red bar). The dark blue bar indicates the 8 min period used in our experiments. The light blue bar indicates the minimum resting time between consecutive stimulations. See footnote 71 for experimental details.**

By stimulating the HOG pathway in this manner, and using pSTL1 expression as a reporter of gene expression in response to osmotic stress, we can in fact simplify the global view of the HOG pathway given Figure 34 to keep only the components which should matter in our experiments as represented in Figure 36. By designing dynamic experiment, we can simplify the HOG signaling cascade by removing (or at least neglecting) the feedback coming from adaptation. As it will be exposed in subsection III.2.a, this will allow us to abstract out all the HOG signaling cascade with a very simple, deterministic model.

[72] *i.e.* we have 30 min between the start of two pulses, the minimal period of normal osmolarity is therefore 22 min.

[73] It should be noted that switching cells back to normal osmolarity before complete adaptation leads to a very rapid cell size and turgor pressure restoration. Accordingly, The HOG pathway will shut down rapidly. In addition, any newly produced glycerol in the cytoplasm will leak out of the cells rapidly to ensure osmotic balance. Therefore, this resting period is a precaution rather than a hard bound.

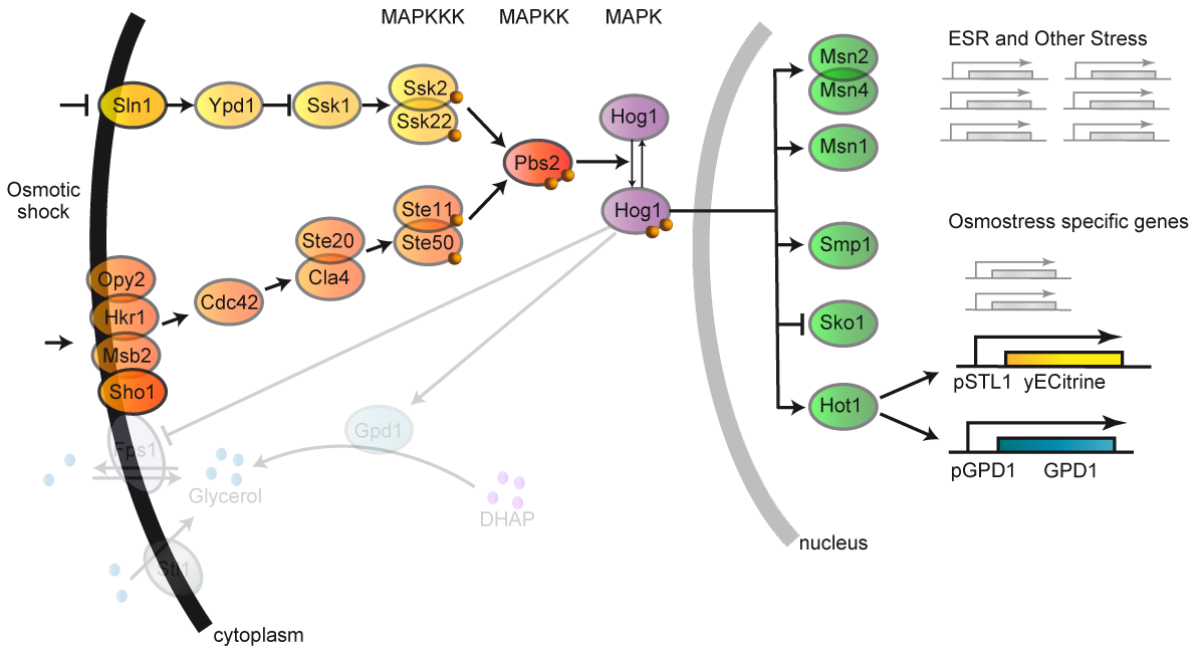



**Figure 36 - Subset of the HOG pathway which is relevant to the presented study. From the more complete picture of the HOG pathway given in Figure 34, we here show the effective impact of using only short pulses of osmolarity along with replacing the STL1 endogenous gene with a yECitrine fluorescent protein gene.**

***Variability in pSTL1 expression***

Cellular systems and gene expression in particular are subject to distinct *flavors* of variability. In Figure 37 we show yeast cells (yPH91) expressing yECitrine under the control of the pSTL1 promoter. We witness a large amount of cell-to-cell variability.

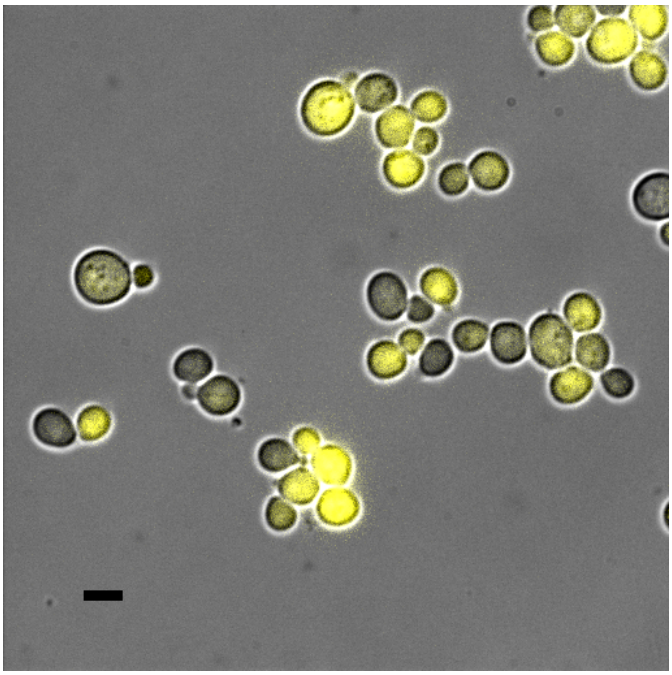


**Figure 37 - Color enhanced microscopy image of cells bearing an pSTL1-yECitrine construct (yPH91). Image (100x) taken from experiment 040813, 30 min after their first 8 min stress pulse (1M sorbitol). Scale bar is 5 µm.**

A common distinction when it comes to gene expression variability is the aforementioned difference between intrinsic and extrinsic noise[74]. In the case of brief pulses of hyperosmotic stress (unless otherwise stated, we always mean as induced by 1M sorbitol), fluorescence levels for pSTL1-yECitrine is shown in Figure 38. We can see that there is an important cell-to-cell variability in gene expression as depicted by the grey area. In addition, when looking at single cell traces, we can also notice that for a given cell, the response to individual pulses can be very different. This random cellular response for distant shocks is reminiscent of intrinsic noise at the STL1 promoter.

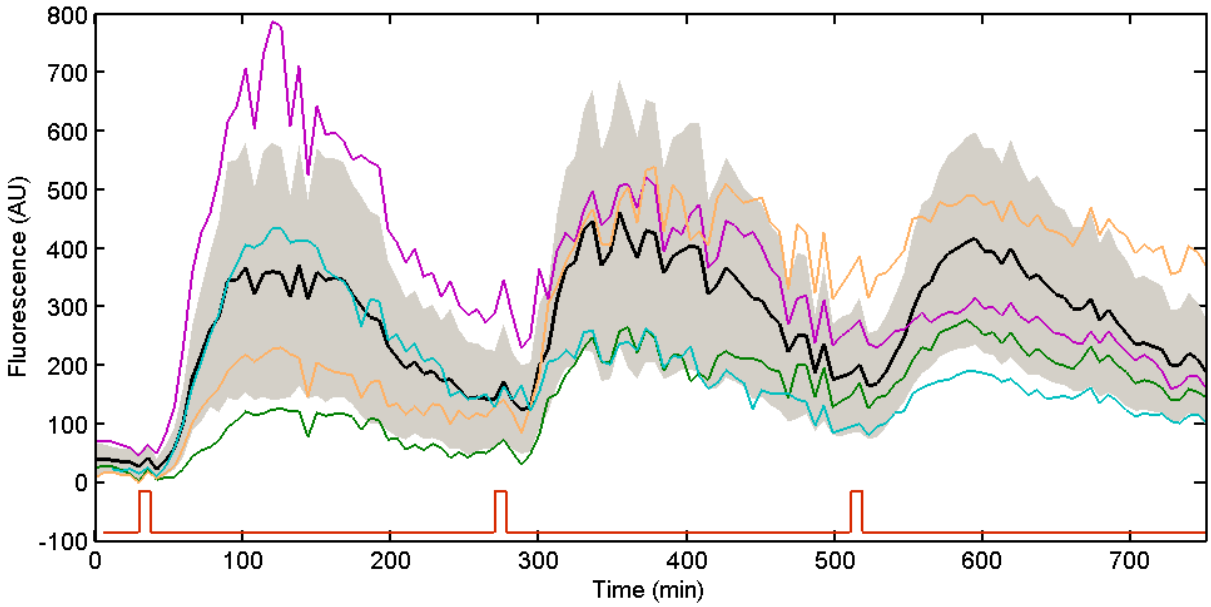



**Figure 38 - pSTL1-yECitrine fluorescence in response to spaced 8 min osmotic pulses (red line). Solid black line is the mean value; grey shaded area represents one standard deviation in each direction. Colored solid lines are single cells fluorescence. Experiment 221012.**

### b. Representing variability in pSTL1 gene expression with stochastic models

Within systems biology an important modelling effort has been dedicated to designing, simulating and estimating (*i.e.* fitting) stochastic models of gene expression. A general and powerful mathematical formalism allowing formulating chemical reactions as stochastic processes is the famous Chemical Master Equation (CME[75]). In such formalism, the exact number of molecules of each reactant is accounted for as an integer value and chemical reactions are defined in a probabilistic manner. In practice it means defining an equation for the time evolution of the probability of the system to be in a given state[76] $\boldsymbol{x}$ at a given moment $t$ knowing its initial state $\boldsymbol{x_0}$ at $t_0$. The equation giving the time evolution of $P(\boldsymbol{x},t|\boldsymbol{x_0},t_0)$ depends on the stoichiometry of the

[74] We use here the typical terminology where *intrinsic* refers to a gene or promoter and not to a cell. See I.1 for a detailed discussion.
[75] A rigorous derivation from chemical mechanics is provided in the landmark paper from Gillespie (167).
[76] Here, the *state* of the system is the precise number of each type of molecule involved in any reaction. Therefore, it is usually an integer-valued vector **x**.

reactions along with how the elementary probabilities of each reaction is modelled[77]. Here we present different studies where this approach was employed on the system we are interested in.

Stochasticity in Hog1-induced gene expression and in pSTL1 in particular was mainly studied experimentally by Pelet *et al.* in (88), where it was shown that for very mild stress (0.1M NaCl, so around 0.2M sorbitol) stochasticity in gene expression could lead to bi-modal distribution of expression (some cells expressing pSTL1 and others not). In addition, a dual reporter assay found intrinsic noise to be dominant for pSTL1 at this mild level of induction and still contributing to around 50% of total noise at higher levels (0.4M NaCl). It should be noted that this stochastic behavior was not due to differences in HOG signaling (as quantified by Hog1 nuclear enrichment in time). Instead, ChIP[78] and mutant experiments were rather indicating that pSTL1 was undergoing remodeling of its chromatin state and that this process, rather than binding events from Hog1 and Hot1, was responsible for the noisy expression pattern (*i.e.* pSTL1 has a *bursty* type of noise).

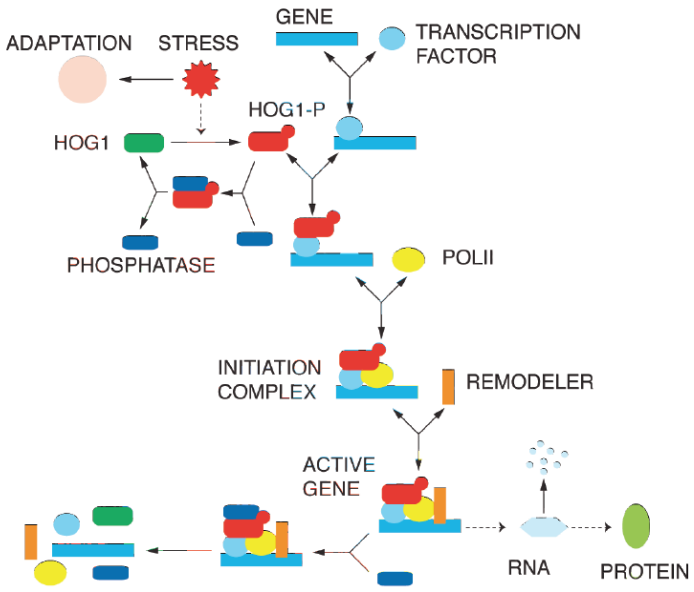



**Figure 39 - Sketch of the stochastic model representing pSTL1 stochastic behavior from (88).**

From this set of experimental results, the authors proposed a stochastic model depicted in Figure 39 which could capture some experimental features. As we can see, this model abstracts out signaling and adaptation and represents in a more detailed fashion processes acting at the promoter level and leading to transcription. In this model, all cells were considered to have the same dynamical parameters although reactions being stochastic, their state originated from different realization of this same stochastic process. Despite the fact that this model could account for several experimental findings qualitatively, it relied on many parameters which cannot be measured and which were set

[77] Most of the time elementary reactions are modeled as exponential jumping times. The propensity for a given reaction increase linearly with reactant abundance for monoreactant reactions. and as the product of reactants' abundances for a multi reactant reactions. Under the classic assumption than reaction rates do not depend on time, the CME defines a Markovian process. Relaxation of this assumption can leads to semi-Markovian processes it rates are time dependent.

[78] Chromatin Immuno-Precipitation experiments allow to measure the proteins bound to a specific portion of DNA.

by hand using both reasonable assumptions and probably some trial and error. It can be expected that this model would display a significant amount of non-identifiability if its parameters were estimated from data alone. In this respect, this model was more useful in demonstrating that the hypothesized mechanisms can indeed produce part of the observed behavior (*e.g*. bi-modality at low induction) rather than in proposing a quantitatively predictive tool.

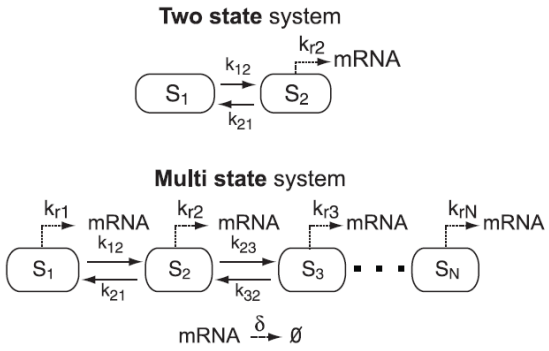



**Figure 40 - Schematic representation of a basic random telegraph model of gene expression (top) and of the generalized random telegraph model (bottom) selected to represent STL1 expression in** (89)**. Note that transition rates can depend on Hog1 nuclear abundance Figure from** (89)**.**

Several subsequent studies tried to propose stochastic models which could be inferred from data in order to be more predictive quantitatively speaking. In (89), single cell data of STL1 mRNAs was acquired by FISH. This allowed the measurement of distributions of mRNA abundance within a cell population at different time points. The authors took a more abstract view to model stochastic gene expression by considering a generalized random telegraph model (GRTM)[79] with various possible states for the pSTL1 promoter. They used a cross-validation approach to conduct model selection. This allowed selecting a GRTM of a specific size by operating a trade-off between improved fit (which is naturally provided with larger models) and minimizing over-fitting (which leads to non-identifiable parameters and lowered prediction capability). This computationally extensive method indicated that the four states GRTM depicted in Figure 40 (bottom) provided the best compromise. In addition, a model selection refinement determined that having Hog1 nuclear abundance impacting the transition $k_{21}$ alone was sufficient, yielding a model with 13 parameters. Therefore, this model has a fairly high number of parameters which may lead to non-identifiability issues. Arguably, since fitting was performed on the full population distribution of mRNA abundances and at several time points, the data used for inference is way richer than simply mean and variance in the population. Nevertheless, consideration of experimental replicability and measurement errors make it difficult to assess fully experimental distributions robustness.

[79] The random telegraph model (RTM) is a classic model of gene expression in bursts. In this model, the gene can be in several states (OFF and ON in basic RTM, see Figure 40 top) which have defined transition propensities which in this context depend upon Hog1 nuclear abundance and each state can produce mRNA with different propensities (In basic RTM, the OFF state does not allow mRNA production while the ON state does). In its generalized form, more than two states are possible (but transition is sequential, *i.e.* transitions are only possible between states I and i+1) and one or more of them can produce mRNAs with possibly different rates.

In both models for STL1 expression presented until now, all cells have always been considered as different realizations of a same stochastic process. In addition, the total variability in transcription was accounted by stochastic effects at the promoter level alone. Nevertheless, both in theory and as measured in (88), variability in STL1 abundance should include both an intrinsic and an extrinsic noise component. Building stochastic models where every cell has the same parameters usually amounts to neglecting extrinsic variability[80] . When inferring parameter values by fitting total variability, it somehow *forces* a stochastic model to represent what it is not supposed to capture.

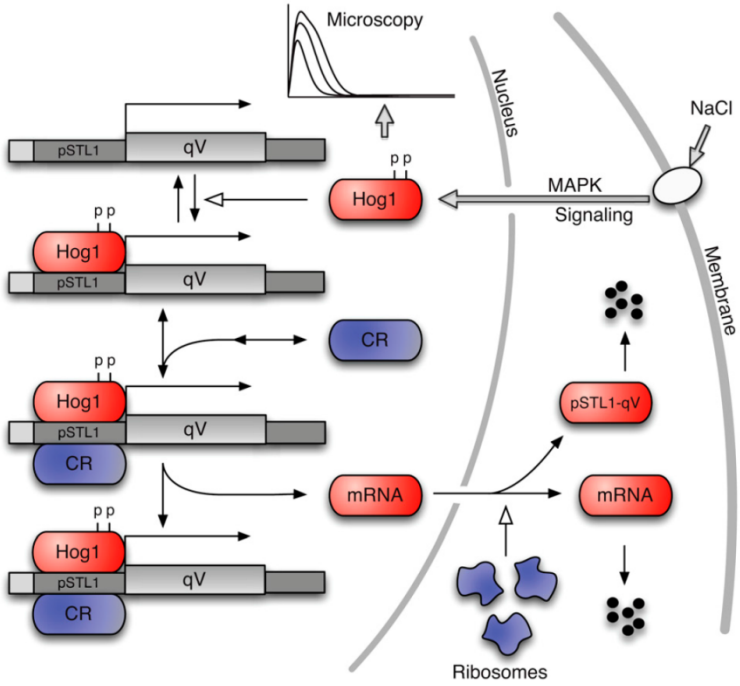



**Figure 41 - Schematic representation of the stochastic model of gene expression of pSTL1 including extrinsic variability (elements in blue can vary between cells). Figure from** (87)**.**

In (87), the authors took this consideration into account and proposed a stochastic model of gene expression for pSTL1 which included both intrinsic variability and extrinsic variability as depicted in Figure 41. To some extent, it resembles a GRTM with three promoter states in which, among 12 *primary*[81] parameters, two were able to have different values from one cell to another and are depicted in blue in the figure. From a theoretical perspective, this framework is therefore much more satisfying than those presented before. Indeed, it allows modelling both intrinsic and extrinsic noise components in a rigorous manner. Yet, the model was estimated on distributions of fluorescence obtained by flow cytometry, therefore the dataset did not include longitudinal information about single cells trajectory. In consequence, the repartition of intrinsic to extrinsic noise along with the position of extrinsic variability along the typical stochastic model cannot be validated.

[80] Unless all the molecules involved in gene expression such as RNA Pol II, free ribosomes etc. are also modelled as variables of this stochastic model, which would be intractable from a computational point of view.
[81] Since two primary parameters can vary, the real number of free parameters to be estimated was 15 since variance of the extrinsic parameters and their covariance also needed to be estimated from data.

### c. Identifiability of extrinsic and stochastic models of gene expression at the single-cell level

In the examples presented in the previous subsection, the first model was parametrized by hand, using reasonable assumption and trial and error so as to represent data at least qualitatively. Such hand-made parametrization is usually not predictive at all and cannot tell more than *this* hypothetical construction can produce *this* typical behavior and as such is more of a reasoning tool. The other two examples used parameter inference in order to extract parameters values from data. This in turns not only yields more predictive models, but can also answer to some extent to the question "*what matters in this system?"* which is central in systems biology as presented in I.3. In fact, performing parameter estimation requires dealing at some point with the issue of the identifiability of the parameters of a model given some data. A model is identifiable if all of its free parameters[82] can be defined completely and univocally from data alone. Usually, this is not the case and only confidence intervals on parameter values are accessible. In some cases, because of the nature of data itself or because of its quantity or quality, parameters will be non-identifiable[83]. This means that there can be an infinite number of combinations of parameter values which, although much different, will produce indistinguishable outputs given data precision. Identifiability is challenging, but it also indicates when modeling becomes too precise for the sole description of a given dataset.

#### *Identifiability of stochastic models and single-cell data*

Although all models are usually written at the single cell level (*i.e.* they aim at representing what happens within a cell), many are in practice used on population level data only. For example, if we consider the stochastic models presented in the previous section, we need to consider that what was measured and compared are dynamic distributions coming from populations. When it comes to genuine temporal single cell data, as provided by time-lapse microscopy, the direct application of stochastic models leads to difficult estimation problems. This is because within a stochastic framework, a single cell trace is only one possible outcome among many. Given that gene expression, like most biological processes, is not *ergodic*[84] (114, 115), estimating parameters of stochastic processes absolutely requires having several realizations of such process. Yet, in practice, a cell only features a single realization[85] and therefore renders single-cell parameter estimation much more problematic than it is the case for population data.

---

[82] Here a free parameter is a parameter which is to be estimated from data. It is possible to build a model and assign by hand some values while estimating others. Yet, fits and predictions are as good as the value set by hand which is rarely very accurate.

[83] The concept of non-identifiability is crucial, yet regularly overlooked. It is related to the idea of *overfitting* which is more common. Its precise discussion is outside the scope of this thesis but readers can refer to (166) for a graphical introduction. Also, supplementary information of our paper in Annex 6 includes a non-identifiability analysis applied to the model of gene expression we used.

[84] From a statistical point of view, an ergodic process is stationary and when observed for long enough in one system (particle, cell, gene etc.) its time distribution of states matches that of a sufficiently large population of such independent systems. I our case, a single cell would be ergodic if the distribution of its fluorescence level in time would produce the same distribution than a span shot of fluorescence in a population. Yet, because of extrinsic variability which may only change on long time scales, this is not the case.

[85] It can be argued that using dual or multiple reporters could in fact give several independent realizations per cell. Yet, this would impose some additional experimental difficulties and more importantly, may show systematic biases between the different reporter (because of the importance of the genetic context which

This pinpoints an important aspect for modelling variability. Identifiability is related to the model structure (*i.e.* its mathematical formulation), the parameters to be estimated and qualitative and quantitative aspects of data. In this respect, longitudinal single-cell data as obtained by time-lapse microscopy is qualitatively different from data obtained at the population level as provided by flow cytometry of FISH. As a consequence, identifiability of a given model is expected to differ significantly whether it is used on longitudinal single-cell data or on population-level single-cell data.

This should motivate us in making a clear distinction between models that are designed and identified at the single-cell level, and those which are designed at the single-cell level but used on population data only. In this respect, the approach we present in this chapter allows to distinguish both clearly and to define properly how they are related.

Besides mathematical clarity, this overlooked distinction is crucial as biology becomes more quantitative. Building models at the single-cell level and estimating parameters on population data is not an issue *per se*. The problem is when interpreting a parameter value as *the* value (or the average value) for single cells. Actually, parameter values estimated at the population level are not directly applicable to all the single cells composing the experimental data. Rather, they define implicitly some form of *average* or *virtual cell* and *virtual populations* composed of identical cells.

***Combining extrinsic and intrinsic variability***

In the last example of the previous subsection (from (87)), authors presented a model including both intrinsic and extrinsic variability. Although they took care in making their model identifiable[86], no rigorous model selection was proposed regarding which parameters in the model would be subject to extrinsic variability. In fact, model selection can itself be a non-identifiable task, meaning that from data alone we cannot always choose among several model structures with different parameters. Although the biology of the HOG pathway and STL1 expression are quite documented, there is not any information which ensures the proposed sources of extrinsic variability are the major ones at play.

In this study, we are interested in cellular non-genetic identity which falls clearly under the extrinsic label. Therefore, we wanted to impose as little constrain as possible regarding where variability could be present (while still ensuring our model is identifiable). Also, it was important to be able to actually estimate single-cell parameters which represent cellular identity and to test their biological relevance. If we had used a framework combining intrinsic and extrinsic variability, the estimation of single-cell parameters from single-cell data not only would have presented the two types of difficult identifiability issues mentioned, but also the blending of both. In other words, given a single cell trajectory, many combinations of specific single-cell identities (or context) and specific possible realizations from a single identity could yield a equally-plausible fit to our data.

Witnessing that validating a model encompassing both types of variability against data is still very difficult given current experimental possibilities (116), we propose to explore a different

would either be different for all reporter, either artificial if reporters were placed sequentially along the genome). Although challenging, it is nevertheless a promising direction for future investigations.

[86] For example, in their model, ribosomes levels and translation rates could not be estimated separately because all that matters for the data they have is the effective translation rate (i.e., the product of translation rate per ribosome and abundance of free ribosome).

approach in which variability is represented only as stable differences between cells (*i.e.* extrinsic variability). Accordingly, we place ourselves in an experimental situation where intrinsic variability is mild and we neglect it in our modelling by using deterministic (ODE) models of single-cell behavior. This still leaves room for a challenging estimation problem as it will be discussed in the remainder of this chapter. Overall, our simplifying assumption is a necessary first step towards a congruent representation of the total variability in gene expression, and can be readily applied to other biological processes in which extrinsic variability dominates or when the focus lies on cellular identity.

## 2. Mixed effects models of pSTL1 expression

### a. Building a single-cell model of pSTL1 expression including cell-to-cell variability

Here, we propose to represent lasting differences in single-cell gene expression using single-cell parameter values. Basically, every cell in the population is represented by the same model of gene expression but the parameters of this common model can take different values for different cells. Nevertheless, for both practical and empirical reasons, all parameters of this common model are not necessarily variable across cells. Therefore, it should be kept in mind that some parameters can vary in the population while some are common to all cells. In addition, as it will be described in the next subsection, not all parameters are estimated from data. In this subsection we will present the common model and explain the assumptions it is built upon.

In subsection III.1.b, we presented several models which were proposed to represent pSTL1 expression. In particular, the model from (87) served as starting point upon which we performed many iterations to obtain the model we present here. As it was discussed in the previous subsection, we decided to represent single cell gene expression using deterministic models (ODEs). Since the original model in (87) is stochastic, we started by using a deterministic version of it[87]. Nevertheless, this initial model had too many state variables and parameters to be correctly estimated from data at the single-cell level. In addition, as it was stated in the previous paragraph, we seek to enforce the lesser constrains as possible on which parameters can differ between cells. This increases the effective number of parameters to estimate (as it will be explained in the next subsection) and therefore raises identifiability issues. Therefore, we iteratively simplified the model until the point it was identifiable enough while keeping most parameters variable in the population.

The transcriptional response of individual cells is described mainly by two state variables. Denoting with $m$ and $p$ the cellular concentration of mRNA and fluorescent protein, respectively, we have the following ODE system which represents transcription and translation.

$$\begin{cases} \dot{m}(t) = k_m u(t) - g_m m(t) \\ \dot{p}(t) = k_p m(t) - g_p p(t) \end{cases} \quad (3)$$

The production and decay rates are denoted $k_m$ and $g_m$ for the mRNA, and $k_p$ and $g_p$ for the protein, respectively. We assume that these are the parameters which can differ between cells. Although each of these parameters represents the overall contribution of several biological processes, it is possible to propose biological interpretations of why such parameters would differ from one cell to another.

[87] From stochastic models expressed in terms of Chemical Master Equation, it is possible to derive ODEs for all the moments (mean, variance etc.). Yet, in general, this system of ODE is not in a closed from as equations for the moment of order N typically include the moment of order N+1. Using a moment closure approach (170), we use assumptions to make the system closed form. Here variances and any higher moments were set to zero.

Dynamic experimental stimulation of gene expression is encoded by the input function $u(t)$ which represents the whole HOG signaling cascade and more precisely the phosphorylation and nuclear accumulation of phosphorylated Hog1. This abstract representation is represented in Figure 42. We control the valve switching between isotonic and hyperosmotic media and call $u_v(t)$ the valve status. From $u_v(t)$ we compute the effective osmolarity in the microfluidic chambers: $u_c(t)$. The applied transformation essentially represents the delay for the fluid to travel from the valve to the chambers along with the mild mixing during this travel along with diffusion limited washing of osmotic medium. At last, $u_c(t)$ is the input for a first order kinetic representation of the HOG signaling cascade which can be seen as related to Hog1 activity and nuclear localization and is represented by $u(t)$. See Figure 42 A for a schematic representation and Figure 42 B for the typical time course of $u_v(t)$, $u_c(t)$ leading to $u(t)$. See Text S1 and Table S1 of Annex 6 for details and reference on this aspect of our model.

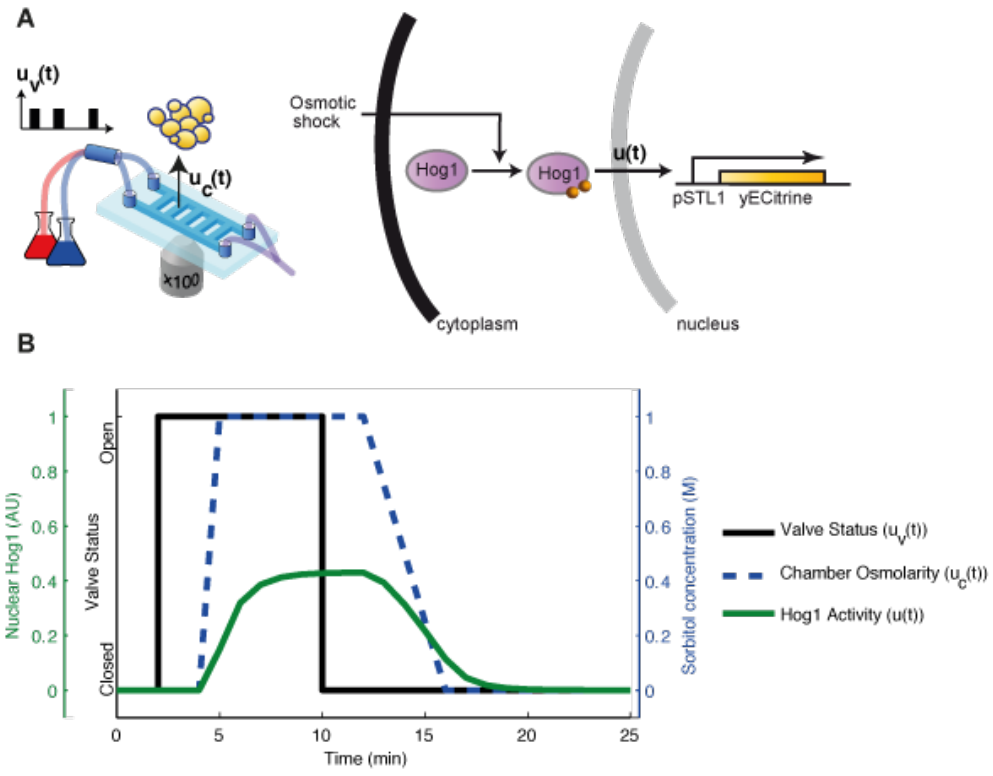



**Figure 42 – Abstracted representation of the HOG signaling cascade in the input function $u(t)$. A. Schematic representation of the transformation from $u_v(t)$, the effective input we impose, to $u(t)$, the input term in our gene expression model. B. Corresponding time course for a standard pulse of 8 min.**

It can be noted that we assume that all cells have the same signaling dynamics as the input function $u(t)$ will be the same for all cells. This is motivated by the fact Hog1 nuclear localization is only mildly variable within the population (see Figure 35). In addition, the absolute value of $u(t)$ is somehow arbitrary and what is relevant to a single cell is $k_m u(t)$ which can indeed be different for every cell.

The model presented so far cannot be compared to experimental data yet. To relate fluorescence measurements $f(t)$ to the protein concentrations $p(t)$, we account for protein folding

and maturation time using a delay $\tau$. In practice, this captures the fact that measurable fluorescence increase will be visible only 40 min after a shock. To faithfully account for the fact that while a protein matures, it is still diluted, we use equation (4) to relate $p(t)$ to $f(t)$. See Text S1 and Table S1 of Annex 6 for details.

$$f(t) = e^{-g_p \tau} p(t - \tau) \quad (4)$$

Given empirical observation in our data and the fact that fluorescent protein maturation rates were reported to display low (~10%) cell-to-cell variability (117) the protein maturation delay $\tau$ is also assumed to be the same for all cells in an experiment but its precise value will still be estimated from data.

At last, in order to estimate parameter values from data using likelihood, we need to asses measurement errors (*i.e.* the expectable deviation of the model from data). To this purpose, we assumed that multiplicative and additive white Gaussian measurement noises were affecting fluorescence measurements, whose strength is the same for all cells (see Text S1 and Table S1 of Annex 6 for details).

### b. Representing extrinsic variability with using Mixed-effects models

In the previous subsection, we presented a common gene expression model which included parameters which value could vary from one cell to another. In this section, we will present how we can relate single- cell models to a model of the whole population using Mixed-effects models. In fact, as every single-cell model can have different parameter values, a straightforward model for a population of cells consists in a collection of as many models as there are cells. Yet, if an experiment has N cells, and considering that 4 parameters can differ at the single-cell level, such representation of the whole population would have roughly 4N effective parameters. This is obviously problematic for several reasons. We expect that for large enough populations, the overall behavior will not depend on each and every cell or on the precise number of cells. Also, by having so many parameters describing an overall population we can hardly determine what is important for the population behavior. One simple way to propose a more concise representation of the population behavior is to further assume that single-cell parameter values follow a given distribution across the population. Basically, this assumption forces variability to be constrained for the purpose of population representation to some class of multidimensional distribution.

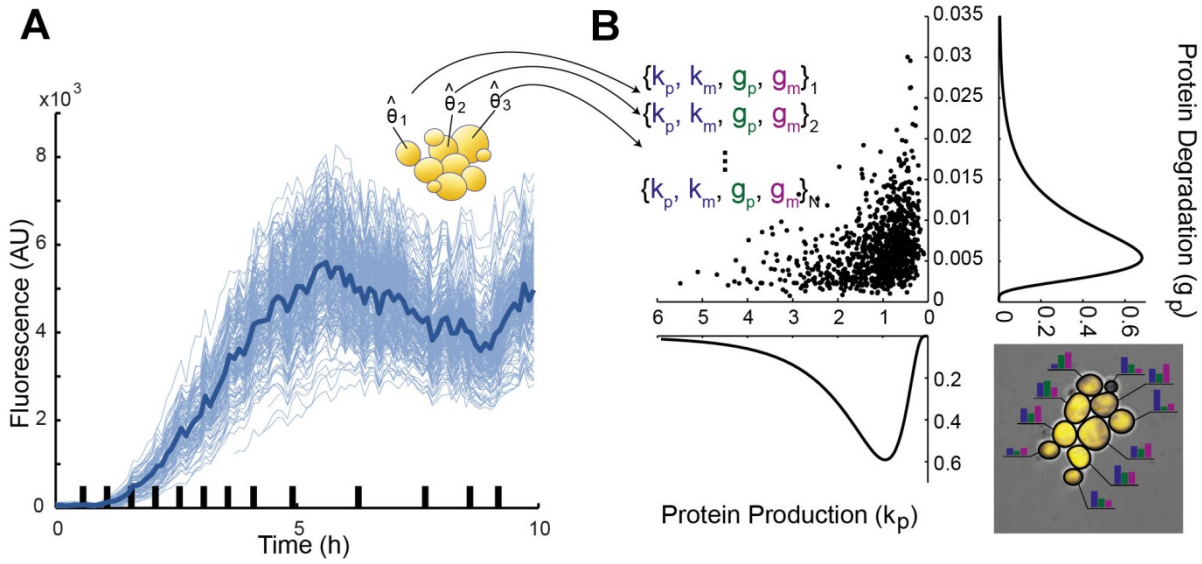



**Figure 43 - Visual representation of the Mixed-effects approach employed here. A. Single cell fluorescence (thin blue lines) in response to several pulses of hyperosmotic medium (black bars). Median behavior is shown in dark blue. B. In ME, each cell has its own set of parameters value as represented in the small insert at the bottom right. Here we represent two out of four parameters which vary across the population. In the scatter plot, each dot is a cell and we provide the marginal distribution for these parameters.**

Mixed-effects (ME) models are a class of statistical models introduced to describe the response of different individuals within a population to known stimuli. In our context, we consider that $k_m$, $g_m$, $k_p$, and $g_p$ vary within the population as represented in Figure 43. We assumed that these parameters were log-normally distributed across the population: $\theta = (k_m, g_m, k_p, g_p)$ with $\ln(\theta) \sim \mathcal{N}(\mu, \Sigma)$, where $\mu$ and $\Sigma$ correspond to a vector of means and a covariance matrix, respectively (see marginal distributions for $k_p$, and $g_p$ in Figure 43 B). This assumption ensures the population is represented in a much more concise and general manner than what would be possible by representing it as the sum of every cell observed in an experiment. In this framework, each cell has a set of parameters ($\theta_i$ for a cell $i$), and the population is described by a set of *meta-parameters* ($\mu$ and $\Sigma$) which describe the distribution of parameters across all cells[88].

### c. Estimating population and single-cell models and validating them

As it was described, in a ME approach we can derive models at both the population and the single-cell scale which are related as single-cell parameter values form a distribution over the population. In this subsection, we are interested on how we can estimate from data, the parameters values for single cells and the parameter distributions for the population. Since the model was written for single-cell data, it is possible to fit each individual fluorescence trace in order to obtain single-cell parameters. The question remains on how we can estimate the population distribution of parameters and how single-cell fits should be conducted.

Concerning the population model, we are looking for a multidimensional distribution defined by its center of mass (*i.e.* a vector of mean values) and its spread (*i.e.*, a covariance matrix) across the

[88] Here we omit parameters which are common to all cells such as $\tau$ for simplicity since they are the same for each cell and for the population.

population. A simple, intuitive manner to tackle this problem is to search for the different parameter values that best describe each individual cell (using single cell likelihood), and then compute the statistics (mean and covariance) of the underlying distribution from the set of single-cell parameter estimates. We refer to this method as the '*naive approach*' since it is the natural starting point and that it is only by testing it that its limitations appear.

An alternative is to conduct the estimation procedure the other way around and start by estimating the population distribution. To do so we use a state-of-the-art algorithm for the identification of ME models: the Stochastic Approximation Expectation Maximization (SAEM) algorithm. SAEM is a stochastic approximation version of the well-known expectation–maximization algorithm and has been developed for the inference of population models in presence of limited available information (118, 119) Notably SAEM is the reference approach in pharmacokinetics/pharmacodynamics studies (120, 121). However, it has not yet been applied to time-lapse single-cell data. The SAEM algorithm directly searches for multivariate distributions by alternating (i) an estimation of (an approximation of) the likelihood of the parameters and individual observations given the current best estimate of the parameter distribution in the population and (ii) an update of the current estimate of the parameter distribution. In a second step, *a posteriori* estimates of the individual cell parameters are obtained from the inferred parameter distribution and individual data (maximum *a posteriori* estimate, MAP). This way, the fact that all parameters share (hidden) traits of the common population is explicitly taken into account. The naive and proposed approaches are graphically represented in Figure 44.

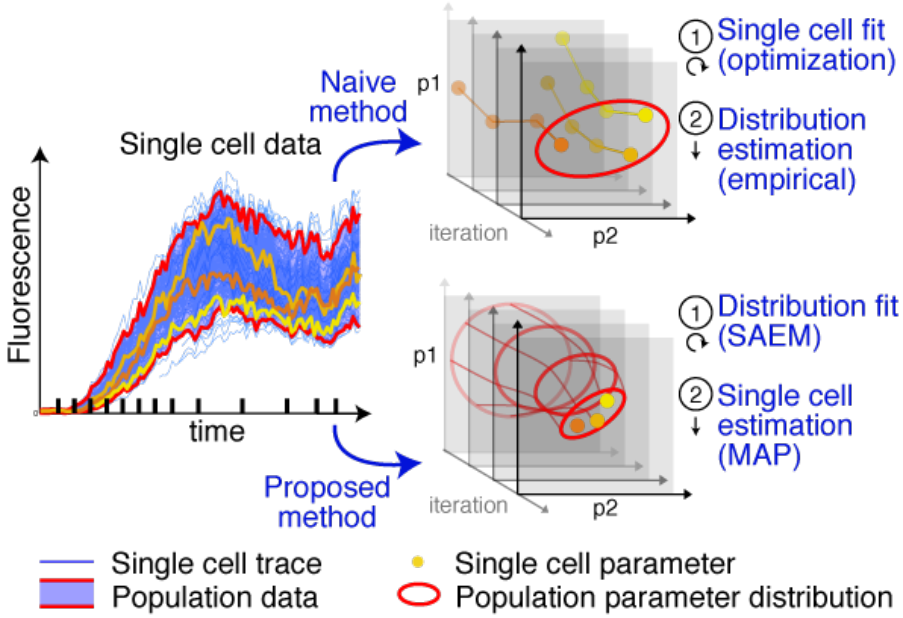



**Figure 44 – Schematic representation of the statistical inference methods for single-cell and population parameter estimation.**

To summarize, in the naive approach, optimization is applied to seek -for each cell- parameter values fitting the individual behavior of the cell via residual minimization[89] (Figure 44 top, step 1). The distribution, describing all of the estimated parameter values, is empirically deduced afterwards (Figure 44 top, step 2). In the proposed method, the SAEM tool is used to infer directly a distribution that explains the set of individual behaviors (Figure 44 bottom, step 1). Parameter values for single cells are then estimated based on the particular behavior of the cell and the inferred distribution for the population[90], using maximum *a posteriori* estimation (Figure 44 bottom, step 2). More details are given in Text S1 from Annex 6.

***Fit and validation***

Both the *naive approach* and the SAEM-based estimation method were applied to an experimental data set comprising more than 300 cells observed during several hours. Despite the significant diversity in the behavior of individual cells (Figure 43 A), both the *naive approach* and the SAEM estimation method were able to find single-cell parameters that fitted well[91] the set of observed single-cell behaviors (Figure 45 A and B).

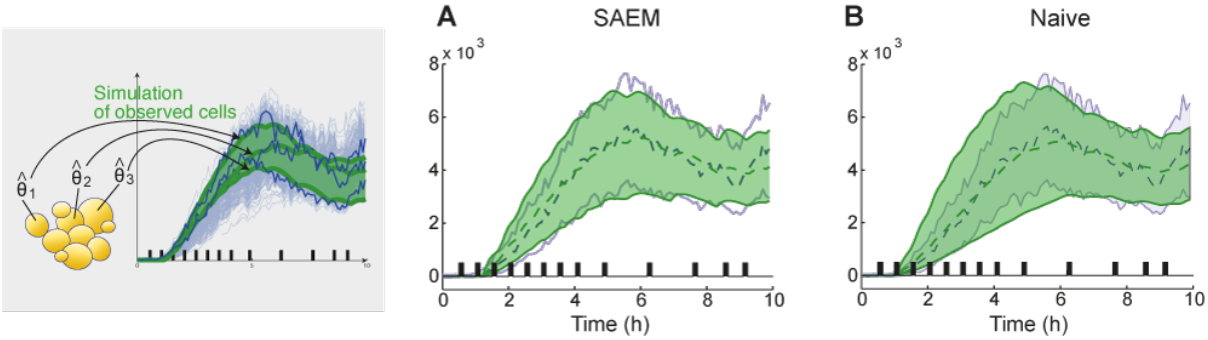



**Figure 45 – Fit performance of single-cell models estimated with the naïve or SAEM method. Envelopes represent 95% of the fluorescent traces and dashed lines represent the median. Green envelopes are simulations for each cell from time 0 to t=10h based on single cell fit results. Blue envelope is data. A Simulated behavior obtained when using the parameters of each observed cell in the dataset (325 cells) inferred with the SAEM approach. B. Simulated behavior obtained when using the parameters of each observed cell in the dataset (325 cells) inferred with the naïve approach.**

We then evaluated the capability of the obtained parameter *distributions* (*i.e.* population meta-parameters) to actually describe the behavior of the cell population (mean and spread). To do so, the parameter distributions obtained using the *naive* and the SAEM approaches were randomly sampled, thus creating two different virtual *'cell populations'* for which the corresponding sets of behaviors were computed from our model of gene expression. The SAEM-inferred parameter

[89] Given the error model is Gaussian at each time step, estimation of parameters by minimizing residuals (*i.e.* the mean squared error) or maximizing the likelihood is equivalent.

[90] This is done in practice by using the population distribution as a prior for parameter values.

[91] It may be surprising to see that the naive method yields slightly larger envelopes than SAEM, given than it has much more flexibility as no prior is enforced on single cell parameters. Yet, as some cells used in the experiment are not born yet at time 0, we see that the naive method attributes parameters to these young cells which produce less realistic simulations at early time points. On the other hand, SAEM does a better job as parameter prior includes information of early time points. When compared only on the data points were cells exist, naïve and SAEM give a relative error of 8.6% and 8.3% respectively which show they perform equally good and in agreement with the assumed measurement errors.

distribution accurately reproduced the observed behavior of the real cell population (Figure 46 A) whereas the *naive approach* failed to do so (Figure 46 B). Therefore, although both approaches were able to identify a set of single-cell parameters that reproduced well the behaviors of observed cells, only the SAEM-based method was able to infer a parameter distribution at the population level which is consistent with the observed heterogeneity in gene expression.

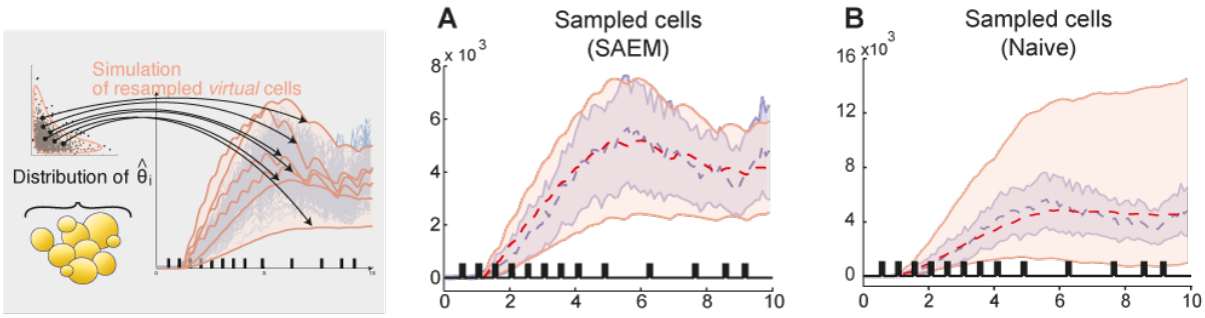
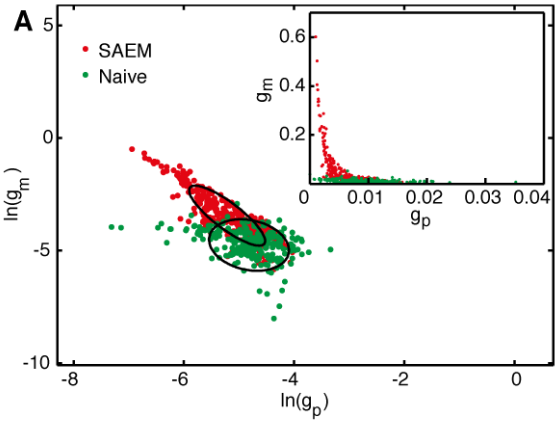



**Figure 46 - Fit performance of population models estimated with the naïve or SAEM-based method. Envelopes represent 95% of the fluorescent traces and dashed lines represent the median. Blue envelope is data, pink envelope is population simulation. A. Simulated behavior of 10,000 cells when resampling the population joint distribution inferred with SAEM. B. Simulated behavior of 10,000 cells when resampling the population joint distribution inferred with the naive approach.**

To investigate the causes of the marked differences between the predictive power of the ME models inferred using either the naive approach or the SAEM-based method, we compared the corresponding parameter distributions. In both cases, the mean values of the parameters were comparable and within the expected ranges (see Table S1 for parameter values and Text S1 for literature values in Annex 6). However, the distribution obtained with the SAEM algorithm was visually more compact and structured as visible on a 2D projection in Figure 47 A. This was confirmed using two metrics to quantify compactness and structure: SAEM yielded distribution with a smaller volume[92] in the parameter space, (see "parameter distribution spread" in Figure 47 B); along with higher cross-correlations on average[93] (see "parameter distribution structure" in Figure 47 B).

[92] The *volume* of parameter distributions is computed as the volume in parameter space of the 95%-confidence ellipsoid associated with the covariance matrix. This yields a measure of the typical volume of parameter space occupied by the parameter distribution, and therefore, quantifies the spread of the parameter distributions

[93] Structure in the parameter distribution is quantified using the average of the coefficients of the variation matrix (*i.e.* of the off-diagonal terms $cov_{ij}/(\mu_i\,\mu_j)$) between the parameters of the model.




B

| | Parameter distribution structure | Parameter distribution spread |
|---|---|---|
| SAEM | 0.62 | 6.7 |
| Naive | 0.19 | 37 |

**Figure 47 - The distribution that better describes the entire population is more compact and more structured. A. 2D plot describing the distribution of the (logarithm of) single-cell parameters for two parameters (insert: same data shown in natural scale). The ellipses represent the region in which 50% of the parameters are distributed. B. Two metrics were computed to quantify the difference in the structure of the parameter distributions at a more global level. See text for details.**

This strongly suggested that the structure of the parameter distribution is essential in order to capture the population behavior. Both the individual statistics of each parameter, and their covariance, describing mutual relationships, contain essential information to properly account for the observed cell-cell variability. And indeed, when using a parameter distribution with the same individual parameter statistics (mean and variance) as the distribution inferred using SAEM but with null cross-correlations (*i.e.* using the marginal distributions), the model lost its capability to represent the behavior of the population (compare Figure 46 A and Figure 48).

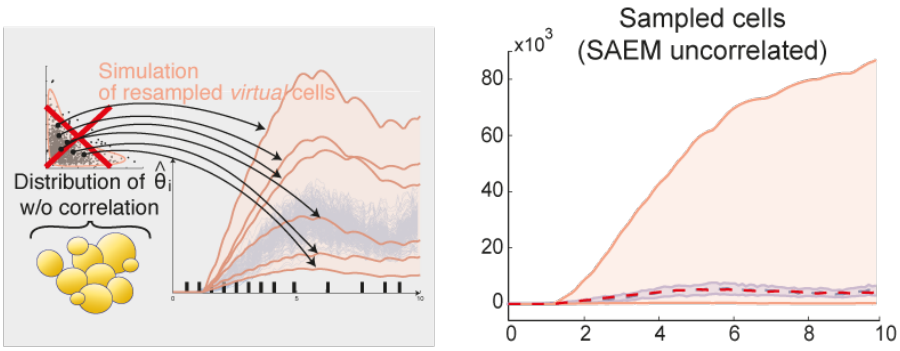



**Figure 48 - Impact of population parameter distribution structure. By removing the covariation information from the SAEM inferred parameter distribution, the fit quality is totally lost which stresses the importance of the parameter distribution structure. Envelopes represent 95% of fluorescence traces and dashed lines are median. Blue envelope Is data and the pink envelope simulation of 10 000 cells.**

Our understanding is that in the *naive approach*, all cells are fitted individually and are subsequently *casted* into a multidimensional distribution. In contrast, SAEM allows finding equally good single-cell parameters while favoring a compact multidimensional representation of the overall population. The difference in performance between these two approaches is rooted in the fact that even with a simple model of gene expression the information contained in a single trajectory is too scarce to constrain the inferred parameter values in a satisfactory way. Using the population distribution as a prior for single cell fits, we actually allow each single-cell fit to use information about the overall population, which ensures coherence between the representation of the population by distributions and the representation of single cells with specific parameter values. Having demonstrated that the SAEM-based identification approach captures the behavior of the cell population, from here on we focus only on the results obtained using this method.

In addition to the fits presented in this subsection, we performed several tests about the prediction capability of the inferred models. For the population model, we tested the prediction of a different experiment or of the same experiment when learning with a shorter time horizon. Also, we tested the prediction capabilities with a shorter horizon for single cells. At last, we investigated the robustness of this approach as the number of single cell traces is reduced and found that decent population parameter distribution could be obtained with as few as 36 cells. We refer to Annex 6 for details.

***Identifiability in single cell and mixed effects models***

Another important consideration for parameter estimation in the case of ME models, and in particular when using SAEM is the flexibility in the status of each parameter regarding estimation. As it was mentioned, some parameters can be fixed for all cells from the start (such as those underlying the definition of the input function $u(t)$), other parameters are equal for all cells but can still be estimated from data (these are called *fixed effects* which include here $\tau$ and the noise model parameters), and at last, some parameters are variable across the cell population and are estimated from data (these are called *mixed effects* or *random effects*).

A non-identifiability analysis of our model revealed that the parameters $k_m$ and $k_p$ were structurally non-identifiable at the single cell level (see Annex 6 TextS3). This means that from a single-cell estimation perspective, only the product of these parameters is observable from the dynamics of the fluorescence. Therefore, we define $k_{mp} = k_m k_p$ which will be in fact the parameter which is estimated[94] for all the following results which concern single-cell parameter values.

Nevertheless, it is non-trivial to assess what are the consequences in terms of non-identifiability for the population model. This is because as it was stated, the population model is actually different from the single-cell model, although derived from it. For the population, $k_m$ and $k_p$ not only show up in the vector of means (in logspace) $\mu$, but also in the covariance matrix $\Sigma$ with variance and covariance terms. Although a conservative reflex would push towards fixing all population parameters related to one of the non-identifiable ones, we considered that it was indeed

[94] An equivalent and practical way to do so is to fix one of these parameters to an arbitrary value and to estimate only the other.

possible that having both parameters varying may result in a different identifiability perspective[95]. Therefore, we decided only to fix the component corresponding to $k_p$ in $\mu$ but to leave its variance and covariance with other parameters for estimation from data. This corresponds somehow to a *fixed, mixed effect* and was carried in all the results concerning the population model presented previously.

Another sanity check which should be carried when using ME has to do with shrinkage. As it was stated, using SAEM we first compute a distribution of parameters for the overall population of cells and we subsequently attribute to each cell parameter values by MAP using the population parameter distribution as a prior. Shrinkage will appear if overall, the single cell parameters do not *repopulate* the distribution which was inferred at the population level. This is not the case here as visible on Figure 49.

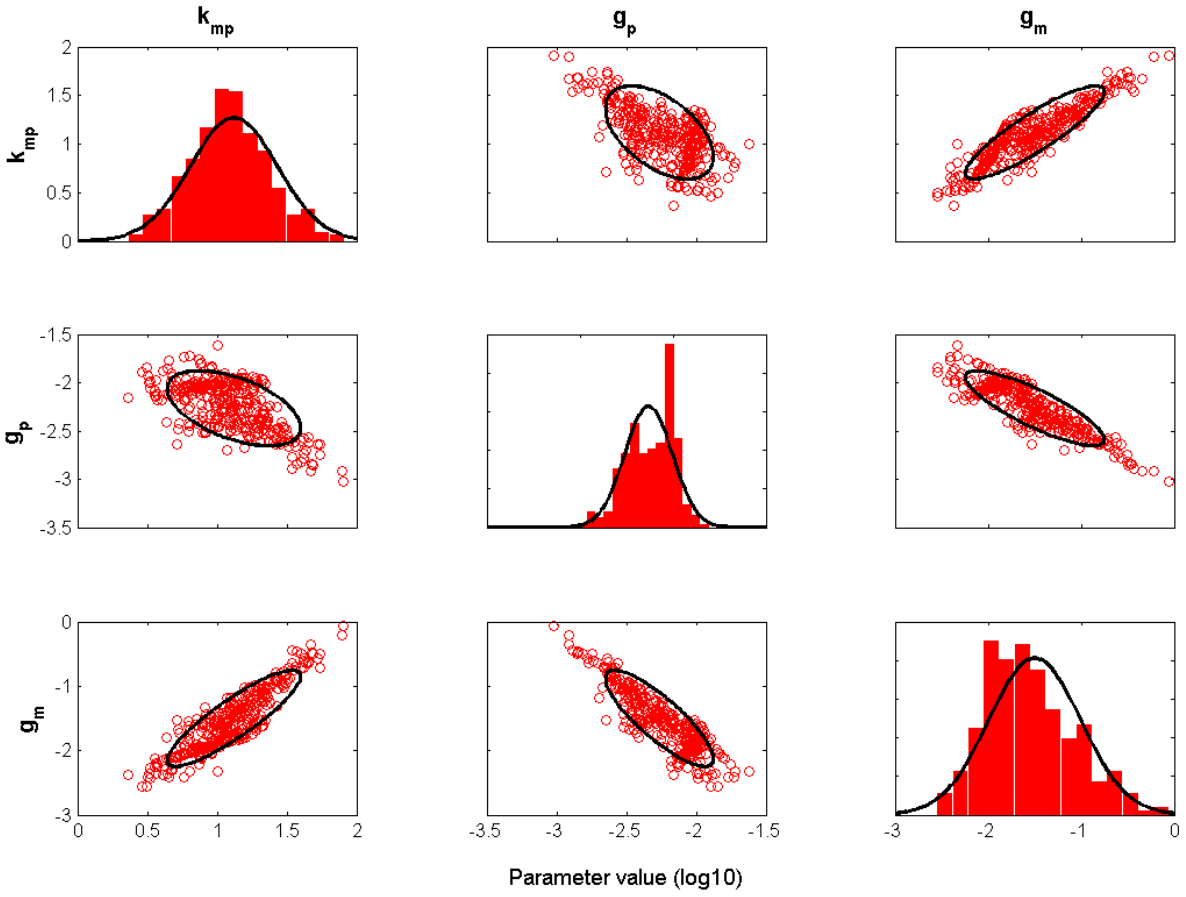



**Figure 49 - Comparing single cell parameters and population distributions. Plots on the diagonal represent the marginal distributions (black line) estimated from SAEM for each parameter along with a histogram of all the single cell parameters estimated by MAP. The plots off-diagonal show the covariation of parameters one with another. In red are represented the inferred single cell parameters and black ellipses correspond to one standard deviation (68%) confidence intervals from the population distribution inferred with SAEM. 325 cells from experiment Di were used.**

[95] A publication having for subject the identifiability of mixed effects models should be published soon by M. Lavielle. We hope it will include a more general and detailed analysis of this interesting question.

# 3. Cellular identity and gene expression

## a. Relations between gene expression and cell physiology

Several features of the cell physiology and local environment were speculated to be related to stable cell-to-cell variability in gene expression (109). Such features notably include cell division rate, cell size, cell age, and local cell density. Thanks to specifically designed image analysis algorithms, these features can be measured or estimated for each single-cell based on bright-field time-lapse imaging. In consequence we tested in a systematic manner for empirical evidence of such relations between cellular features and the parameters that describe cellular individuality in gene expression.

First, we searched for a correlation between the protein decay parameter, $g_p$, and the cell division rate. Indeed, as the fluorescent reporter we used has a long half-life, one should expect that its observed decay comes mostly from dilution due to cellular growth. Therefore, we quantified for each cell its division rate averaged over the observation period (as described in II.3.b) and, as expected[96], found a significant positive correlation between the measured average single-cell division rate and the protein decay parameter $g_p$ (Figure 50). Stated differently, using exclusively the fluorescence profile of individual cells and the inferred parameter distribution for the cell population as an *a priori*, the inference approach attributed statistically higher dilution rates to cells that grow faster.

| | Mean Density | Mean Division rate | Mean Size | Mean Age |
|---|---|---|---|---|
| $g_p$ | -0.03 (0.61) | **0.20** ($2.1\times10^{-4}$) | 0.04 (0.44) | -0.10 (0.16) |
| $g_m$ | **0.21** ($3.5\times10^{-4}$) | **-0.23** ($2.6\times10^{-5}$) | **-0.21** ($1.3\times10^{-4}$) | -0.06 (0.39) |
| $k_{mp}$ | **0.35** ($4.4\times10^{-11}$) | **-0.19** ($7.4\times10^{-4}$) | **-0.30** ($2.4\times10^{-8}$) | **-0.19** ($2.7\times10^{-3}$) |
| PC1 (87%) | **0.22** ($7.2\times10^{-5}$) | **-0.23** ($4.2\times10^{-5}$) | **-0.21** ($1.5\times10^{-4}$) | -0.05 (0.46) |
| PC2 (12%) | **0.34** ($4.5\times10^{-10}$) | 0.05 (0.35) | **-0.24** ($1.4\times10^{-5}$) | **-0.24** ($1.5\times10^{-4}$) |
| PC3 (<1%) | **-0.13** (0.02) | **-0.12** (0.03) | -0.02 (0.76) | 0.00 (0.98) |

**Figure 50 - Correlations between these single-cell features and the single-cell parameter estimates and their principal components are provided with their corresponding p-values. The proportion of variance accounted for by each principal component is indicated in parenthesis.**

Several other highly significant correlations between single-cell parameters and the above-mentioned single-cell measured features were observed (Figure 50). Note that all measured features were averaged across time to allow the comparison with the time-invariant model parameters (Text S1 Annex 6). Although it is difficult to attribute in a systematic manner a direct and unambiguous biological interpretation of the observed correlations between coarse-grained model parameters and

[96] The fact that the actual correlation coefficient is not very high reflect probably some mis-estimation and measurement errors but also can arise from the fact that here we consider the average division rate over several hours and that division rate is related but not equivalent to growth rate in terms of volume which is the proper metric for dilution rate.

cell features, one can nevertheless observe (i) that cell density appears to have a pronounced influence on the protein production rate, suggesting that - even in microfluidic growth chambers - the environment of the cells should not be assumed to be perfectly homogeneous, and (ii) that the correlations of the protein production rates and mRNA degradation rates with every measured feature always have the same sign, corroborating the presence of mechanisms for the joint regulation of these processes in our system.

More generally, one wonders how the different measured cell features relate to the overall (multivariate) parameter variability. We conducted a principal component analysis (PCA) of the set of inferred single-cell parameter values. This yielded a new parameterization of the model (new parameters being called principal components PC1, PC2 and PC3) which is particularly relevant to investigate variability as, unlike natural parameters, each principal component is uncorrelated to the others. A visualization of what these principal components may represent in terms of actual single cell data is provided in Annex 7. The analysis showed that the first two components PC1 and PC2 represented 87% and 12%, respectively, of the overall variance in single-cell parameter values, and that these principal components correlated very significantly with measured cell features (Figure 50). We then ranked the various features based on their correlation with the variability captured by the inferred ME model. For a given feature, this is defined as the weighted average correlation with the different PCs, with weights equal to the importance (*i.e.*, explained variance) of every PC. It appeared that local cell density was the most important factor (average correlation: 0.23), followed by cell size (0.21) and the division rate (0.2). Quite surprisingly, from our data, age was not associated with a significant variability in parameter values. Taken together, our results show that, for quantitative studies, features other than culture medium or colony growth rate should be taken into account when comparing experiments.

### b. Inheritance of phenotype and gene expression features

Finally, we investigated inheritance of single-cell parameters. Statistical tests showed that the parameters of mother and daughter cells were significantly closer to each other than the parameters of 20 000 random cell pairs (Text S1 in Annex 6 and Figure 51). However, this comparison does not exclusively test the effect of lineage. The fact that mother and daughter cells share a similar environment may also explain this result.

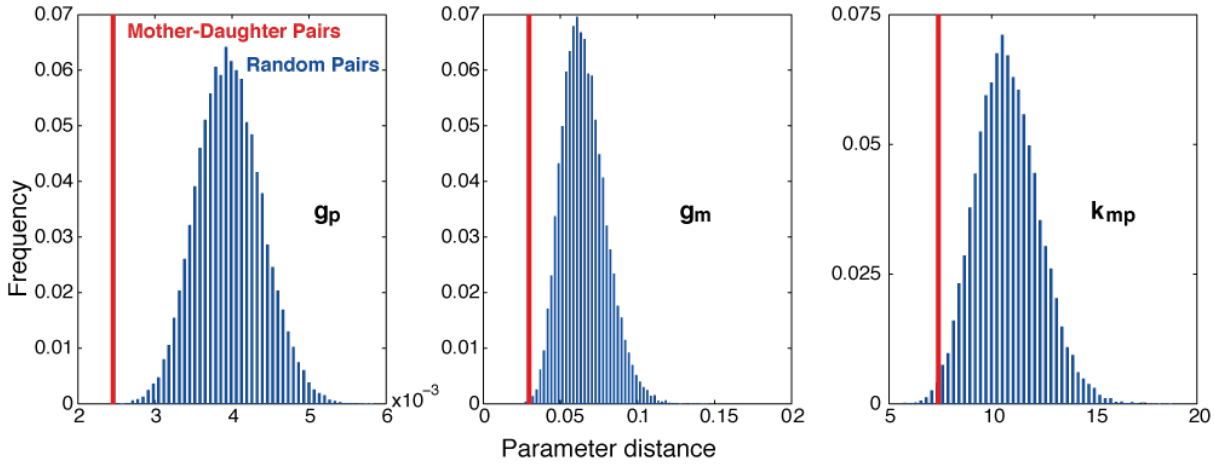



**Figure 51 - Mother and daughter cells have closer parameter values in average than random pairs of cells. For each parameter, we report the average value of mother-daughter distance in parameter values (red lines) and compare it to the distribution of parameter distance of 20 000 couples of cells picked at random (blue histograms).**

To study the specific influence of lineage, we compared the parameter values between pairs of cells that either were mother and daughter (related mother/daughter pairs, called MD) or were a mother and the unrelated daughter of another mother cell (non-related mother/daughter pairs, called nMD), with all cells growing in the same microfluidic chamber so as to limit environmental bias. Comparing the average parameter distance revealed that mother and daughter had closer parameters values and this for all parameters (compare the blue and red large vertical bar in Figure 52).

Yet, as the parameter distance varied importantly among particular pairs, we wanted to test this hypothesis statistically. As the empirical distribution of the distance between parameters for all MD pairs was not reasonably approximated by common distribution shapes, we employed a bootstrap approach to compute empirically the probability distribution of the estimator of the mean (which is Gaussian as soon as enough samples are drawn). This allows us to derive p-values for the inequality of the mean between MD and nMD average parameter distance[97].

As shown in Figure 52, the parameter values of individual cells are indeed statistically closer to the parameters of their own mother cell than to the parameters of another mother cell. More precisely, they are 16% (resp. 14%, 10%) closer in genuine mother/daughter pairs for $g_p$ (resp. $g_m$, $k_{mp}$). Although mild in absolute terms, the bootstrap test showed the presence of a statistically

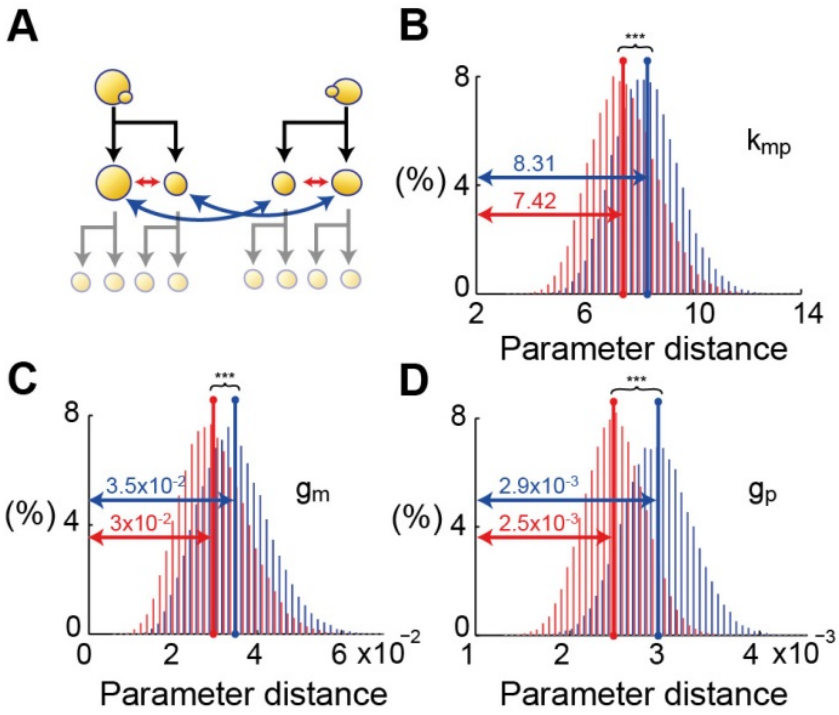



**Figure 52 – Statistical analysis of single cell parameters inheritance. Comparison of the mean distance between parameters of either a mother with a daughter (MD, red histograms) or a mother with a non-related daughter (nMD, blue histogram). Details are in the text.**

[97] In the example plotted, 50 000 bootstraped sets of 40 pairs (either from MD, or from nMD) were drawn to derive each histogram and p-values were computed using classical two sided t-tests for mean inequality of Gaussian distribution with different variance.

strong inheritance effect ($p$-values < $10^{-15}$ for all parameters, see Text S1 Annex 6).

***Biological interpretation of inherited single cell parameters***

When we apply the same inheritance testing to some of the single-cell features which were shown previously to correlate with single cell parameters, we found that several ones were in fact *anti-inherited*, meaning that mother and daughters were significantly more different that non-related mother and daughters. As we can see in Figure 53, this is the case for the intensity of the perceived shocks, with daughter cells being 14% more sensitive than their mother on average. This might be related to size effects (as mothers are larger than daughters) or to properties of the cell wall (daughter cells having a newly synthesized cell wall which differ from the mothers).

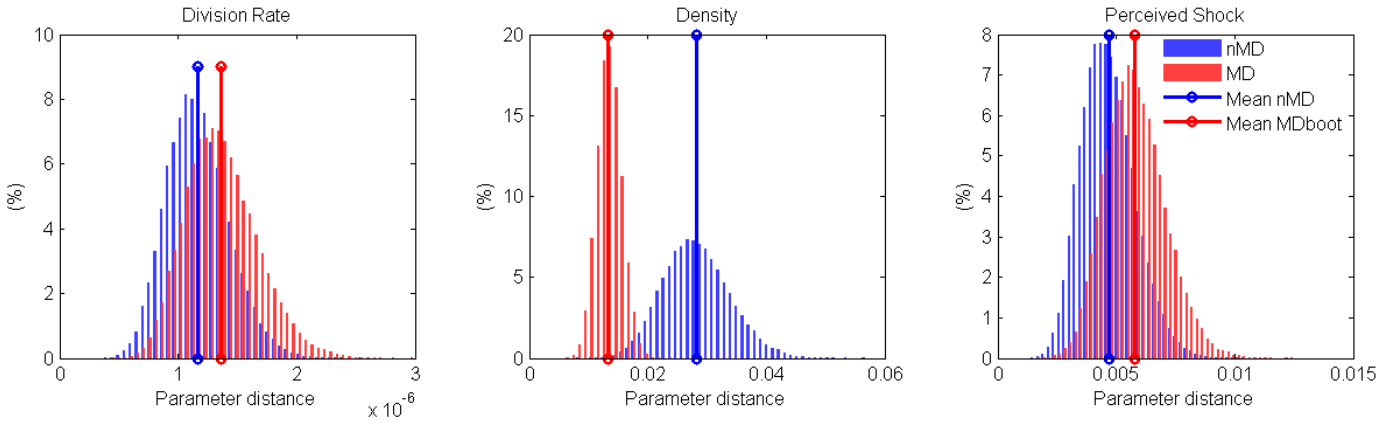



**Figure 53 - Statistical analysis of single cell features inheritance using the same methodology as in Figure 52. NB: perceived shocks are defined in the next subsection. MD stands for Mother and Daughter Paris and nMD for non-related Mother and Daughter pair.**

At last, concerning density, it is expected that mother and daughter being spatially close will experience a more similar local density. As cellular density was demonstrated to be the most correlated cellular feature with parameter values, inheritance of density could bias the previous results on parameter value inheritance. As we can properly assess such density inheritance, we were able to construct a more refined set of non-related mother and daughter pairs which featured nearly exactly the same distribution of density distance as MD pairs. Running the previous inheritance test with such set of nMD pairs gave nearly exactly the same results[98] corroborating the idea of a genuine partial inheritance of single cell parameters..

As a conclusion, we can see that parameter values, along with cellular features, exhibit some degree of inheritance (or anti-inheritance) from a mother to a daughter. We can wonder what biological features could be the basis for such inheritance of non-genetic identity. Such an observation leads to many possible speculations, some of which we share thereafter.

[98] It can also be noted that testing inheritance in terms of principal components rather than natural parameters yields very similar results.

Mother and daughter are usually living in proximity and therefore will experience similar microenvironments. Despite our efforts to mitigate such potential influence from our measurement of inheritance, we cannot completely refute this simple hypothesis.

We can also propose that epigenetic effects would lead to inheritable changes in gene expression features over the course of several generations. In the specific case of the HOG pathway and pSTL1, we know that chromatin remodeling complexes such as the SAGA complex are recruited and alter the chromatin organization(62). Also, pSTL1 is located close to telomere regions of the chromosome which are known to be subjected to epigenetic silencing. The SIR complex (and Sir2 in particular) is involved in silencing in these regions. Sir2 silencing can be affected by the Redox balance (122). Knowing that glycerol production also plays an important role in maintaining the Redox balance, one can wonder if this would indirectly relate response to osmotic stress and epigenetic silencing. Yet tentative experiments using TSA or Nicotinamide (which are inhibitors of several epigenetic silencing processes) were inconclusive.

Although we always consider epigenetic phenomena when it comes to non-genetic inheritance, other mechanisms should be considered as well. As gene expression networks form large dynamical systems, they may possess many local dynamic equilibrium states. Therefore, we could envision a situation where a mother cell transmits, along with its DNA, the current *state* of her GRN, including its local equilibrium which will change slowly.

Yet another possibility would come from the consideration of the impact of osmotic stress on the cell cycle (see chapter IV). This could lead to a situation where mother and daughters are mildly synchronized in their cell cycles by some form of lock-in due to the frequency of stimulation we applied here. The measured inheritance being reminiscent of cell cycle effects in the response to osmotic stress.

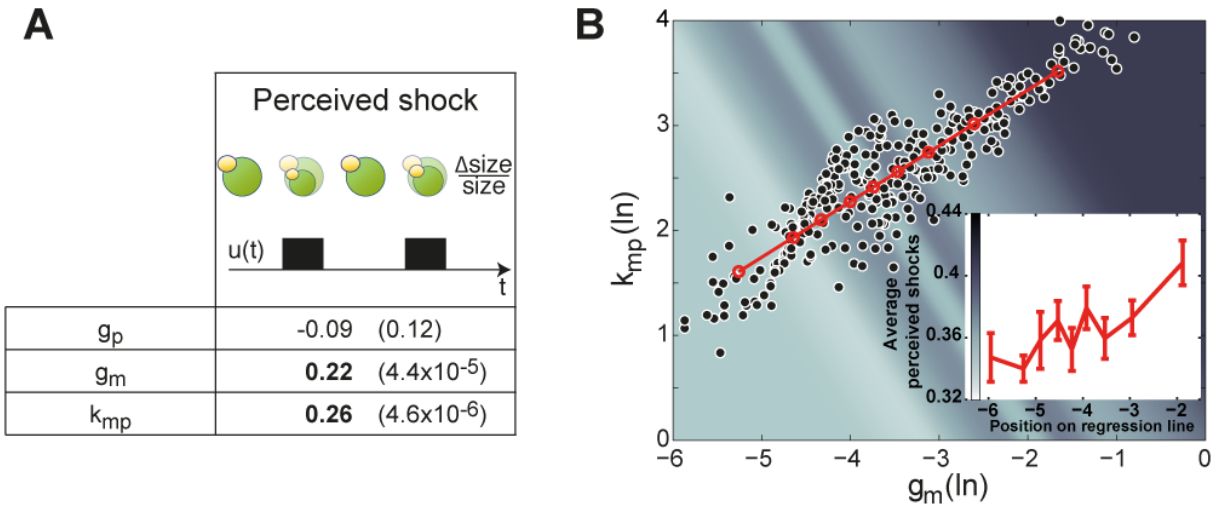
Concerning anti-inheritance effects, we might recall that several proteins are known to exhibit an asymmetric repartition between mother and daughter cells (123) which could therefore create such effect. Considerations of the cell wall properties which differ in mothers and daughters and are also related to the mechanical aspects of osmotic stress might also provide candidates for this effect.

Therefore, we see than when it comes to cell-to-cell variability, many factors can introduce biases in analysis and many known biological factor could equally play out. Rigorous statistical testing is a necessary tool to try to avoid bias issues but it is not possible to protect against all bias. From these exploratory analyses, hypothesis driven experiments are needed to confirm the importance and decipher the underlying mechanisms behind such inheritance relationships.

### c. Listening to the noise: harvesting natural cell to cell variability

Having identified single-cell parameter values, one may wonder whether they can be used to retrieve known facts or discover new ones on the physiology of the cell response to hyperosmotic shocks. In our model, hyperosmotic shocks affect all cells identically (in terms of signaling cascade). However, from a physical point of view, the intensity of the shock perceived by different cells varied, as evidenced by differences in the reduction of cellular volume following shocks. Therefore, one might expect that protein production parameters inferred for the most severely impacted cells are statistically higher than average. We thus estimated the perceived shock intensities as the time-

averaged reduction of cellular volume following shocks, and compared for all the cells the inferred parameter values and the perceived shock intensities. We found a strong correlation between protein production rates and shock intensities in agreement with our hypothesis. Moreover an equally-strong correlation was also found with mRNA degradation rates (Figure 54 A). This second feature, obtained by our framework without any additional measurements or hypothesis, is consistent with the known global destabilization of mRNAs observed after hyperosmotic shocks (77). Lastly, the high correlation between protein production rates and mRNA degradation rates (Figure 54 B) indicates that these two processes are jointly regulated in response to hyperosmotic shocks. Note that the direct experimental identification of such co-regulations would be very challenging. This shows the interest of extracting and analyzing distributions of model parameters for co-regulation identification.




| | |
|---|---|
| $g_p$ | -0.09 (0.12) |
| $g_m$ | **0.22** ($4.4\times10^{-5}$) |
| $k_{mp}$ | **0.26** ($4.6\times10^{-6}$) |

**Figure 54 - Effects of hyperosmotic shocks on intracellular processes involved in gene expression. A. Correlations between the perceived intensity of hyperosmotic shocks and single-cell parameter estimates are provided with their corresponding p-values (Text S1 Annex 6). B. Estimated values for protein synthesis rates $k_{mp}$ and mRNA degradation rates $g_m$ for each individual cell. Their strong correlation (Spearman coefficient: 0.88; *p*-value<$10^{-15}$) together with their mutual increase with perceived shocks intensity indicates that these two processes are jointly regulated in response to hyperosmotic shocks. Insert plot and colored background represent perceived shock intensity for 9 groups of 35 cells along the regression line.**

# 4. Conclusions on: Individuality in the transcriptional response to osmotic stress

***Personal contributions***

The study presented in this chapter has been done in close collaboration with the co-authors of the article provided in Annex 6. For the purpose of a fair evaluation I should state my personal contributions to this study. These includes: the realization of experiments, image analysis both for single-cell fluorescence extraction and for the quantification of all other cellular features of the microenvironment or of cellular physiology. I performed the inheritance analysis along with correlations of cellular features with single cell parameters. I contributed to the construction of the single-cell model and to the parametrization of SAEM runs (although not performing them myself). I took part in the mathematical and biological analysis of single-cell parameters and population distributions. At last, review and discussion of all analysis, along with redaction of the article was done collectively.

***Conclusion***

In this chapter, we used a classic reporter of the transcriptional response of the HOG pathway along with a specific design of experiments using dynamic hyperosmotic stress in order to reveal long-lasting differences in single cell genetic expression. In sharp contrast with previous work on cell-to-cell variability on this system, which mainly focused on the intrinsic aspect of gene expression noise, we focused on the extrinsic part only. From our early discussion of identifiability issues arising when estimating the parameters of models of gene expression at the single-cell level, we can now better assess the potential issues arising when both intrinsic and extrinsic noise are considered jointly. Arguably, the information content of our experiment is to some respects qualitatively richer[99] than what was used in previous studies. Yet, identifiability considerations in our study indicated that only coarse-grained models could be used to infer extrinsic variability and assign precise values to each cell. At the same time, as it was investigated in (124), mixed-effects models and stochastic models using moments (as in (87)) were found to be equally good at explaining population data only (as it would be obtained by flow cytometry) but only ME allowed decent single-cell fits. Therefore, ensuring the identifiability of a detailed model including both intrinsic and extrinsic variability should probably require even more informative experiments and require a very careful estimation method. A noticeable trial in estimating models with both intrinsic and extrinsic noise from single cell longitudinal data and using a different gene expression system in yeast (116) has shown that indeed, large stochastic models (as used in purely intrinsic models) are not adapted when extrinsic variability is also considered.

We presented an approach for capturing the biological variability observed in single-cell time-lapse microscopy experiments of gene expression by *distributions* of parameters. By doing so, we address a fundamental issue encountered in the vast majority of quantitative studies where parameters of deterministic or stochastic models of intracellular processes make sense at the single-

[99] This is because from many single-cell traces it is possible to reconstruct at each time points distributions of single cell fluorescence, but such distributions as given by population snapshots experiments (*e.g.* flow cytometry or FISH) cannot provide any single-cell temporal information. Yet we do not use varying levels of stress or mutants which might bring even more information.

cell level but are estimated from population data and therefore are actually representing a virtual '*mean cell*' rather than actual cells. As it was discussed in the introduction of this thesis, variability is ubiquitous in biological systems. Although reasoning in terms of *average* effects is useful, and for a long time was mandatory given the experimental possibilities, it appears that variability somehow *combines* with the common non-linearity of biological processes. It results in systematic estimation bias when variability is neglected and only average values are considered. This is detrimental to the generality and reproducibility of quantitative research in biology. Therefore, more care should be taken to underline the distinction between genuine single-cell measurements or estimations and information coming from population information.

Our analysis was based on the mixed-effects (ME) modeling framework and two inference approaches were evaluated. The use of adapted estimation methods and advanced algorithms, like SAEM, was essential to properly capture the variability of biological parameters across the population in a simple manner, including most notably the correlation among them. With this approach, the information on each and every cell is jointly used to calibrate the population parameter distribution and in the end, constrains single-cell parameter estimation. This approach alleviates the problem of limited observability and noisy observations encountered at the individual cell level. This explains the surprisingly low number of single-cell data which is sufficient to represent population variability in our system (see Text S2 in Annex 6) and explains the robustness of this approach.

Although we have tested how the inference method scales when fewer cells are used, we did not study directly how important was the use of complex gene induction patterns for the estimation of relevant parameters. In this respect, from our principal components analysis of single-cell parameters, we can derive *eigen cells*. Eigen cells are virtual cells which represents independent modes (or types) of parameters combination which can *summarize* the observed variability in gene expression. Looking at simulated traces of these eigen cells (Annex 7) we can see *a posteriori* that it is only on the latter part of our experiments, (when random pulses are applied) that the first and second eigen cells' trajectories diverge. Therefore, using complex temporal stimulations of the HOG pathway was indeed instrumental to discriminate the two major modes of variability encoded by the two first principal components. Although the design of dynamic stimulations was based here on informed, yet intuitive guessing, it is highly relevant to propose a more systematic experimental design methodology as more complex models are used.

As it was presented, the practical application of mixed effects model estimation requires an adequate modelling effort if identifiability is important[100]. In particular, a precise modelling construction methodology should be used to find the proper modelling scope which satisfies the tradeoff between detail and identifiability. The fact that parameters can be variable or not (*fixed* vs *random* effects), estimated or not leads to a great flexibility in this approach. Yet, a consequence of such flexibility is that many possible combinations are possible and should be screened during systematic model selection. Further investigations could be performed so as to find the number of minimal sources of variability in the model as well as their precise position.

[100] It should be reminded that if parameter estimates are not used for biological interpretation but only serves to calibrate a model which is therefore used as a black box model which is used for prediction, identifiability issues can be ignored to some extent. Such approach can be sufficient to implement a model predictive controller for instance but is not relevant to describe biological processes.

A significant contribution of this work comes from the demonstration of the biological relevance of the inferred cell-specific parameters. Our single-cell parameters which are based on fluorescence measurements only were found to contain information which overlaps with biologically relevant measurements on the same cells and which were not included in any manner in our model. These include genealogy, micro-environment and single-cell physiology which, taken together, constitute a set of corroborating evidence supporting two distinct claims:

i. **The proposed method allows extracting biologically relevant information at the single cell level from dynamic fluorescent measurements only**. Indeed, if our model or our estimation of variability could not be related to independently-assessed biological features, any biological interpretation would be shadowed by a suspicion of an *artifact based* variability representation[101].

ii. **Although simplistic, our description and estimation of extrinsic variability allows a prospective interpretation where single-cell parameters define some form of cellular identity in gene expression dynamics**. Such interpretation should be taken with care as we only consider here a particular biological process within a single experimental framework. Yet, the fact that single-cell parameter values are related to known contributors to cellular identity (*i.e.* size, physical properties of the cells, micro-environment history etc.) and seems to be inherited differently than these known determinants might suggest that gene expression features and physiological features are related, yet not identical, aspects of what makes every cell unique.

***Perspectives***

The first immediate perspective of the study presented here is its application to other processes than pSTL1 expression in order to assess to what extent our methodology can be generalized. In addition, hypothesis driven experiments aiming at providing direct evidence of some covariation between gene expression features on one hand and environmental and physiological features on the other hand are necessary to strengthen our findings. Nevertheless, experimental limitations make some of these validations technically impossible at present times.

Another interesting perspective concerns the broad question of *"what matters?"* for single-cell variability in gene expression. As it was discussed, a typical distinction is made between random fluctuations emerging from stochastic gene expression (intrinsic noise) and more stable differences which form the basis of some form of cellular identity. Although much work has been done in the past concerning the former, more and more studies now quantitatively assess the later. Extrinsic variability in fact is here defined in opposition with its intrinsic counterpart. Yet, such negative definition makes it too vague for a meaningful discussion (18). A more relevant breakdown of individuality in gene expression should focus on the causes of stable differences and include several properties of variability.

In particular, it was proposed in (109) that extrinsic variability could be to a large extent caused by single cell physiology and micro-environment. A genome wide and high-throughput systematic analysis of both mRNA transcript counts, physiological state and micro-environment features in

[101] Our discussion in the introduction of this chapter explains why we can have such suspicion concerning models of pSTL1 expression forcing variability to be fully intrinsic.

mammalian cells, relying of advanced image analysis of FISH data (125) was carried recently. This studied allowed single cell transcript abundances to be predicted with high accuracy from phenotypic descriptors only[102], highlighting the deterministic nature of gene-expression variability. Yet, the precise causal relationship between what we could call *gene expression identity* and *augmented phenotypic identity* (which includes information about the micro-environment) cannot be resolved in general from snapshot data of unperturbed cell populations[103]. Advanced experimental design combined with longitudinal experiments will be necessary to resolve causality hierarchy in the overall cellular individuality. In this respect, the importance of covariation among single-cell parameters raises important experimental challenges as most of the time it is not possible to measure directly parameters at the single-cell level, let alone several of them simultaneously. An important part of stable individuality is expected to come from the levels of proteins, yet given the number of putative relevant elements, it is not possible using fluorescence tags to systematically screen for all of them or worse, for possible covariations among them. In this respect, an extremely challenging, yet technically possible experimental system which would combine time-lapse fluorescence microscopy followed by in-situ single-cell proteomic measurements is probably the best prospective to relate observed variations in gene expression dynamics to absolute protein abundance and to screen putative molecular sources of extrinsic variability in a systematic manner.

The vision of gene expression identity which comes with stable differences in gene expression features should also be considered from a dynamic perspective. As it was discussed in I.1.b, the cellular identity in gene expression at a given instant is probably the results of structural and molecular features fluctuating with different time-scales. This in turns calls for distinct interpretations of cellular identity as different time scales are considered. In the experiments which were presented here, cell identity was defined implicitly as a time averaged effect over a few cell-cycles which *de facto* removes extrinsic components whose fluctuation is faster. Given previous studies on extrinsic noise dynamics (126) in *E. coli*, it is surprising to find that some aspects of such identity were stable enough to produce biologically significant parameters. Still, there is direct evidence than several core features of the augmented phenotypic identity were changing during the course of our experiment. This means that if the proposed method was to be applied to longer experiments, it might require at some point either to split single cell data into fixed periods which would correspond to different cellular identities (*i.e.* taking into account changes in identity in a discrete manner), or to account explicitly for continuous time-changing identity, combined with the contribution of cellular division as a fundamental discrete event affecting identity[104] (through inheritance and anti-inheritance that affects both the mother and the daughter cells). This latter option would in principle allow a beautiful, yet extremely challenging, unification of intrinsic and extrinsic noise were intrinsic noise over all the genome constitutes the basis for slow identity fluctuation upon which dynamics of the micro-environment (including contribution by neighbor cells) would act.

[102] Work by L. Pelkmans' group communicated at the EMBO/EMBL symposium on Cellular Heterogeneity in Heidelberg, 15-18 April 2015, still unpublished at the time of writing.

[103] In fact, from snapshot data only, distinguishing from A causes B, B causes A or X causes both A and B is a chicken and egg problem.

[104] From M. Lavielle comments on this work, it is worth noting that in fact, in its current form, mixed effects models assume individuals composing a population to be independent one from another. Accordingly, our finding that some inheritance effects are present already requires a modification of the general theoretical framework we employed to include some form of dependence between single-cell individuality.

Interestingly, division and growth, which are tightly coupled processes (so as to ensure cells size to be bounded), are essential components of both the augmented phenotypic identity and of gene expression identity (as most proteins decay is driven by dilution). Although it was considered constant in time (yet variable among cells) for our study until now, the division rate of cells is actually changing in time as repeated stress is applied. Division and growth rates are central components of cellular identity. Moreover they provide a measure of single cell fitness which is the missing piece allowing single-cell and population measurements to be related in the previously described dynamic representation of time-changing cellular identity. For these reasons, and for the purpose of improving our integrated vision of the cellular response to fluctuating stress, we decided to study the impact of repeated osmotic stress on cellular proliferation as it will be presented in the next chapter.

# IV. The impact of repeated stress on cellular proliferation

In the general presentation of yeast's response to osmotic stress (see I.4) we mentioned how deep and global is the effect of hyperosmotic stress on cellular activity. In fact osmotic stress and cells' subsequent adaptation have an impact that is not limited to the HOG pathway and glycerol production. For instance, the brutal change in size occurring during a sudden hyperosmotic shock affects the membrane and cell wall and osmotic stress triggers the Cell Wall Integrity (CWI) pathway (69) both after hyper and hypo osmotic shocks. In this chapter, we are concerned with the impact of repeated osmotic stress on a central activity of cells: proliferation. Focusing on the consequence of osmotic stress on this particular aspect of cellular physiology is motivated by several factors.

Proliferation, by means of genome duplication and cellular growth, affects globally the cellular context which is relevant to gene expression. Indeed, as growth drives the dilution of the cell's content, it actually determines one of the key parameters of gene expression dynamics. As long as genes are studied at a coarse-grained level, in isolation and under conditions which do not alter much replication, assuming constant and homogeneous replication may be a fair assumption. Yet, as we try to better understand dynamic changes at the level of gene regulatory networks which are related to fundamental steps of cellular life (like metabolic changes upon nutriment change, adaptation to environmental aggression, differentiation, mating, sporulation, senescence etc.) we will find that all of them come with important changes in growth and cell cycle. Therefore, any quantitative approach to these questions will have to abandon at some point the useful assumption of constant and homogeneous proliferation. In the context of the study of osmotic stress, the investigation of how growth is affected by repeated stress may therefore also possibly help the study of gene expression dynamics and regulation upon stress.

The impact of environmental conditions on proliferation is obviously a very old research topic in biology. Yet, most traditional experiments only measured this impact in stationary conditions. Dynamic changes in environment give access to transient effects. It is generally possible to perform some form of dynamic perturbations in batch and chemostats. Yet, given experimental possibilities with these systems, applying precise and repeated changes is limited in practice to a single change or low frequency perturbation (because of the typical time required to remove a previously added chemical, by dilution or centrifugation). At last, such studies capture only the overall population growth which is insufficient to characterize replication at the single-cell level. Using microfluidics and automated image analysis, we can actually measure proliferation at the single-cell level in a large range of dynamical perturbations frequencies.

In chapter III, we demonstrated that the use of specific temporal patterns of osmotic shocks could both simplify our study of the HOG pathway by making short-term adaptation feedbacks negligible, while providing rich information of single-cell dynamics. In the investigation of the consequences of osmotic stress on proliferation, we will again rely on dynamic stimulations. This serves two purposes: on the one hand, it allows us to explore how cells react to repeated stress which is a fairly uncovered topic (while there is a profusion of studies on single stress). On the other

hand, by repetitively stressing cells at different frequencies we hope to separate cellular processes which have distinct dynamical typical time-scales. For instance, in (55), several typical time-scales were identified for the physical, signaling and transcriptional response of cells to hyperosmotic conditions (from the faster to the slower). Here we will investigate lower frequencies which are relevant for slower components of the cellular response: *i.e.* proliferation and metabolic adjustments. At last, repeated stimulation may allow in theory to amplify transient effects which would be otherwise inaccessible to direct measurements.

# 1. An integrated view of the response to osmotic stress

## a. How osmotic stress affects growth and division?

Sudden osmotic upshift is known to impact proliferation which is stopped or slowed down and resumes once cells have adapted (*i.e.* the HOG pathway is deactivated) (61). Replication involves mostly two connected sets of processes. On the one hand the cell cycle orchestrates DNA replication, budding, nuclear separation and finally cytokinesis. On the other hand *growth* (although the term is sometime used to describe proliferation as a whole) relates more specifically to the production of cellular mass (i.e. the replication of other cellular components than DNA) which is required to *fill* the daughter cell without having the mother cell getting smaller. Cell cycle and growth are obviously connected so as to ensure that cells keep sizes within a viable range. Although how this coupling is ensured is still an active research question, it seems that division accommodates to effective growth but not the opposite (127, 128).

Recent work which will be detailed in 2.a demonstrates that in fact phosphorylated Hog1 directly controls cell cycle arrest. From a general point of view, osmotic stress, via Hog1p, is able to delay or stop the cell cycle in different phases, namely in G1, S and G2. This is believed to allow cells so as to prevent deleterious effects of stress on some delicate steps of the cell cycle and eventually give time for possible repair. By applying repeated stress to cell cultures, we were indeed able to spot one harmful effect osmotic stress can have for cells in M phase for which yeast seems not to have a protection against and which will be described in 2.b.

As it was presented in I.4.a a central aspect of osmoadaptation is the production of glycerol to allow water influx and restore pre-stress size. As it will be presented in 3.a, production of the required quantity of glycerol to allow adaptation is costly, both in terms of energy and in terms of carbon resources. This in turns will impact growth in terms of speed (growth rate) and in terms of yield.

In our study, we will focus on these two sets of mechanisms (direct cell-cycle arrest and altered carbon metabolism) to provide a system level description of the impact of osmotic stress on proliferation. Obviously, other mechanisms should be at play and will be hereby neglected. For instance, indirect effects may influence cell cycle under osmotic stress like the sudden loss of turgor pressure happening in hyperosmotic conditions which is normally necessary for bud formation. Also, osmotic stress triggers several pathways and transcription events which may alter the state of cells indirectly affecting proliferation. An example being the activation by osmotic stress of Sfp1 which controls the production of ribosomal protein (129) and is possibly associated to cell size control (127).

Under a single osmotic stress, both mass growth and cell cycle are affected as the cell produce glycerol (which requires glucose) and pauses its cycle to protect itself against potentially harmful effects. In Figure 55 is represented at an abstract level the two main mechanisms of acclimation[105] to

[105] Acclimation is the phenotypic response of a given organism to a sudden change in environment. It can be reversible to some extent. Adaptation is used more broadly and is often confused with acclimation as it is the case here. Yet, adaptation often refers to long term changes in both genotype and phenotype on several generations.

hyper osmolarity. Importantly, the action of Hog1 is by essence transient as the HOG pathway is inactive in adapted (acclimated) cells.

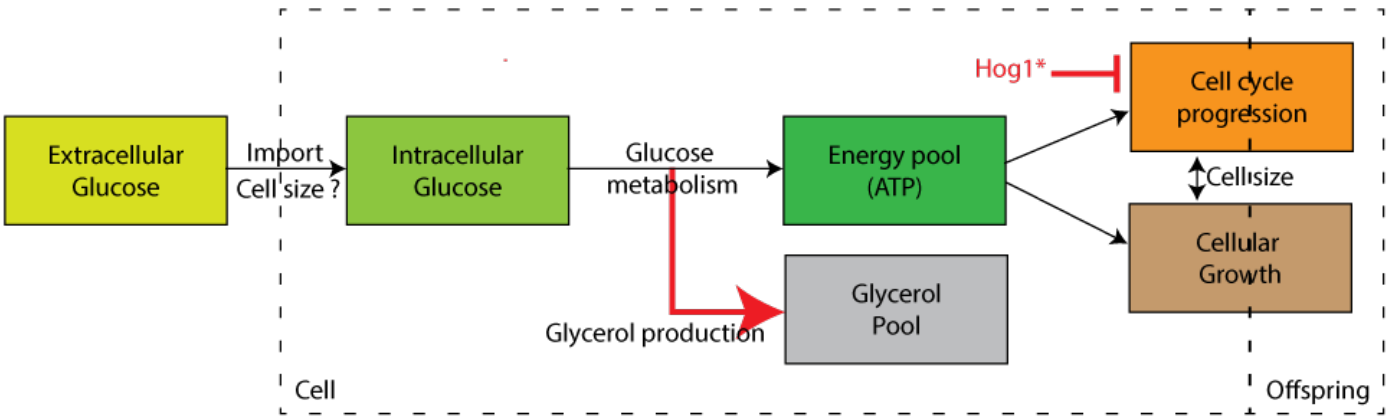



**Figure 55 - Schematic representation of the acclimation to a hyperosmotic stress. Red arrows indicate the modifications due to hyper osmolarity. Hog1* stands for phosphorylated Hog1.**

Once acclimated, proliferating in a hyperosmotic environment requires constantly producing glycerol in order to maintain the proper osmotic balance (turgor pressure being mandatory for budding). Although glycerol can be recycled as a carbon source in practice it will not be the case because it will either be diluted upon daughters divisions or it will be leaked out from the cell to the environment. In fact, *S. cerevisiae* sustained growth is achieved by over producing glycerol and having the glycerol channel Fsp1 acting as some form of pressure-sensitive valve which leaks the surplus out. This situation is depicted in Figure 56.

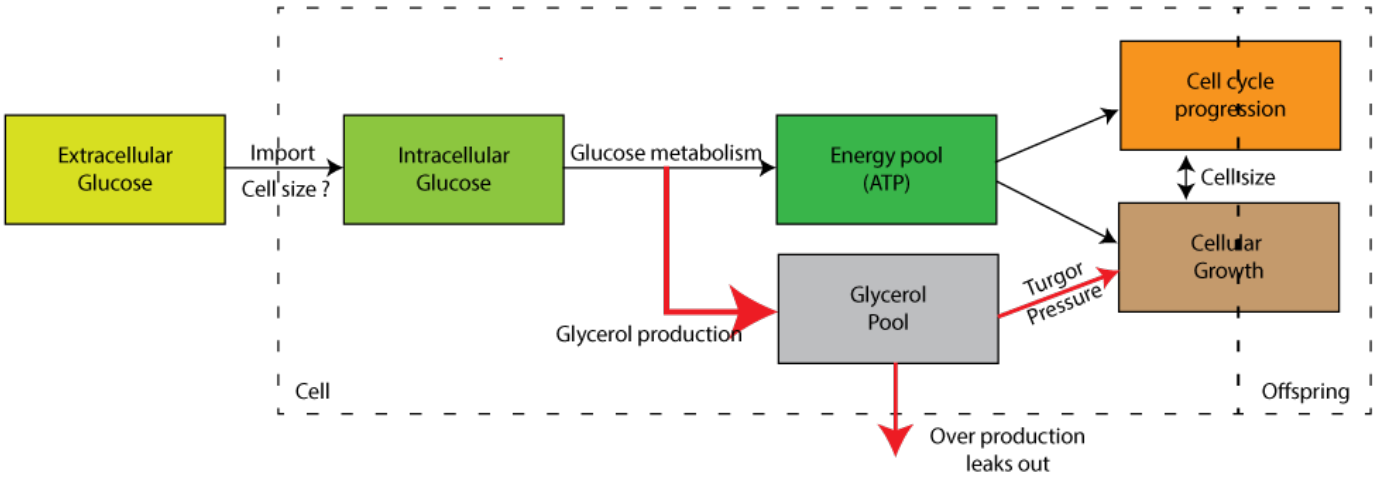



**Figure 56 - Schematic representation of modifications related to proliferation in hyperosmotic environment.**

Both transient and persistent effects of hyperosmolarity are at play and hardly distinguishable when a single osmotic up-shift is applied. Accordingly, it is usually by using various kinds of mutant strains that the current knowledge on the impact of osmotic stress on proliferation has been

gathered. Despite their indispensable contribution to eliciting the molecular mechanisms at play, the study of mutants in single upshifts hardly allows assessing quantitatively and dynamically what is at play in wildtype cells.

### b. Proliferation quantification at the single cell level: a matter of point of view

As observed for gene expression, when considered at the single-cell level, proliferation displays a richer behavior than what is accessible from population-based measurements. In most studies, proliferation is quantified using a single parameter: a population growth rate. Because cells grow at different speeds, single-cell metrics of proliferation are needed.

We presented original image analysis methods that allow quantifying proliferation at the population level (II.3.a) and at the single-cell level (II.3.b). In the proposed single-cell measurement of division rate, the focus is on a cell which undergoes several cell cycles. Accordingly, we quantify the average time it takes for a cell to divide over several divisions. As we can see in Figure 57, there is an important heterogeneity in individual average division time with some cells dividing on average two times faster than others.

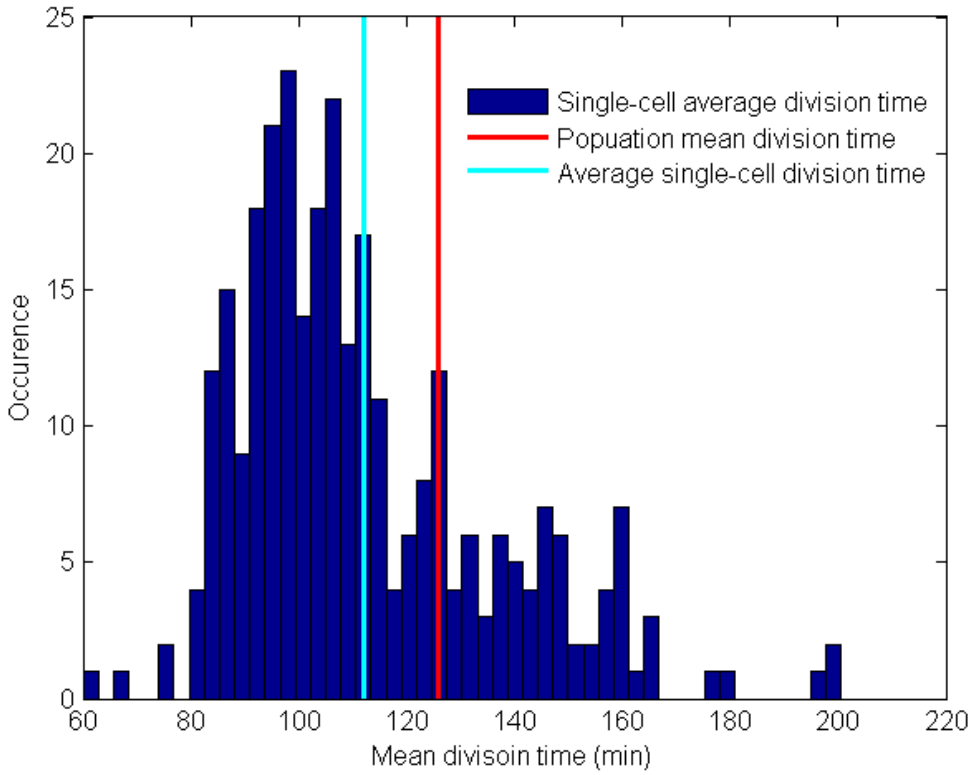



**Figure 57 - Comparing estimation of cell proliferation in population or at the single cell level.. Experimental data from 041013.**

From these single-cell average division times, we can compute an overall average division time by averaging over all cells which is represented by the light blue line in Figure 57. If we compare such average to the average division time as computed from population data (red line in Figure 57,

estimation details in II.3.a) we find dissimilar values[106]. The observed discrepancy comes from the different representations of cellular proliferation which we used. A population division time or division rate does not provide a precise representation of proliferation at the single cell level. Only if all cells had an identical division rate that the population measurement would be applicable to single cells.

If we see the population representation of Figure 57 as a static picture of cells division speed (with fast and slow cells) the fact that single-cell average division time is smaller than the population average division time is surprising. In fact, fast cells should divide more during a given experimental time window and therefore, population growth rate should be determined primarily by these fast dividing cells. To resolve this seemingly strange result, we need to take into account not only the fact that cells are different, but also the precise definition of what our system and our measurement are. In fact, because we consider here mostly cellular identity, we look at cells for several cell cycles. Collecting data for a given number of cells which we *follow* during an experiment amounts to studying a cohort of cells. Studying cohorts is convenient from an experimental perspective[107] and also puts the focus on cellular identity. In order to relate our population measurement with the average division time of a cohort of cells as plotted here, the individual history of cells matters. As *S. cerevisiae* undergoes an asymmetric division, newborn cells are smaller and tend to be at first slower to divide than their mothers. This feature of budding yeast division is represented in Figure 58.

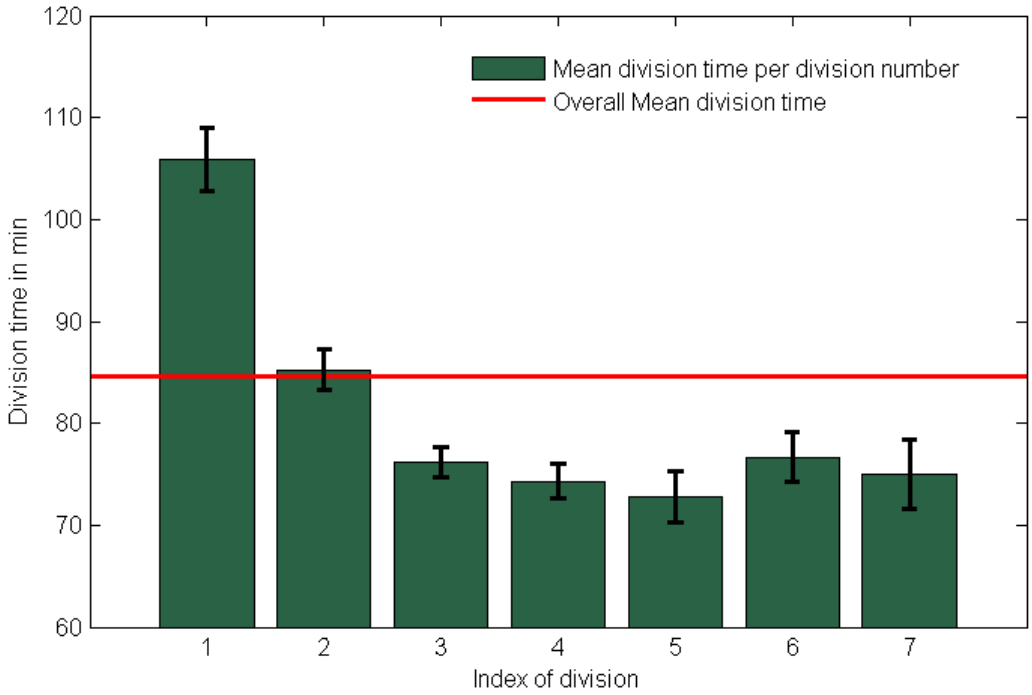



**Figure 58 - Average division time given for the first 7 divisions of a newborn cell. Here average is done over all recorded division events for a given division index. Data extracted by hand from bud appearance on 50 cells (196 divisions) from experiment 041013. Error bars represent the standard error on the mean.**

[106] Note that using other metrics such as the geometric mean, or median leads to similar discrepancies.
[107] Computing a single cell average growth rate requires a cell to be observed for more than one cycle. Note that cells in our cohorts do not have the same age.

This dissection of division time by groups of cell of the same age unveils the structured (in terms of age) aspect of a growing yeast population. This measurement confirms the statement made in (130) that 4 generations should be taken into account in order to have a faithful representation of the age structure in a budding yeast population since after that point, division time seems to be stable (up to a certain point where aging impacts division time). Going back to the apparent paradox between population and single-cell estimates of division rates, we can now see that any cell which divides gives birth in fact to a slow cell. This means that at any time, half the population is composed of slow cells. When we pickup cells which we then follow in time, we distort the instantaneous picture present in a population. This selection process allows performing measurements on what a *typical* cell is during its life, but in terms of quantitative representation of a population, it suffers from a bias towards old cells.

In conclusion we see here that quantifying replication at the single-cell level may yield different results than population based measurements. The precise definition of the studied system (a collection of cells considered during one division only, a cohort of cells undergoing several cycles) along with the selection process (constructing cohorts controlling for age structure or at random, building cohorts by age or not) influence the results we can obtain and the conclusion we can draw from them. At last different metrics used to quantify proliferation (*e.g.* division rate, mean cycle time, growth rate, volume or mass doubling rate) are only equivalent (*i.e.* can be deduced one from the other) for simplistic models of proliferation. Considering dynamical changes and/or cell-to-cell variability will usually require more complex representation of proliferation which can involve other quantities such as cell volume, volume at division, cell mass, division asymmetry parameters or age structure in a population.

## 2. The impact of osmotic stress on the cell cycle

### a. Osmotic stress can trigger phase-dependent arrest of the cell cycle

The rapid change in physiology following a hyperosmotic stress can have harmful effects on cell proliferation. Among others, water potential is altered, therefore modifying biochemical reaction rates. Given that the cell cycle requires a precise sequence of biochemical reactions to be performed in the correct order, altered water potential may lead to errors in completing the cell-cycle program. The cellular response to osmotic stress itself (for instance the rapid transcription of hundreds of genes) can be harmful (131). In addition, budding requires a high-enough turgor pressure, a condition which cannot be met unless the cell has recovered its size and accumulated enough glycerol. In order to prevent dangerous progression in the cell cycle before adaptation has occurred, several direct control mechanisms of the HOG pathway on cell-cycle have evolved in *S. cerevisiae*. Recent studies have unveiled some molecular details of such interactions which are summarized in Figure 59.

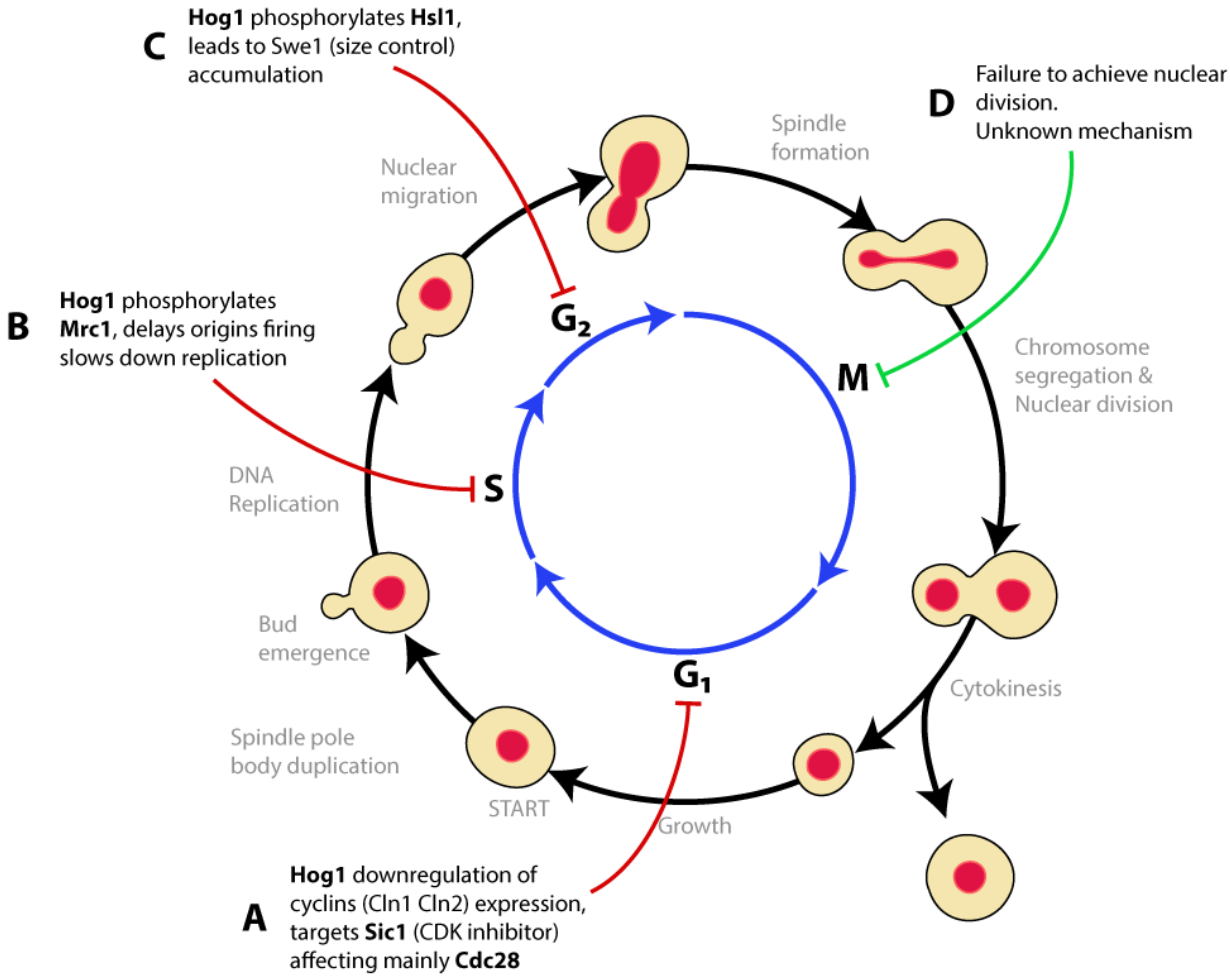



**Figure 59 - Overview of the direct effects of the HOG pathway on the cell cycle of *S. cerevisiae*. Red arrows represent known mechanisms and the green arrow represents a novel mechanism.**

During G1 (Gap1) the cell increases its mass and volume until it will eventually commit irreversibly to division through a molecular checkpoint called *Start*. The *Start* transition is made only if a certain number of conditions are met (sufficient size, nutrient availability, absence of

pheromones, absence of DNA damage etc.). It is clear that hyper osmotic stress affects cell-cycle regulation by down regulation of the transcription of cyclins CLN1 and CLN2 (the G1/S cyclins[108] in *S. cerevisiae*) (132). Yet, the direct role of Hog1 is a bit controversial in this respect as (132) finds that for normal stress (0.5M KCl in their case) this phenomenon was Hog1 independent, while in (133), activating Hog1 artificially without osmotic stress lead to the same cell-cycle arrest, therefore suggesting CLN1 and CLN2 downregulation is Hog1 dependent. Both agree upon the fact that recovery from such cell-cycle arrest requires functional Hog1. Nevertheless, CLN1 and CLN2 down-regulation is not enough to prevent *Start* (as over expressing them did not removed the delay) but may rather be a consequence of a lesser activity of their main regulator: CLN3 (the G1 cyclin in *S. cerevisiae*) which acts in association with Cdc28 (the only Cdk[109] in *S. cerevisiae*). As the level of transcription of CLN3 (which is always constitutively expressed) is not affected by hyperosmotic stress, other inhibitory mechanisms were expected to be active under osmotic stress. It appears that Hog1 directly interacts with Sic1 which is a Cdk inhibitor[110] and stabilizes it whereas it is normally targeted for degradation by Cln-Cdc28 complexes. This in turns prevents cells under osmotic stress from committing to *Start*. This regulation is depicted as **(A)** in Figure 59.

During the S phase, the cell starts to bud and proceeds to DNA replication. As it was presented in I.4.b, the HOG pathway triggers the rapid transcription of hundreds of genes. This situation is potentially harmful for cells replicating their chromosomes as it can lead to collisions between the replication and transcription complexes. Such collisions are prone to recombination events and accordingly threaten genomic integrity. To reduce the probability of such events, it has been shown that Hog1 can in fact directly delay early and late replication origins' firing (134). This protective action is achieved by phosphorylating Mrc1 which is a member of the DNA replication complex. This phosphorylation is directly attributed to Hog1 and appears to be distinct from other mechanisms affecting Mrc1 as in response to hydroxyurea exposition for instance. This regulation is depicted as **(B)** in Figure 59.

During the Gap 2 phase (G2) the bud grows and the spindle is assembled to prepare mitosis. Hog1 can delay G2 exit to allow the cell to adapt before entering M phase (135). A failure to do so would lead to deleterious effects and abnormal morphology. This control is exerted by Hog1 on Hsl1 (a checkpoint kinase) which in turns leads to Swe1 accumulation. Swe1 is believed to ensure the G2/M size checkpoint and its sustained level inhibits the activity of the Clb2-Cdc25 complex which triggers mitosis. This regulation is depicted as **(C)** in Figure 59.

At last, some influence of the HOG pathway on cell exit from mitosis (M/G1 transition) was reported in (136). Yet, this effect, which is visible in specific mutants and involves Hog1 once again, does not lead to a quantifiable delay in *wildtype* cells. In addition, it is unclear whether this phenomenon happens differentially under osmotic stress compared to isotonic environments. Therefore, the aforementioned impact of Hog1 may well correspond to a function of Hog1 which is

[108] Cyclins are proteins which oscillate during the cell cycle (with the notable exception of CLN3). G1/S cyclins are required for *Start* and proceeding to the S1 phase of the cell cycle.

[109] Cyclin dependent kinases (Cdk) have crucial roles in orchestrating the cell-cycle and their activity is conditioned by their association with cyclins. In S. cerevisiae, there is a single Cdk whereas other organisms have several.

[110] Cdk inhibitors inhibit the activity of cyclin-Cdk complexes. In s. cerevisiae, there are two of them: Sic1 and Far1, the later having been ruled out as a potential actor in cell-cycle arrest by hyperosmotic stress.

separated from its central role in the response to hyperosmotic stress. For these reasons, I did not include such regulation in Figure 59.

Unlike the experiments performed in all of the studies mentioned previously, which used artificially synchronized populations reaction to a single hyperosmotic upshift, we systematically studied the reaction of unsynchronized cultures to repeated osmotic stress of varying frequency. As it will be presented more precisely in the next paragraph, we observed that when stressed at a precise moment during mitosis, yeast cells have a risk of failing to divide their nucleus properly among daughter cells. This can cause a delay in the progression of the M phase as the cell try again to separate its genetic material among daughter cells. Interestingly, such even can in rare cases induce polyploidy, aneuploidy or have lethal effects. This impact of hyperosmotic stress on the cell cycle is depicted as **(D)** on Figure 59.

### b. Nuclear separation is perturbed by osmotic stress

Within the M phase of the cell cycle, the genetic material is separated among mother and daughter cells. This crucial part of mitosis is usually a relatively brief event. Applying repeated osmotic stress on yeast strains harboring a nuclear tag, we observed that stress could significantly delay the M phase with cells repeatedly trying to separate their chromosomes for hundreds of minutes (therefore it seems that such issues happen during anaphase although it is difficult to discern exactly phases composing the M phase). In Figure 60 we report a photomontage of an experiment[111] where periodic phases of high osmolarity were applied (1M sorbitol for 45 min followed by 45 min of isotonic medium). The depicted cell is in anaphase when the first hyperosmotic stress is applied. Whereas usually nuclear division lasts approximately 10 minutes, it will take more than two hours for this cell to finish its cycle as it is stressed again 90 min after the initial stress.

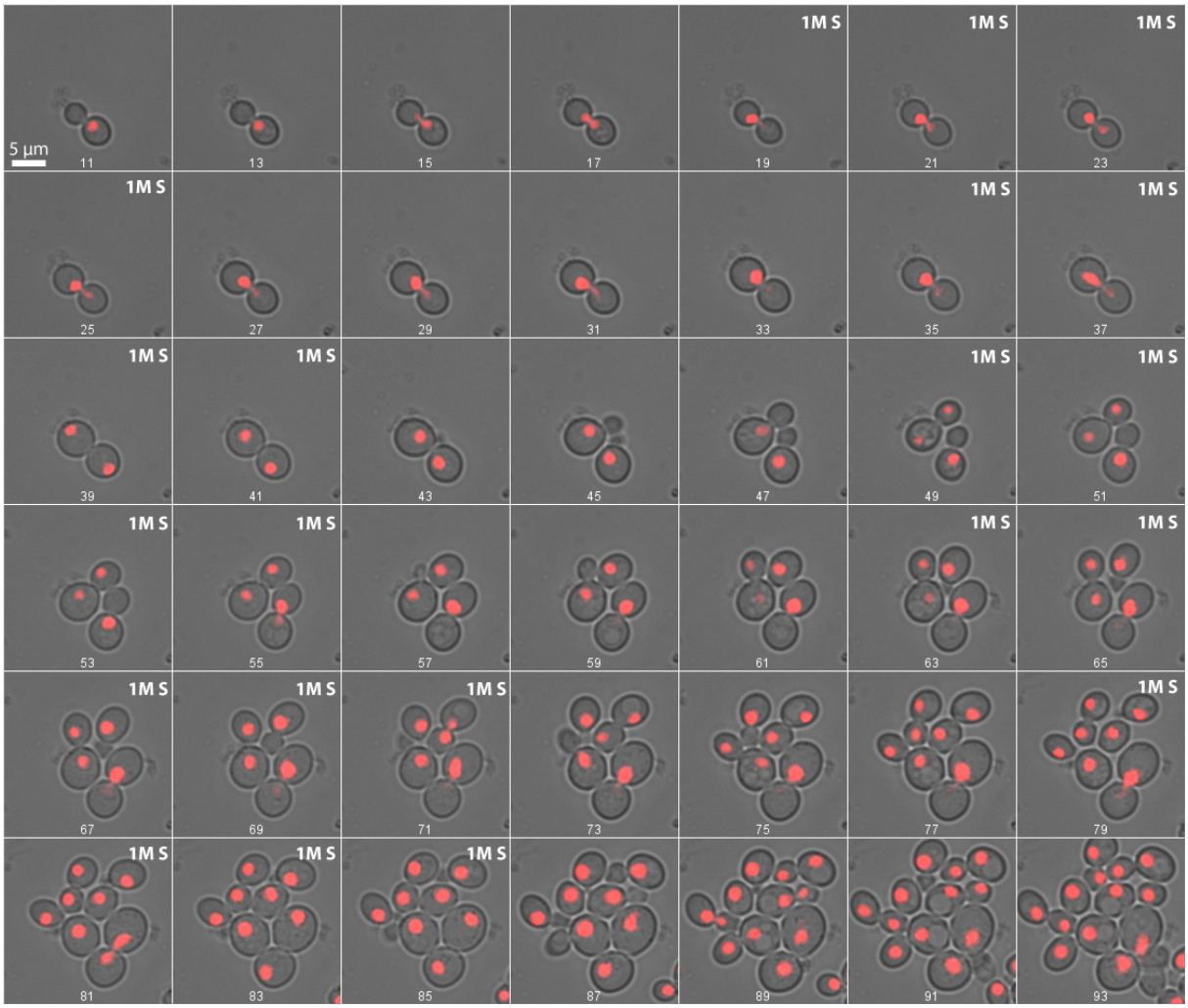



**Figure 60 - Photomontage showing a cell stressed during the M-phase which endures a significant delay in nuclear separation. Note that the mother cell will experience again difficulties for its next division whereas its daughter cell is not affected. The frame number is indicated in white, frames being taken every 6 min (therefore there is 12 min between each vignette). Periods of hyperosmotic stress are indicated with a 1MS white label in the top right part of a vignette. See text for experimental conditions.**

[111] Experiment 180215 used strain yPH15 which has a HTB2-mCherry nuclear marker. Cell from exponential cultures were let to grow in a H-shaped microfluidic device with SC medium (2% glucose) for 105 min before subjecting them to repeated phases of osmotic stress (1M sorbitol SC medium) and normal medium of 45min each.

While such events usually end up by cells finally dividing, as the cell followed in Figure 60, it can also lead to many deleterious effects. For instance, such events can lead to ploidy anomalies. In Figure 61, we show an example taken later from the same experiment (cells being unrelated). In this case, the mother cell of Figure 61 had been stressed in late G2/early M phase at the beginning of the experiment. It displayed an abnormal M phase but managed to divide several times until it divided leaving a bud nearly empty of genetic material (green arrow in the first vignette of Figure 61). During its next M phase we can observe that the genetic material is no longer well separated as three distinct clusters of chromosomes are visible (frames 94 to 98, blue arrows Figure 61).

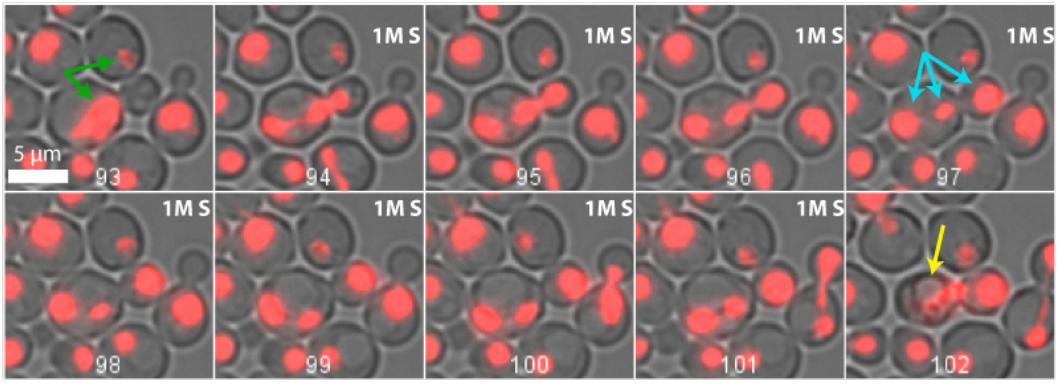



**Figure 61 - Photomontage showing a cell displaying ploidy anomaly (blue arrows) and which will finally die upon hypo-osmotic stress (yellow arrow). The cell shown here had previously endured difficulties in performing M phase which ended in budding a cell with only a very little portion of regular genetic material (green arrows in the first vignette). The frame number is indicated in white, frames being taken every 6 min. Periods of hyperosmotic stress are indicated with a 1MS white label in the top right part of a vignette. See text for experimental conditions.**

Although we also observe ploidy defects which are non-lethal, they often lead to altered morphologies (cells being usually larger than the average), complete stalled or extremely slow cell cycle or cell death as it is the case in Figure 61 (yellow arrow).

We wonder what dysfunction (or safeguard mechanism in the case of delayed yet recovering cells) is at play in such anomalies. It seems plausible that mechanics of chromosome repartition are altered by osmotic stress. This could come from the impact of osmolarity on microtubules which play a central role in chromosomes separation. In fact, studies carried on *S. pombe* showed that osmotic stress could lead to severely altered microtubule function, including during M phase (137). Cytoskeleton dysfunction could have a mechanical or a molecular origin. Another hypothesis would be an influence of osmotic stress on the APC (anaphase promoting complex) and its action on securins and cohesins which hold sister chromatids together because the reported anomaly shows chromosomes stuck at the equator.

Adaptation to the level of stress used in our experiments takes around 15 min (and therefore osmolarity is rapidly balanced and the HOG pathway is deactivated) while we witness much longer delays. Therefore it appears that whatever mechanism is behind this M phase arrest, it require some time for the cell to recover once the triggering stimuli has been removed. To gain more insight on this phenomenon so as to narrow down the range of possible explanations, we tested if such anomaly was related to the duration of stress or to the frequency of stress. We quantified the

number of different M-phase anomalies happening in different dynamical stress conditions. Sensible delay in nuclear division is called a M phase anomaly. A cell showing an abnormal number of chromosomes clusters such as in Figure 61 is called a Ploidy anomaly and observable cell death is called a lethal anomaly.

In Table 1 we report our results concerning three experiments conducted in similar conditions (H-shaped microfluidic device, using the yPH15 strain, constantly flowing fresh SC medium without or without 1M sorbitol and using the same imaging settings). As a negative control we quantified the number of anomalies occurring in a growing population in absence of osmotic stress. Then, we considered two experiments where we repeatedly stress cells. Because we apply symmetric repeated stress (*i.e.* the hypertonic and isotonic phases are of equal durations), cells spend an equivalent proportion of time in hyper osmotic medium. Because anomalies are a fairly rare events we need many cells and divisions to observe enough events so as to obtain a meaningful quantification. Observing multiple separate microfluidic channels in parallel and performing experiments of 20h for the no stress and of more than 10 hours for the T=90 and T=20 min conditions, we could observe ~40000, ~6000 and ~1300 divisions[112] in no Stress, T=90 and T=20 conditions respectively.

| | No Stress | T=90 min | T=20 min |
|---|---|---|---|
| Observed M phase anomalies | 27 | 30 | 49 |
| Observed Ploidy anomalies | 0 | 2 | 13 |
| Observed Lethal anomalies | 0 | 2 | 3 |
| Observed division events | 41653 | 5901 | 1251 |
| Average number of cells per frame | 828 | 725 | 146 |
| M anomaly per newborn | **6,48E-04** | **5,08E-03** | **3,92E-02** |
| P anomaly per newborn | 0,00E+00 | 3,39E-04 | 1,04E-02 |
| L anomaly per newborn | 0,00E+00 | 3,39E-04 | 2,40E-03 |
| M anomaly per cell per hour | 6,48E-03 | 3,48E-03 | 1,76E-02 |
| P anomaly per cell per hour | <2,4E-05 | 2,32E-04 | 4,68E-03 |
| L anomaly per cell per hour | <2,4E-05 | 2,32E-04 | 1,08E-03 |
| M anomaly per cell per stress | | **5,21E-03** | **5,87E-03** |
| P anomaly per cell per stress | | **3,48E-04** | **1,56E-03** |
| L anomaly per cell per stress | | **3,48E-04** | **3,60E-04** |

**Table 1 - Quantification of stress-induced anomalies in M phase. Based on anomaly counting by visual inspection of Experiments 140214 180214 and 020813 (SC medium, 2% Glucose, 0 or 1M sorbitol). T stands for the period of hyperosmotic repeated stimulations which consists of T/2 time in 1M sorbitol followed by T/2 in isotonic medium. In later rows of the table, P stands for Ploidy and L for lethality See text for more explanations.**

We found that an anomaly in M phase occurred on average once every ~1500 divisions (Table 1, M anomaly per newborn). Such anomalies are rather mild as we could not witness any ploidy or cellular death when no stress was applied. The rate of a given anomaly per division or newborn is simply the number of recorded anomalies divided by the total number of births estimated on an

[112] The number of divisions is estimated by segmentation and tracking of cells nuclei which are marked with the fluorescent HTB2-mCherry tag.

overall experiment (all individual channels being pooled). When we consider cells being stressed, we observe a nearly ten-fold increase in the rate of anomaly per division for T=60 with an anomaly every ~200 divisions approximately and an even higher probability of anomaly in intense repeated stimulations (T=20) with one anomaly every ~25 divisions (Table 1, M anomaly per newborn).

Although the number of anomaly per number of division gives a good vision of the impact of this phenomenon at the single cell level (because it compensates for variable division rate between conditions), we tried to compute metrics more adapted for a synthetic view of anomalies occurring at the population scale. Accordingly we estimated the probability for any given cell to experience an anomaly within an hour (Table 1, anomaly per cell per hour). From this, we can factor in the frequency at which we impose osmotic stress which yields the rate of anomalies per cell and per stress (more precisely, per period T which is composed of T/2 stress followed by T/2 isotonic environment). In Table 1, (anomaly per cell per stress) we can see that in fact, the probability of M phase anomalies per stress and per cell is very similar for T=90 and T=20. This in turns calls for a mechanism which depends mostly on the number of stress received by a cell. The higher number of anomalies in T=20 compared to other conditions when measured per unit of time or per division appears therefore as originating from the combined effects of imposing a larger number of stress events per unit of time and having a lower division rate.

Although we could not observe enough Ploidy and Lethal anomalies in mild or no stress conditions to be conclusive, it appears that severe defects are more likely to happen in frequent stress. This could mean that following an initial anomaly, it is additional stresses in a given time window which triggers deleterious effects. At last, it should be considered that ploidy modifications, although possibly hazardous for cells are paradoxically also an interesting potential mechanism to evolve important mutations quickly in response to various stress, including those induced by NaCl (138, 139).

### c. Timing of cell cycle arrest and partial lock-in

Many published experiments reveling cell-cycle delays in response to hyperosmotic stress were usually conducted on synchronized cultures (using α factor release or nitrogen deprivation for instance) with usually low time sampling. Synchronization is a necessity for most molecular biology assays but may lead to artifacts when it comes to retrieving dynamic information. For example, several experiments in the literature were conducted on mutants conditionally activating the HOG pathway using temperature based inducible genes therefore applying a heat stress. Some rough estimates of delays can still be obtained and are reported in Table 2

| Phase affected | Measured delay | Experimental condition | Reference |
|---|---|---|---|
| **G1** | 30 min | 0.5M KCl batch population growth | (132) |
| **G1** | 30-50 min | Synchronized cultures (α factor release) wildtype vs 0.5M KCl | (132) |
| **G1** | 40 min | Synchronized cultures (α factor release) wildtype vs 0.4M NaCl | (133) |
| **S** | 20 min | Temperature sensitive mutant : 25°C - α factor release – 37°C synch and release at 25°C. 0.4M NaCl | (134) |
| **S** | 40 min | Same procedure as over, 0.8M sorbitol | (134) |
| **S** | 10-30 min | Delay for replication origin firing | (134) |
| **G2** | 80-100 min | Synchronized cultures (α factor release) wildtype vs 0.4M NaCl 50 min after release. | (135) |

**Table 2 - Literature values for cell cycle arrest in response to hyperosmotic stress.**

Besides possible bias coming from experimental methods, cell cycle delays have always been measured in single osmolarity upshifts. Therefore we do not know if once triggered, the molecular mechanisms involved in these delays will proceed alone or if sustained Hog1 activation is necessary. Yet, we know on the opposite that for constitutively activated Hog1, cells seem not to divide at all (133) so the G1 arrest can be maintained as long as a cell is not adapted. At last, these measurements only give information on the average delay but not on the cell to cell variability in this delay.

In order to quantify the delay induced by an hyperosmotic stress at the single cell level, we performed an experiment (220715) were cells bearing a nuclear fluorescent tag (strain yPH15) were grown in a microfluidic chip in 2% glucose SC media constantly renewed. After 300min of free growth, a short (15 min) osmotic stress was applied using the same media with an addition 1M sorbitol and cells were imaged for another 300 min. Our custom lineage detection algorithm

presented in II.2.b was used to retrieve division instants (defined as the time of nuclear separation) for 218 cells (442 divisions).

We then sorted all the observed cell cycles in four categories. *Pre-stress* concerns cell cycles which have ended before the osmotic stress was applied. *On stress* regroups cell cycles which have some overlap with the applied stress. *Post Stress 1* contains the cell cycles which have begun after the osmotic stress was removed but which immediately followed the stress. At last, *Post Stress 2+* concerns all subsequent cell cycles. As we see on Figure 62, a 15 min stress leads to a ~50 min increase in the average cycle time, a value which is coherent with the literature values reported in Table 2. Since when we return to an iso-osmotic medium the HOG pathway is deactivated very shortly, we see here that once arrested the cell cycle requires some time to resume.

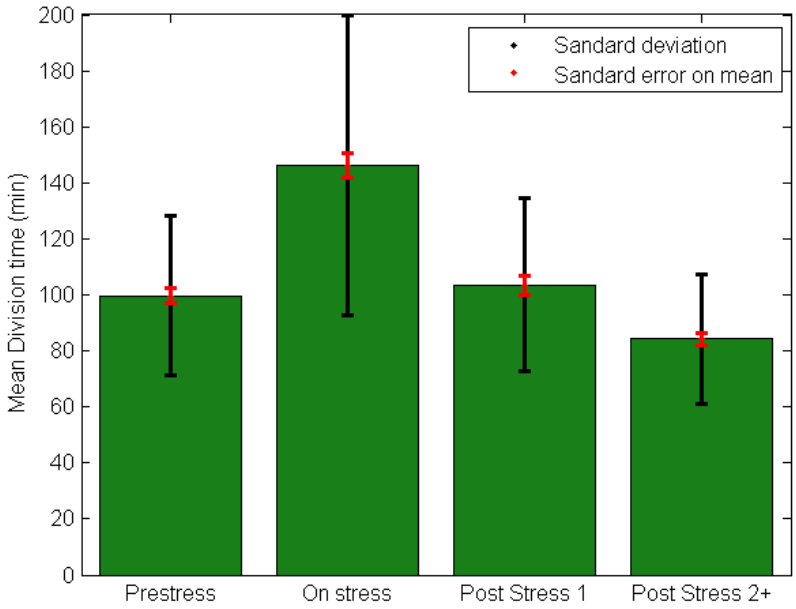



**Figure 62 - Average cell cycle duration before, during and after a short hyperosmotic stress. Average is performed on all cells without discrimination.**

***Influence of age on cell cycle delay***

Considering the important standard deviation in cell cycle durations reported in Figure 62, we additionally distinguished cells' cycles based on their age to see if this variability comes from age differences between cells. In other words we distinguish the first cycle of a newborn from the first cycle of a mother (mother n°1), the second cycle of a mother (mother n°2) etc. (see sketch in Figure 63). As we see on Figure 63, new born cells have significantly longer cell cycle times when stressed than mother cells. Yet as newborns are always slower than mothers, we computed the time difference between the normal division time as in pre stress conditions, with that of under and post stress. As depicted in Figure 64, we find that the length of the average delay is actually similar for all ages, amounting to approximately 50 min.

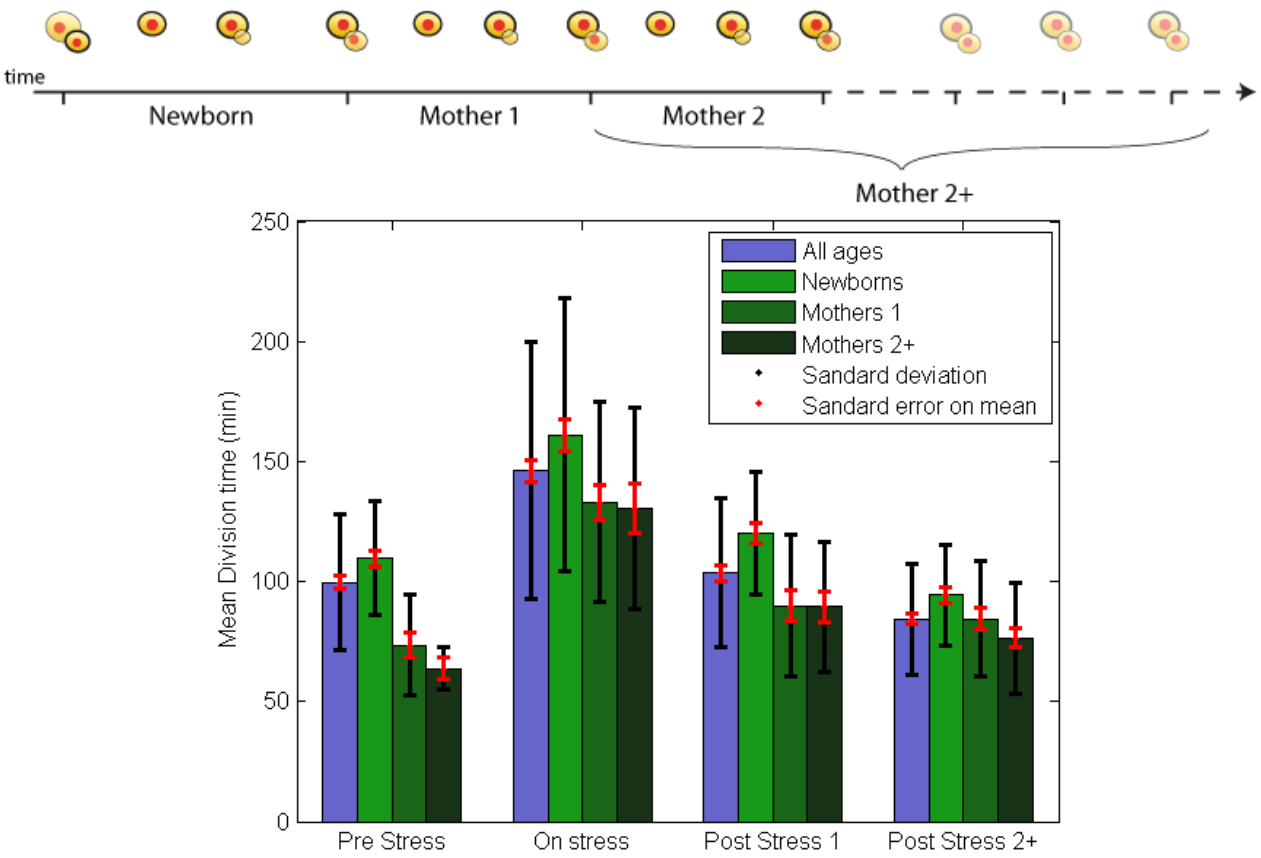



**Figure 63 - Comparative effect of age on the average cell cycle duration before, during and after a short hyperosmotic stress. The average for a given experimental period (*e.g.* post stress 1) and a given age class is derived from the cells that had the age of the given age class during the given experimental period. Age classes are depicted in the sketch.**

We can also note that older cells have a slightly longer delay (Figure 64) which in turns means that relatively to their normal cell cycle length, a fixed delay represents a much more important slowdown than for newborns as visible in Figure 65. In fact for a mother cell, the delay upon osmotic

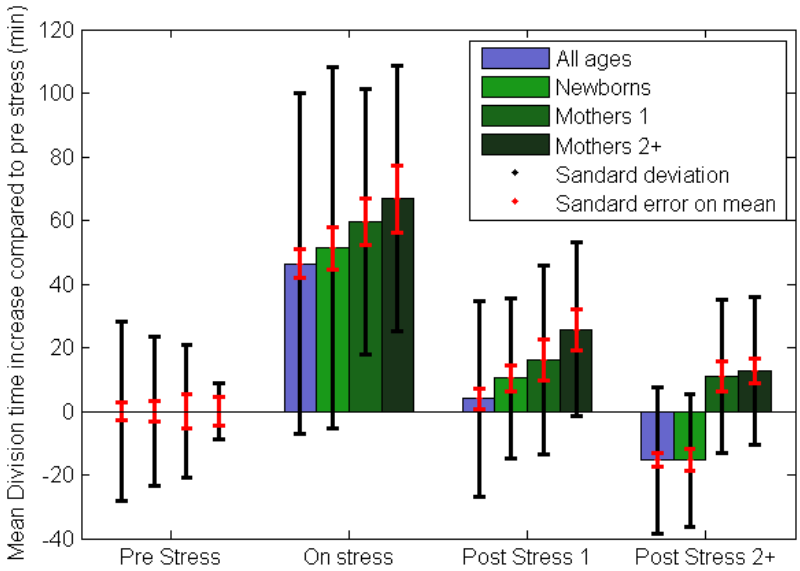



**Figure 64 - Difference in division time relative to pre-stress condition for several age classes.**

stress can get as long as its normal division time. Interestingly, mother cells seem to require more than one cycle to recover their normal cell cycle time. This could be seen as if cells had some form of inertia in their proliferation. It would be interesting to investigate more precisely if such inertia is related to the cell cycle or to cellular growth. In other words, is inertia due to properties of the biochemical network governing cell cycle or is it rather a consequence of metabolic adaptations?

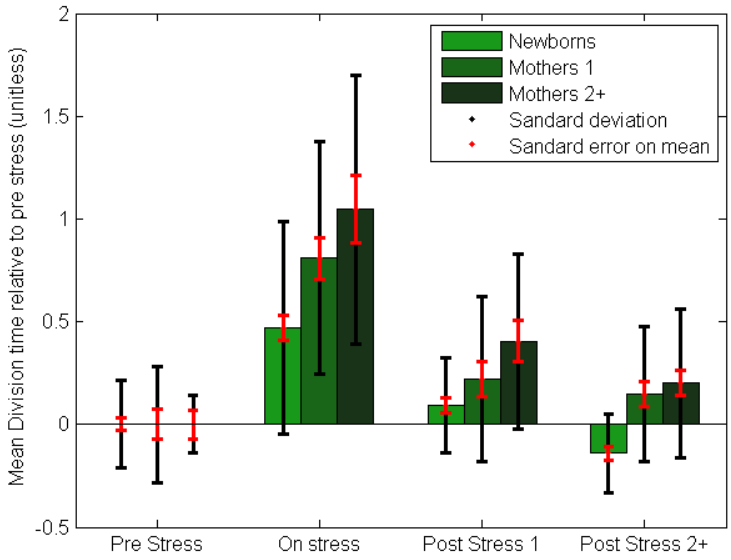



**Figure 65 – Relative variation in division time relative to pre-stress condition for several age classes.**

***Influence of cell cycle position on cell cycle delay***

Instead of proceeding to a synchronization of our yeast culture, we use single-cell information to perform a virtual synchronization a posteriori of the cells present in our experiment. We divide the duration of each single-cell division in five portions with equal duration. By considering the average duration of each phase of the cell cycle for newborns and mothers[113] (140), we can propose the following approximate mapping:

- For mother cells: 0-20% :G1, 20-60% S, 60-100% G2/M
- For newborns: 0-40% G1, 40-70% S, 70-100% G2/M

In Figure 66 we show the delay and relative delay of cell-cycles perturbed by osmotic stress compared to pre-stress conditions. In contrast to the indicative values from literature in Table 2, we find the effect of stress to be the most important when occurring in the late G1, early S phase.

[113] We round the estimates of cell cycles phases to be 20 min, 40 min, 40 min (G1, S, G2/M phases respectively) for mother cells and 60 min, 40 min, 40 min for daughter cells.

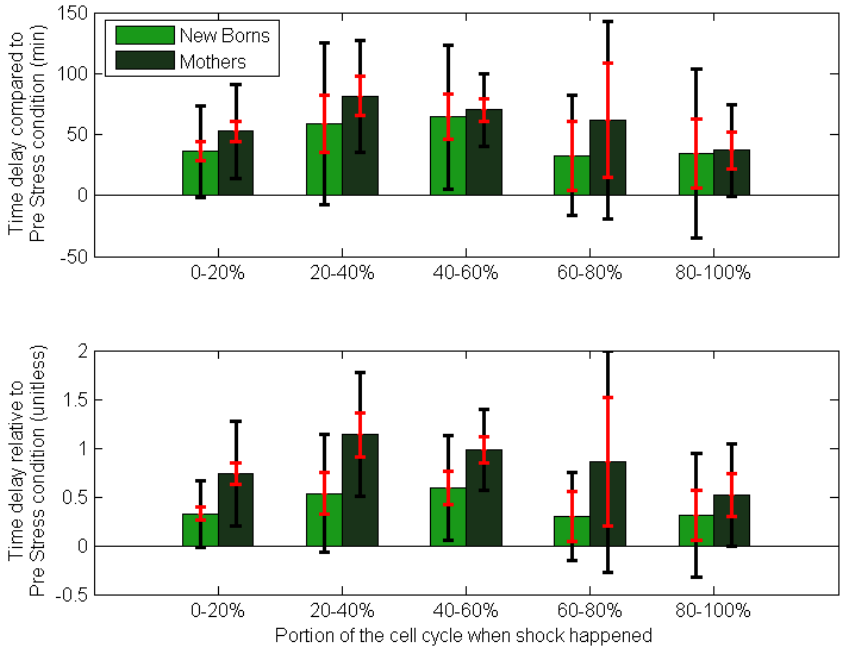



**Figure 66 - Impact of cell cycle position upon osmotic stress induced cell cycle delay.**

### d. Lock-in phenomenon

When a naturally oscillating dynamical system is perturbed with a periodic input, lock-in phenomenon can appear. Lock-in happens when the oscillating system frequency and phase (with a possible lag) will progressively change to match those imposed by the external stimulus. The cell cycle is obviously a natural oscillator. Since osmotic stress can delay the cell cycle, we can expect a population of initially desynchronized cells to be forced into synchrony by applying a regular pattern of stress.

In Figure 67 we show that the division rate of a population of cells oscillates in phase with a periodic stimulation with 1M sorbitol having a period (90 min) close to the cell cycle duration. We used a strain bearing a nuclear tag and window-based methods (see II.3.a) to estimate the instantaneous population division rate (black thin curve in Figure 67). We smooth this noisy[114] curve using a sliding window averaging method where the smoothed signal at time *t* is the geometric average of the signal over a symmetric time window of 54 min centered on *t*. Once smoothed once or twice (green dotted and black thick line respectively in Figure 67) we can observe that indeed, in a fluctuating environment, cells at least partially synchronize and that they tend to divide less in high osmolarity as expected.

[114] We see that with time, the instantaneous division rate becomes less noisy. This is because as our fields of views fill with more cells more divisions occur at each frame.

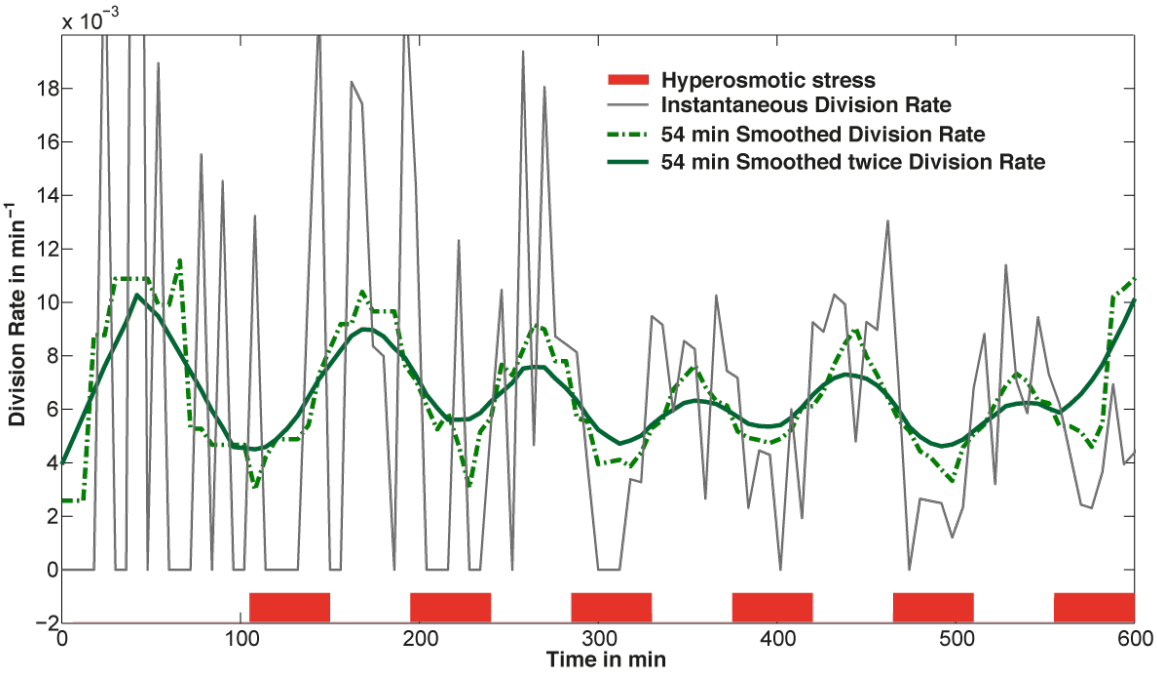



**Figure 67 - Time evolution of the division rate of a population stimulated by periodic osmotic stress (1M sorbitol) with a period of 90 min. Osmolarity within a period is 1M sorbitol for half the period (red pulses) and 0M sorbitol for the remainder. Data from experiment 180214.**

Note that here we used a geometric average because instantaneous division rates are subjected to compounding (like interests) and using a geometric mean produces an equivalent division rate over a given time period (the effective difference with an arithmetic average being more pronounced as many frames are averaged).

In Figure 67 we show synchronization occurring with an input frequency which is fairly close to the natural division rate of mother cells in a standard culture medium. Smoothed rates improve the visualization of such phenomena, yet given the sampling time, it makes it difficult to assess properly the importance of the lock-in in terms of amplitude of the forced oscillations. When we look at the unsmoothed division rate at late time points, we see that the amplitude of forced oscillations is far larger than for the smoothed curve and that in the middle of the hyperosmotic period, division is almost zero.

In order to quantify the amplitude of lock-in, we performed a spectral analysis (using *fft*, fast Fourier transform) of the raw instantaneous division rate of cells subjected to periodic hyperosmotic stress with different frequencies[115]. To compare spectra across different conditions having different average division rates, we centered and scaled the raw signal. Because the average division rate can change slowly during an experiment (an evolution we do not want to capture in this lock-in analysis), we used a 114 min geometric averaging sliding window smoothed division rate for normalization. So

[115] All experiments were carried in microfluidic devices, using strain yPH15 with 2% glucose SC medium with or without 1M sorbitol. Applying an input with period T means that after a resting period of 1 to 2h, cells are repeatedly in hyperosmotic medium during T/2 and in isotonic medium during T/2 for 10 to 20h.

the signal used in fft is the relative variation of the instantaneous division rate compared to the slow time-averaged division rate. Resulting power spectrum densities are reported in Figure 68.

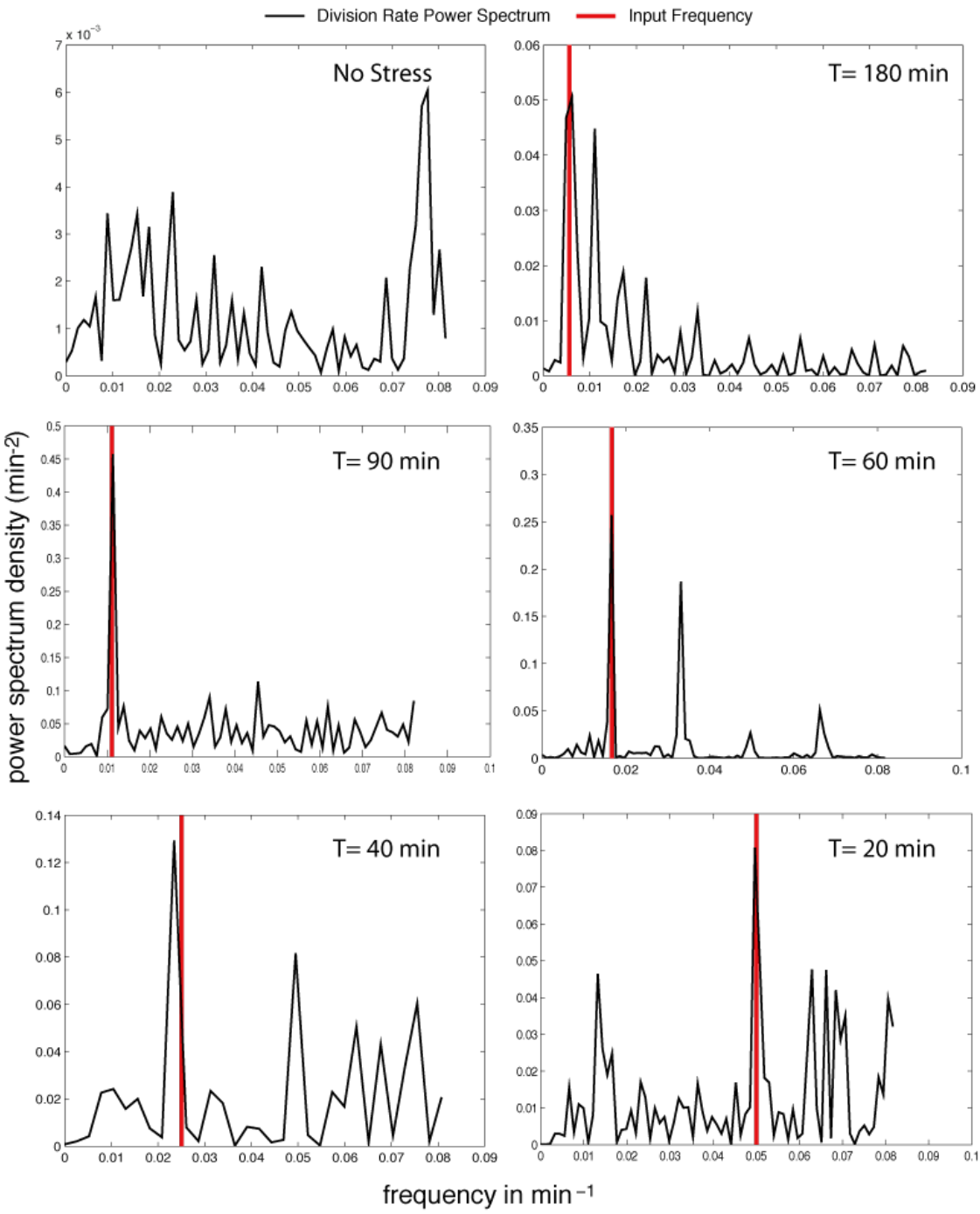



**Figure 68 - Power spectrum density of the instantaneous population division for different experiments where cells are subjected to periodical hyperosmotic stress with different, fixed frequencies. For each panel, the osmotic stress input period is indicated in top right corner and the corresponding frequency in indicated by a solid red line. Note that scales are different across panels. See text for details.**

Observing the frequency content of division rate, we observed that in unstimulated growth (top left panel in Figure 68) only residual frequencies are present (note the $10^{-3}$ factor for the power)

with a relative peak corresponding to twice the sampling frequency (a common sampling artifact). This serves as a baseline of the spectral *noise* we can expect in absence of any stress. It should be noted that as we consider here the overall division rate of a non-synchronized population, we cannot expect to find more than traces of the natural cell cycle frequency.

As it is clearly visible in all conditions the population division rate spectrum shows a significant peak at the input frequency (red lines in Figure 68). This indicates that there is always some synchronization occurring. Yet, in terms of absolute power, we see that the peak corresponding to the input frequency does not carry the same power across conditions as represented by the histogram in Figure 69.

Not only we observe a peak at the input frequency in the spectra of Figure 68, but we can often observe smaller peaks regularly spaced in the spectrum. It appears that in some case, these secondary peaks correspond to harmonics of the input frequency (this is particularly clear for T=60 min in Figure 68). Such harmonics are probably the signature of a shape closer to a triangle or a square signal than a sinewave as qualitatively visible at late time points in Figure 67.

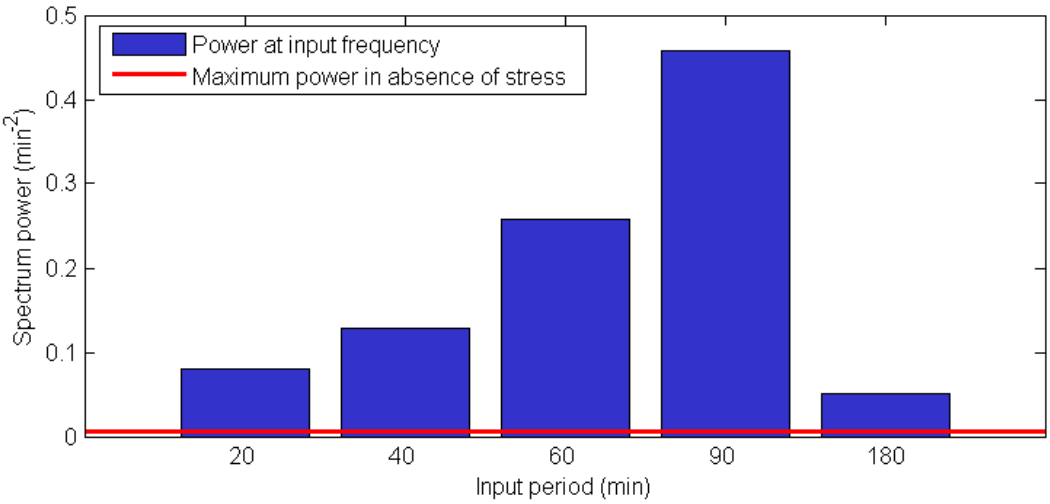



**Figure 69 - Spectral power of the input frequency for different periodic stress experiments. The red thin line represents the maximal power recorded in an experiment with no stress and serves as a baseline.**

As it could be expected, we see that lock-in is the most efficient when we try to synchronize cells at a frequency close to the natural division time. Yet it is surprising to see that at T=180 min the synchronization factor is even smaller than for high frequencies. Future experiments mapping intermediate periods between 60 and 180 min could allow us to determine if we can achieve even higher lock-in and find the *resonance spectrum* of division oscillations. At last, performing a single cell analysis will make it possible to quantify to what extent lock-in is due to changes in cell cycle durations or to synchronization of cell cycles among cells.

# 3. The impact of osmotic stress on metabolism

## a. Metabolism shifts upon osmotic stress

In hyperosmotic conditions, cells produce glycerol from Dihydroxyacetone phosphate (DHAP) in a two-steps reaction, catalyzed by two pairs of isoenzymes, Gpd1/2 and Gpp1/2 as we can observe on the simplified sketch in Figure 70. This implies that glycerol production affects the pool of intermediary metabolites used for energy production via the TCA cycle.

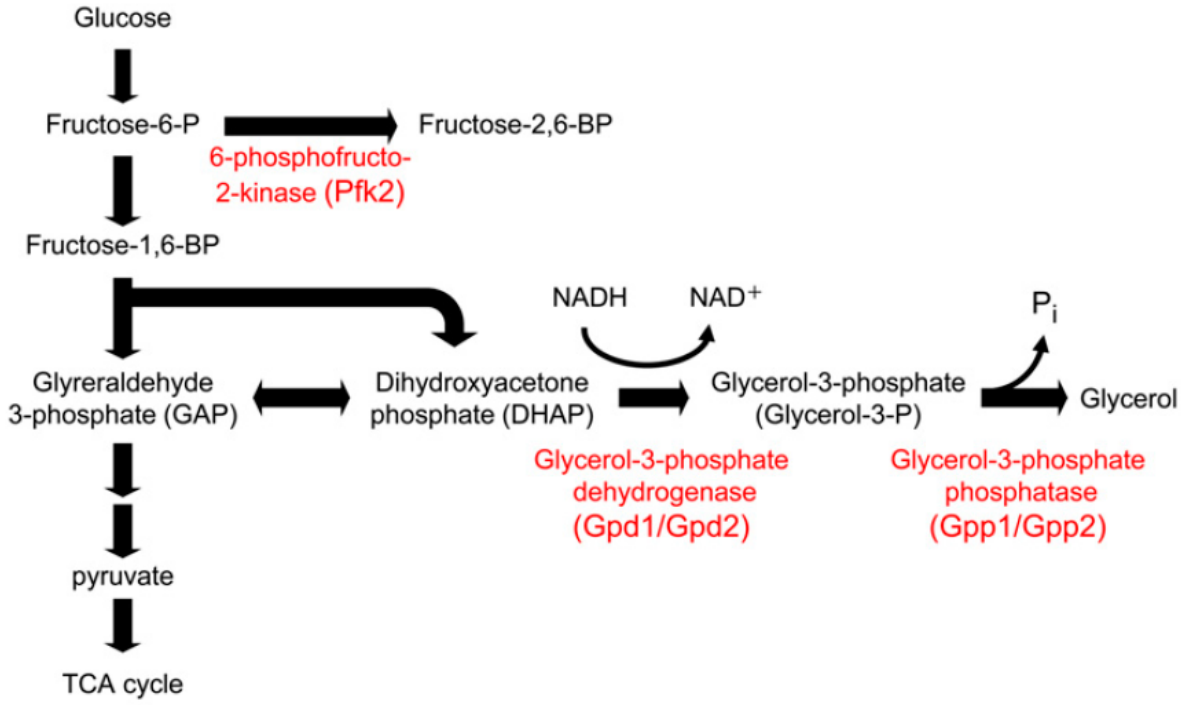



**Figure 70 - Simplified sketch representing the relation of glycerol production with the central carbon metabolism in S. cerevisiae. Enzymes in red are activated by the HOG pathway. Figure adapted from** (62)

When growing in isotonic conditions, yeast cells produce glycerol at a non-negligible basal rate (which serves several purposes such as maintaining the Redox balance of the cell). As it was already mentioned, the genes GPD1, GPP1 and GPP2 are up-regulated under osmotic stress (112). It was reported that Gpp1/Gpp2 is not rate limiting in glycerol synthesis (113) and therefore, expression of GPD1/GPD2 has a more direct impact on cell's adaptation ability. GPD1 plays a major role under aerobic conditions while GPD2 is required and produced in absence of oxygen (113). In addition unlike GPD2, GPD1 is induced by Hog1.

Yet, the transcriptional regulation of glycerol producing enzymes is not responsible for cell adaptation to moderately high osmotic stress (such as those we use here). In fact, although the precise mechanisms are still unclear, it appears that following an osmotic stress, enzymatic activity of glycerol producing enzymes can be increased rapidly (61). In addition to the changes directly related to glycerol production, it appears that osmolarity affects several metabolic routes (61), for instance through the activation of Pfk2 which produces fructose-2,6-biphosphate, affecting rapidly the whole carbon metabolism activity (62).

Here we are primarily concerned with the impact of repeated osmotic stress on proliferation; therefore, we focus on the metabolic aspects which affect growth. Because proliferation and adaptation both use the same resource (glucose and its derivatives), cells under osmotic stress harbor a competition for it between energy production and growth on the one hand and glycerol production on the other hand.

We propose to adopt an economics kind of view of the impact of osmolarity on metabolism. We see hyper-osmotic conditions as imposing a supplementary cost (in terms of glucose mainly, but also in terms of energy) to the basal cost of producing a newborn cell. This cost can be broken down into a fixed cost (*i.e.* a cost a cell has to pay regardless of its proliferation) and a variable cost (which is related to producing cells in a hyperosmotic environment). These costs will be reflected in the growth *yield* of populations in terms of number of cells produced per amount of glucose. Superimposed on this cost aspect is the consideration of flows. In a given environmental condition, the uptake rate of glucose and the rate of production of glycerol and energy are bounded. Considering that in absence of stress cells maximize their proliferation, we can expect that taping into carbon sources for glycerol production will have a similar effect as having less glucose available: it will diminish growth and division rate.

### b. Quantifying adaptation variable cost

Previous work used chemostat cultures with various salt (NaCl) concentrations to determine the cost associated to sustain growth in hyperosmotic conditions (141). They found that the total energetic surplus required for growth in hyperosmotic conditions (0.9 M NaCl) ranged from 28% to 51% of the energetic cost for growth in iso-osmotic medium. In addition, this cost was found to depend on the dilution rate. This cost accounted for both maintenance of cells (what we call a fixed cost) and growth (what we call a variable cost).

In this section, we report experiments aiming at measuring the complete cost of growth in hyperosmotic conditions. Therefore, we are interested in the yield as measured in number of cells produced per quantity of glucose present[116]. In order to quantify this cost, depending on the glucose concentration and the osmotic level, we conducted several batch experiments.

Yeast cells (yPH15) from a fresh YPD plate were grown[117] in liquid cultures containing various glucose and sorbitol concentrations for 48h and the final OD600 was measured[118] as reported in the table below.

[116] Note that this definition of the yield differs from another common definition as the ratio of dry mass of cell produced to the mass of glucose.

[117] Cells were grown aerobically in 3mL at 30°C in a shaking incubator at 250 rpm. Cells were coming from the same YPD plate but not always from the same colony. Nevertheless, the original plate was streaked directly from an isogenic frozen stock of the yPH15 strain.

[118] OD600 was measured in 1mL cuvettes using A SmartSpec Plus spectrometer (BioRad). For each condition, blank was made using the corresponding medium. For OD600 > 0.8 dilutions using the corresponding medium were performed so as to achieve an actual measure below 0.8 on which the dilution factor was applied to find the actual OD600.

| Saturation OD600 | | Sorbitol (M) | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 0,5 | 1 | 1,5 | 2 |
| Glucose (%) | 2% | 9,94 | 8,66 | 6,63 | 3,58 | 3,16 |
| | 1% | 8,39 | 7,69 | 6,76 | 4,37 | 3,66 |
| | 0,10% | 3,60 | 3,15 | 3,00 | 2,57 | 0,71 |
| | 0,01% | 0,67 | 0,67 | 0,63 | 0,65 | 0,22 |

**Figure 72 - Final OD600 for various glucose and sorbitol concentrations. Cultures started from a fresh YPD plate.**

OD600 is a common proxy for cell density. Therefore, a proxy for the total yield of the batch culture can be obtained by dividing the final OD600 by the corresponding glucose concentration[119]

| Yield (OD600 / glucose %) | | Sorbitol (M) | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 0,5 | 1 | 1,5 | 2 |
| Glucose (%) | 2% | 4,97 | 4,33 | 3,32 | 1,79 | 1,58 |
| | 1% | 8,39 | 7,69 | 6,76 | 4,37 | 3,66 |
| | 0,10% | 36,04 | 31,46 | 29,99 | 25,67 | 7,15 |
| | 0,01% | 66,93 | 66,93 | 62,76 | 64,71 | 21,78 |

**Figure 71 - Yield for batch cultures with various glucose and sorbitol concentrations. Cultures started from a fresh YPD plate.**

We note that whatever the glucose concentration is, yield decreases with increasing sorbitol levels. Moreover, as we can see in Figure 73, the average drop in yield due to an osmolarity of 1.8M sorbitol (more or less comparable to 0.9M NaCl) relative to the same glucose concentration without sorbitol is approximately 60% which is comparable to the 28% to 51% overcost measured in (141).

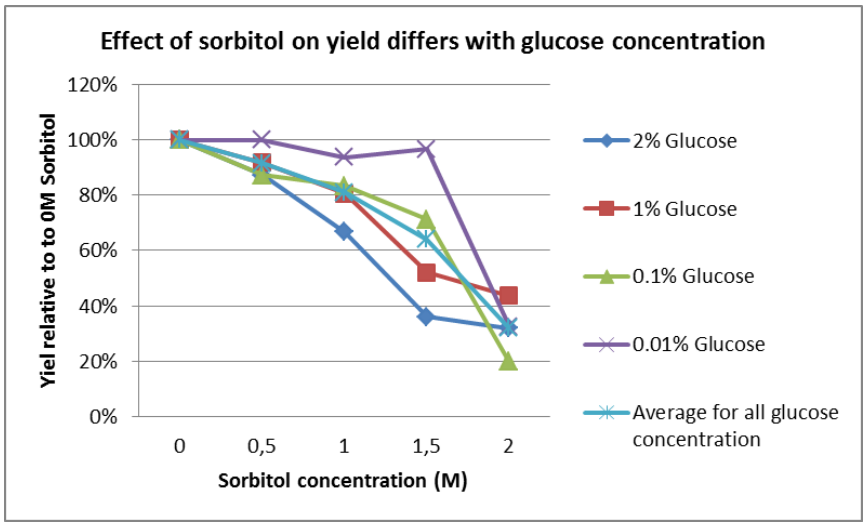



**Figure 73 - Yield (in OD600/ glucose concentration) relative to the yield without sorbitol for cultures started from plate.**

In this first experiment, we started cultures from plated colony of yeast that were not adapted to any of these specific growing conditions. When using the saturated cultures as inoculum for

[119].Yield as measured in batch is not the same as in chemostats with the same glucose and sorbitol concentration. This is because as the cells consume glucose, the exterior concentration in glucose drops.

another experiment (therefore starting effectively from *pre adapted* cells) the results were similar yet, to the exception of very low glucose concentration, more consistent across glucose concentration as shown in Figure 74.

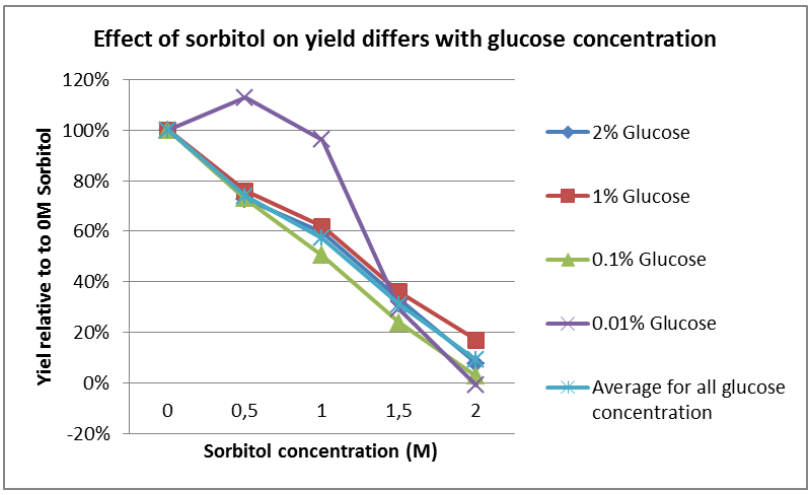



**Figure 74 - Yield (in OD600/ glucose concentration) relative to the yield without sorbitol for cultures started from pre adapted cells. Note that in this case, the average does not take into account the 0.01% Glucose condition which displays a marked difference from the other conditions.**

For this second experiment, the relative decrease in yield due to osmolarity looks more linear. This is coherent with a simple model were growing in high osmolarity requires to produce glycerol as to match the external osmolarity, thereby having a linear increase in the cost with the external sorbitol concentration. Surprisingly, the overall yield reduction is more pronounced in these cultures of *pre adapted* cells that in those starting from plated colony.

As a conclusion, we can therefore summarize these results by stating that each extra molarity of sorbitol in the medium increases the cost in glucose of making a new cell by 45%. Unlike the previous experiment using non adapted inoculum, we find this value to be larger[120] than that reported in (141). Concerning the deviation of the results for very low concentrations of glucose (0.01%), this may result from either:

- The fact that the experimental conditions were inappropriate for such low proliferation. Since absorbance is very low, technical error in determining OD600 is more pronounced. The time span of the experiment may be too short or evaporation of water from medium non negligible any longer before the low concentration of cells.
- Because at these concentration, growth rely mostly on respiration, which is much more efficient than fermentation. Therefore, the cost of growing in sorbitol is less pronounced.

Another question concerns the impact of growing in hyperosmotic medium on the proliferation speed as measured by the division rate. Determination of growth rate in batch cultures proved difficult when measuring OD600 by hand at regular time points because all conditions display different growth kinetics.

[120] By assuming the osmotic potential of NaCl is twice that of Sorbitol (which is not accurate since some Na+ enters the cell) 0.9M NaCl would correspond to 1.8M Sorbitol. At this concentration we find the yield to be reduced by 80% and not 28 to 51%.

Previous data from the lab, acquired with an automated plate reader, give the following impact of osmolarity on OD600 doubling rate in batch (Figure 75).

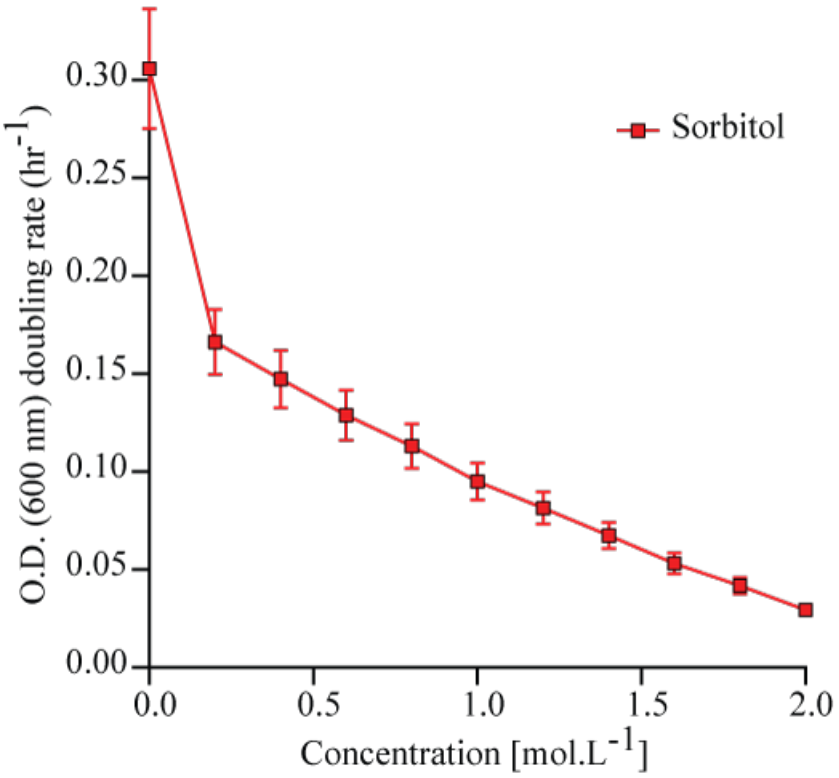



**Figure 75 - Impact of increasing external sorbitol concentration on growth kinetics (OD600 doubling rate) in SC medium with 2% glucose. Data from P. Hersen.**

It seems that growing in a hyperosmotic environment imposes a fixed diminution in growth rate (which corresponds to a 45% decrease with respect to a condition without sorbitol) on top of which we witness an additional diminution which seems linear with respect to the external osmolarity. Fitting the linear part (*i.e.* not taking into account the fixed drop) of the curve in Figure 75 yields that the diminution in growth rate per added molal of sorbitol represents 25% of the growth rate without sorbitol and 46% of the growth rate in 0.2M sorbitol.

Compared to the impact of external sorbitol on yield, the impact on growth rate has a fixed and a variable component. The fixed drop in growth rate upon addition of sorbitol might reflect a qualitative change in growth regime whereas the linear decrease with increasing osmolarity reveals a quantitative change.

We estimated the decrease in yield to be of 45% per extra molal of sorbitol (as quantified in number of cell per quantity of glucose) which matches the proportional decrease of growth rate. This seems to indicate that there is a bottle neck upstream of glycerol production which imposes a slowdown in cellular growth related to the fact that a part of imported glucose is used for glycerol production.

### c. Quantifying acclimation costs of osmotic fluctuation

As we saw previously in this chapter, several factors affect proliferation in repeated stressful environments. In Figure 76 we give a sketch view of the timing which different phenomena can occur, along with notations.

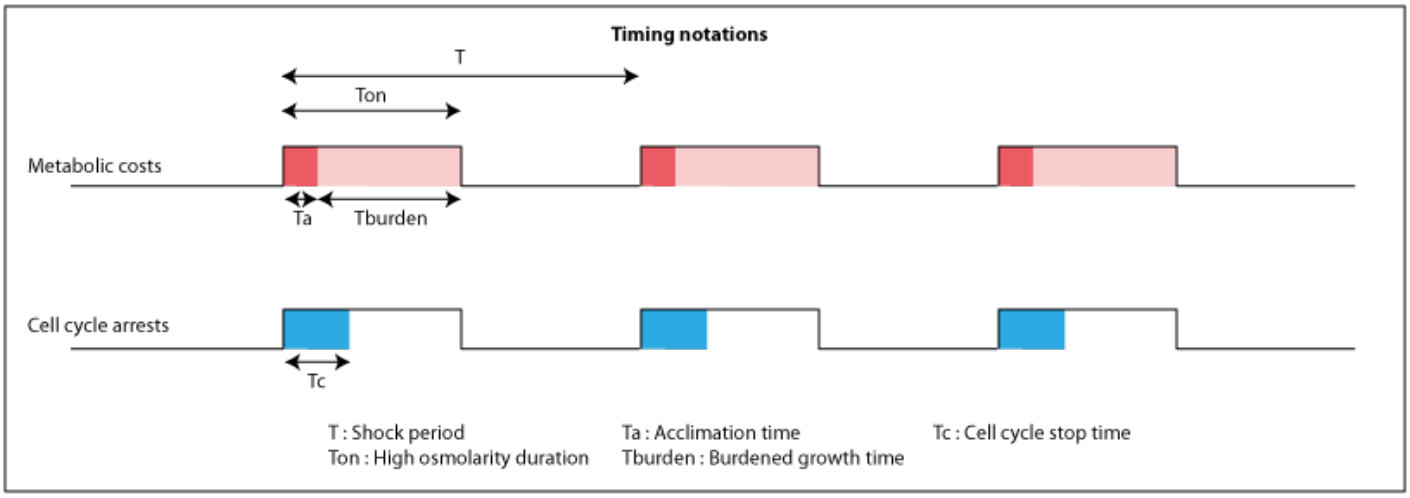



Figure 76 - Notations for the impact of repeated osmotic stress on proliferation.

Upon an osmotic upshift, cells adapt by producing glycerol. This is represented by Ta in Figure 76 and imposes the metabolic cost of glycerol production along with the cost of signaling and modifying cellular physiology. The actual cost of acclimation in terms of energy and glucose is unknown. Once acclimated, cells continue to produce glycerol in order to support growth required for proliferation which will come at an additional cost for as long as the environment is hyperosmotic. This means that during Tburden the cost of producing a cell is ~45% larger than in normal conditions. Once back in isotonic medium (at the end of Ton), nearly all the accumulated glycerol leaks out of cells and is washed away. This resets the glycerol pool to its normal levels. As we saw, cell-cycle can be delayed in several points by 40 to 50 min in average, this delay is noted Tc and in practice is variable among cells and depends on the cell-cycle phase cells are in at the onset of stress. After this delay, the cell-cycle resumes.

We performed several experiments were we subjected cells to symmetric fluctuations of stress (Ton=T/2) at 1 M sorbitol. We varied systematically the frequency of stress (*i.e.* the duration of T) and the available concentration of glucose in the media. We measured the average division rate in the population (computed as the geometric mean of instantaneous division rates like what was presented in IV.2.d and report the results in Figure 77. For a given concentration of glucose, changing the frequency of stress does not change the proportion of time which is spent in hyperosmotic medium. Therefore, the marked changes we observed when the input frequency varies are related to transition effects: cell cycle time delay and acclimation cost. As we see in Figure 77, transition effects can become very important as the fluctuation frequency of the environment increases.

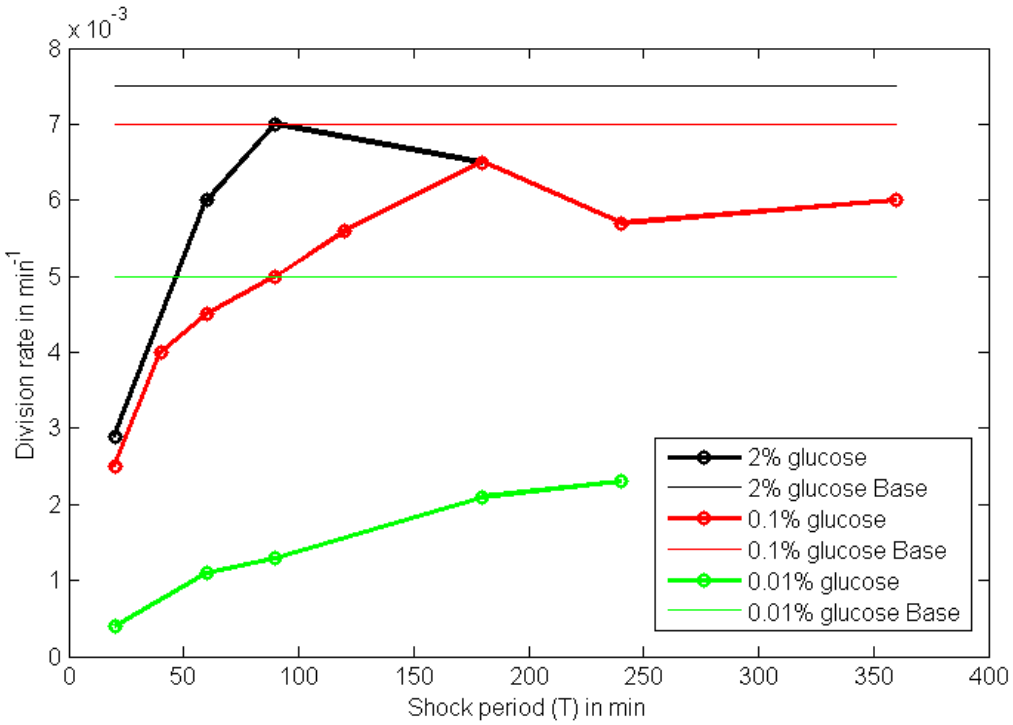



**Figure 77 - Average population division rate for different concentrations of glucose and different stress frequencies. Base lines for each glucose concentration correspond to the division rate in absence of stress.**

As stress frequency increases, a cell will undergo more osmolarity transitions per cell-cycle. This will imply a higher acclimation cost per cycle as well as more cell cycle delays per division time. In terms of division rate, both acclimation costs and molecular delay mechanisms contribute in making each division longer. For as long as a cell is not acclimated, Hog1 is active and continuously triggers cell cycle arrest mechanisms. The effective delay we expect is equal to Tc for sufficiently small Ta and becomes Tc+Ta when acclimation becomes comparable or larger than Tc.

By normalizing the division rates reported in Figure 77 and plotting them in function of stress frequency rather than stress period we observe an apparently linear (affine) relationship of the average division rate with stress frequency as visible in Figure 78. We try to propose an interpretation of such linear relation in the following manner:

We note $T_e$ the effective delay induced by one osmotic transition, $T_S$ the average division time under stress and $T_0$ the average division time without stress. If a cell had a division time of $T_0$, it would experience in average $\frac{T_0}{T}$ periods of stress during its cell cycle. In consequence it would experience a complete delay of $\frac{T_0}{T}.T_e$ and therefore we would have: $T_S = T_0 + \frac{T_0}{T}.T_e$

Noting $f = \frac{1}{T}$ the stress frequency we can rewrite this relation as $\frac{T_S}{T_0} = 1 + T_e.f$

Therefore we can interpret the slope of Figure 78 as $T_e$. Performing linear fits yields the following values:

- For 2% glucose we have $\frac{T_S}{T_0} = 0.86 + 33.f$
- For 0.1% glucose, we have $\frac{T_S}{T_0} = 0.99 + 35.f$
- For 0.01% glucose we have $\frac{T_S}{T_0} = 1.0 + 227.f$

It therefore appears that for 2% and 0.1% glucose we have a very similar effective delay of ~35 min whereas the lower glucose condition gives a much more important delay of nearly 4 h. Because the mechanisms behind cell cycle arrests are unlikely to depend importantly on the glucose availability, a plausible interpretation would be that when glucose is abundant enough, delays in the cell cycle are the main cause of replication slowdown whereas at lower glucose concentrations, acclimation costs becomes more and more important.

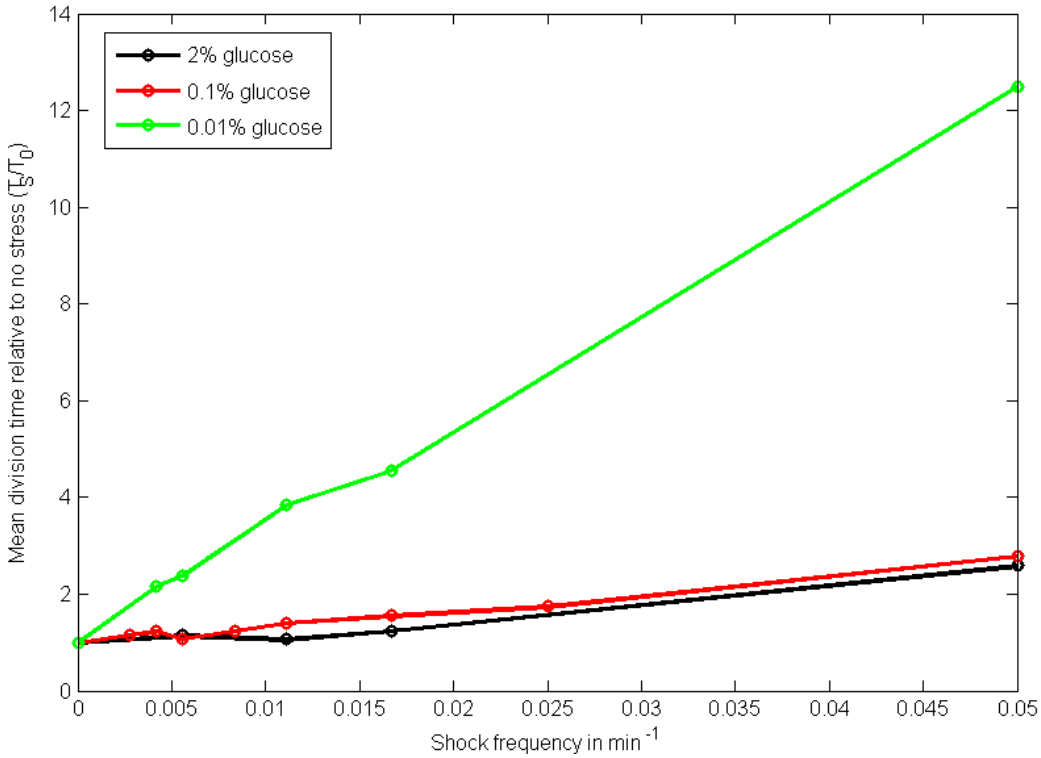



**Figure 78 – Ratio of the average population division rate in absence of stress to the average population division rate for different concentrations of glucose and different stress frequencies. A frequency of 0 means no stress.**

A surprising aspect of the proposed interpretation of the linear relationships visible in Figure 78 which yields the effective delay $T_e$ is that it makes sense as a delay for the number of fluctuations a cell would have received if it had a normal division rate (*i.e.* as in absence of stress) whereas in reality a cell will experience more fluctuations because its average division time is longer. Yet deriving relations using the effective number of fluctuations ($\frac{T_S}{T}$) yields equations which do not match the shapes we observe in Figure 78.

The effective delay that we find is also surprisingly smaller than the average delay in cell cycle upon osmotic stress. This corroborates the surprising fact that we indeed record non negligible division rates for stress periods smaller than the average delay in cycle upon osmotic stress at 2% glucose (~50 min). This indicates than under severe repeated stress, a population of cell is capable of

adaptation which somehow either shortens the cell-cycle delays happening during a single stress event or that they somehow bypass these cell cycle arrests. Partial synchronization of division with stress fluctuations (as shown in IV.2.d) could be part of the answer, with cells progressively adapting their cell cycle so stress occurs in parts of their cycle which is more stress tolerant (or at least deprived of cell arrest mechanism). Another putative explanation would be that cells having slowed down their cell-cycles are less susceptible to cell-cycle arrest.

## 4. Conclusion: The impact of osmotic stress on colony growth dynamics

In this chapter we presented an original approach to characterize the impact of repeated osmotic stress on proliferation. Again the use of dynamic inputs and single-cell longitudinal data proved instrumental to our quantifications. After having reviewed the mechanisms which were known to impact proliferation in relation with osmotic stress we presented several experiments and analysis which aimed at providing temporal quantifications of the processes thereof. Concerning the impact of osmotic stress on the cell-cycle, we observed anomalies occurring at the M phase which had not been reported previously and characterized their occurrence. We used *in-silico* synchronization of single-cell data *a posteriori* to measure cell-cycle arrest without employing potentially harmful and therefore artifact-prone biological synchronization techniques. We characterized the synchronization of cell cycles within a population which occurs under periodic osmotic stress using a spectral analysis of population instantaneous division rates. We considered the major metabolic changes which occur under osmotic stress and which impacts proliferation both in terms of yield and division rate. Focusing on the economic aspect of proliferation, we considered the variable cost of growth in hyperosmotic environments (related to growth and division rates) performing growth rate and yield experiments in batch. We then studied the transitory cost of adapting to changing osmolarity and found a dependence of the average division time upon fluctuation frequency which is congruent with a constant cost per fluctuation. Importantly, our data indicates that this transient cost can be affected by glucose availability.

These findings require further experiments and analysis to obtain a more complete understanding of the quantitative impact of repeated osmotic stress on proliferation. In the months to come, we will complete the picture we have sketched here using quantification of single-cell growth in terms of volume which is a missing piece to understand proliferation as a whole. This addition will allow us to have a minimal working model of proliferation at the single-cell level. This will therefore allow us to assess cell-to-cell variability regarding proliferation in a more coherent manner. Because the effects of metabolic changes and cell cycle arrest can sometimes superimpose one on each other, using mutants or reporters of cell cycle or metabolic activity could provide an additional mean to dissect variability and model it properly. For instance, the amount of Gpd1 is expected to play a significant role in the metabolic part and the localization of Whi5 can serve as a readout of the passage through *Start*.

Proposing an analytical framework summarizing the impact of repeated stress at the single cell level is a challenging task and we are currently considering several possible paradigms. To put it simply, we can start from data and build up a mathematical framework from it or start from available knowledge and use data to adapt and validate its representation as a mathematical model. In the first approach, we would primarily use empirically extracted relationships between our control parameters and observables. This physicist approach has the obvious advantage of providing a compact representation of the information actually present in our data. Nevertheless, we are ultimately interested in relating our observations to the biological knowledge which is available on proliferation and osmotic stress response. In this respect, the difficulty comes in the *a posteriori* biological interpretation of empirical relationships.

Employing the alternative strategy consists in building a mathematical model of the available knowledge which is then fitted to available data. Yet, in such an approach, and given the known complexity of osmotic stress response, growth and cell cycle regulation, we expect severe non-identifiability issues to diminish our confidence in interpreting parameters and predictions. Yet having a model at hand allows data-based model selection techniques along with informed experimental design aiming at improving parameter identifiability overall.

This dichotomy between data-based or knowledge-based mathematical modelling is classic in systems biology. In the problem at hands here, it is the interplay between several already complex processes which is our focus. Recent years have seen several attempts at building *whole cell* models which have the principal advantage of explicitly accounting for such interplay between biological processes. At first, whole cell models seem the extreme version of the knowledge based analytical approach (142). Yet compact whole cell models have also been recently developed which indeed focus on the interplay between processes and abstract most details of any singular biological mechanism (5). In our case, we aim at completing and quantifying known mechanisms on the interplay between repeated stress and growth. Therefore, available whole cell models are not completely adapted. An interesting strategy could be to use a simple enough *abstract* whole cell model as a basis to regenerate empirically inferred relationships between control parameters and observables.

Whole cell models make utter sense when applied to single-cell data which leaves the question of representing cellular individuality with whole cell models. Although we can envision to use mixed effect approaches on concise enough whole cell models, this would only including static individuality (*i.e.* pre-existing differences between cells). As we saw, the increase in cell cycle time under repeated stress is globally regular in its dependence with the frequency of stress. Yet our data on single transient stress show that what happens in repeated stress is not a simple repetition of a single event (otherwise proliferation should be stopped at high enough frequencies, regardless of the glucose concentration). This is possibly indicative of an adaptation mechanism which changes the cell response over time. This in turns might require taking into account time-changing physiology and cellular context on top of the already mentioned individuality.

Nevertheless, our present experimental system may provide insufficient data for a systematic investigation of single-cell individuality dynamics. At some point we need information on how a given cell (in terms of individuality) with a particular history will respond to different environment stimulations. The finer our characterization of cellular individuality, the lesser cells representing each possible case will be observed during a given experiment. Although we observe hundreds of cells over hours, the complete space representing dynamic cellular identity will not be efficiently sampled during fixed experiments. In the following conclusive chapter, we will discuss current possibilities and perspectives concerning real-time, automated, experimental design. The broad array of experimental techniques identified behind such appellation may provide much more adapted tools for the study of dynamic cellular identity.

# V. Perspectives and final discussion

Although isogenic single cells host identical processes, intrinsic randomness in some reactions implies that different cells harbor different realizations of a given process. In addition, cells themselves constitute distinct cellular contexts which will in turn lead to extrinsic, cell-to-cell variability in processes' parameters (which increases the variability of processes outcomes). Even when cell-to-cell variability is not of interest, the fine understanding of biological processes generally requires taking into account variability in time and across a population. Simply neglecting variability by reasoning on *virtual average cell* for instance can introduce systematic bias in quantitative estimation and diminish experimental reproducibility.

A systematic comprehension of dynamical processes which are embedded in a single-cell context could benefit several research areas. A quantitative representation of cellular variability is of high interest for embryology, research on cancer or personalized medicine. In addition, physical limitations and finite resource allocations constrain the cellular context and impose indirect interactions among cellular elements which would otherwise be independents. Such aspects of cellular economics are of high interest for synthetic biology which faces important issues related to orthogonality or modularity of genetic constructs as well as for bioengineering concerning bio-production scaling.

Conducting systems biology at the single-cell level raises specific theoretical and experimental challenges. Compared to aliquots or petri dishes, the physical scale which is imposed by single cell experimentation requires very precise and sensitive technology. Other differences include the fact that whereas for populations, experiments in aliquots allow the use of destructive measurements repeatedly over time, this cannot be done to resolve single cell dynamics. Hopefully, recent progress in instrumentation makes single cells accessible to more and more measurement techniques.

The variable nature of cells induces a profound difference with regular population studies. This is particularly challenging when studying dynamical processes. From a theoretical perspective, we already mentioned many open questions related to cellular differences. In particular, the articulation of intrinsic and extrinsic variability, the notion of identity which can be associated to stable differences between cells and the many interplay and feedbacks between environment, cellular processes and cellular context all require particular analytical frameworks.

## 1. Experimental pipelines for systems biology at the single-cell level

To investigate dynamic processes at the single-cell level, non-destructive experimental methods are necessary. In this respect, fluorescence microscopy is a tool of choice but limits importantly the number of components which can be measured at the same time. Because biological phenomena are usually subject to a large number of interactions, we need to accommodate with partial observations of the complete state of the system. Under traditional modelling assumptions

(*i.e.* a population made of identical cells), this situation already leads to difficult identifiability issues. As cellular variability is taken into account, the scarcity of observations compared to the required amount of information becomes even more pronounced.

***The issue of building a representative sample***

In its simplest form, variability is accounted by a number of discreet cell *types*. As we describe variability in more details, for example considering multi-factorial or continuous differences, the set of possible cell types becomes larger.

While it is possible to use the variability naturally present in a population of cells in order to acquire data on different cell types during a single experiment, increasing the number of criteria defining cellular individuality will decrease the probability to find enough cells of each *type* during a single experiment. In chapter III, variability between cells was represented with few parameters. Performing a single experiment on many cells at the same time and obtaining simultaneously hundreds of single-cell trajectories was therefore sufficient to capture variability.

Yet, the individuality of cells was assumed to be fixed over time. When it comes to the study of the dynamics of the cellular context itself, such as in chapter IV, this experimental procedure is insufficient. In fact, if the past stimulations or the particular history of a cell needs to be taken into account into the cell *identity*, the effective space of cells to sample becomes larger. Simply taking the population of cells present at the beginning of an experiment may not represent a proper sampling of the space of possible cellular context anymore.

Therefore, as we use more refined definitions of cellular individuality, selecting a sample of cells becomes less trivial and careful attention must be paid to it. Depending on the question at hand, using several cohorts of cells having the same age can be relevant whereas in other cases each cohort should be composed of cells of different age in the same proportion as the age pyramid of the complete population.

Given the experimental constrains imposed by dynamical, single-cell measurements we propose three prospective directions to build more informative datasets:

***Improving measurements***

For the dynamical study of gene expression, being able to measure proteins and mRNA at the same time (using a system like MS2 or Spinach) would provide very valuable information compared to proteins only. Such an approach has been recently demonstrated in *E. coli* (143) and my lab is part of a collaborative project aiming at implementing such a system in yeast.

Another type of strategy would be to perform a traditional dynamic experiment which would be followed by a genome-wide (destructive) measurement in-situ (or at least such that we are able to keep track of dynamic and genome-wide measurements concerning the same cell. Advanced microfluidic chips developed in my group would normally allow recovering cells chamber by chamber at the end of an experiment. It may be possible to adapt its design so as to recover only one cell per chamber.

***Improving perturbations***

Throughout this project, we used time-varying stimulations as a tool of choice to obtain informative data on single-cell dynamics. This is made possible thanks to custom microfluidic chips and valves. Yet, all cells in a chamber were always subjected to the same perturbation.

Optogenetics are molecular systems which respond to light. When illuminated with specific wavelengths, some proteins can change their conformation and perform functions (144). For instance, it is possible to trigger the binding of two distinct protein domains in a fast and reversible manner. This allows controlling the cellular localization of a protein (using one domain as an anchor and the other as a target)(145). Such system has been adapted to create light-inducible promoters (56). Although optogenetic systems already constitute interesting induction tools when used globally under the microscope (*i.e.* sending the same illumination all over the sample), it is the possibility to exert single-cell inductions which is the most salient feature of these molecular tools.

In fact, by projecting light patterns through the microscope objective directly on the sample, it is possible to induce single-cells independently (145, 146). Commercial systems allowing single-cell optogenetic stimulation are both expensive and usually run on proprietary software which cannot be programmed. This is an issue when it comes to integrating this tool within an experimental pipeline. During this thesis, I developed an Open Source illumination system which plugs into the microscope and allows single-cell illumination with a higher projection resolution than commercial solutions, and for less than a tenth of the commercial price. This project is described in more details in Annex 8.

***Improving experiments***

In the presented project, we used random or periodic osmotic stimulations to obtain informative data. Complex systems are often composed of interconnected sub-systems which have distinct response frequencies. In the case of the response to osmotic stress, we know that the physical response (*i.e.* loss of water, loss of volume) is relatively fast and therefore will faithfully follow repeated osmotic stress even at high frequency. Signal's transduction by the HOG pathway or nuclear localization of Hog1 both have slower kinetics, acting as integrators at high frequencies. Finally, transcription has even slower typical time (55). Therefore, by applying various frequencies, it is often possible to disentangle various interconnected sub-processes. Yet, such separation of time scales is not always possible. Here we present three more general options aiming at increasing the information content of experiments based on dynamic stimulation.

**Option 1: conditional experiment**

In a conditional experiment, microscopy data is continuously analyzed during the experiment. At each time step, we test a predefined condition based on the data and trigger a predetermined experimental sequence when the condition is valid. This approach is particularly interesting when we study time-changing properties. For instance, in order to assess more precisely the timing of cell-cycle arrests upon repeated hyperosmotic stress, we could want to stress repeatedly a target cell when it is entering the G1 phase. At each time frame, image analysis is performed automatically and the computer determines if the target cell is entering G1 and controls osmolarity if necessary. Because such experiment would only provide data for one cell, it might be convenient to use highly

parallel microfluidic systems with on-chip valve systems and allowing independent control of each culture chambers.

**Option 2: Optimal experimental design**

Given some initial information on a biological system, along with a defined objective (*e.g.* estimating the parameters of a model of the system, being able to perform the best predictions of a given situation etc.), it is possible to optimize the choice of experiments to be conducted. Optimal Experimental Design (OED) is particularly useful when the system under study is already partially characterized as it can only leverage on available knowledge. There are several traditional approaches to OED which differ in the category of models they can handle, the way they predict and define the *utility* of an experiment (*e.g.* using the Fisher information Matrix, using entropy etc.) and the optimization technique used to sample the space of experiments (pure Monte Carlo, biased searches, using heuristic algorithms etc.). In addition to some prior information, one should provide a model of the impact of an experiment. For example, a gene deletion is modeled by fixing the according gene expression rate to 0.

OED is particularly relevant to reverse engineer complex networks of interacting elements for which humans cannot take into account all the expectable consequences of a given perturbation or measurement. Prior to my thesis, I worked on the development of such an algorithm for the estimation of parameters of dynamic models of GRN in the idea of the DREAM 6 and DREAM 7 challenges. The proposed method is original in the fact that unlike many algorithms which base their optimization of experiments on their current best estimate, we designed our algorithm in order to take uncertainty on the system into account explicitly. Also, the proposed method represents OED as a game and adapted one of the most performant algorithm for the game of Go (based on active learning and Monte Carlo Tree Search) (147). Although it was designed for taking uncertainty into account, this algorithm could also be used with populations or single-cell data using as prior information parameter distributions given by mixed effect model estimations. The application of a different OED method on real experimental data proved the ability of such methods to exhibit complex and highly informative patterns of temporal stimulation which would never have been designed by men (148).

***Option 3: Real time control***

A last option consists is performing real-time control of a part or of the whole process. Controlling a cellular process is not necessarily informative *per se*, yet it can be used to *drive* cells towards a state of interest. The possibility to use the HOG pathway for real time control of gene expression has been demonstrated by my research group (86) and a similar control approach was used in (149) with optogenetic inducible systems in batch.

In the context of single-cell systems biology, real-time control can play an important role in ensuring that cells are in the desired state before another experiment would begin. In a way, real time control is a fusion of conditional experiments (where experimental decision is computed in real time) and OED since in some cases performing control requires some optimization on the space of possible experiments.

Real-time control, like conditional experiments, requires real-time, autonomous and reliable analysis and decision making algorithms along with microscopy synchronization and stimulation pilotage. This is made much easier when each of the bricks composing such experimental pipeline has been developed in a modular way and is well documented. Although some commercial software is of very good quality, commercial solutions are most of the time impossible to customize or adapt and cannot be integrated in an experimental pipeline. Usually, Open Source solution exist for the same application, for instance we can quote Micro-Manager, a free software driving our microscopes. The main interest of open source software is not that these are free, but rather that they benefit from modifications, bug fixes, tutorials and customizable plugins by a community of users. At last, they usually include options making integration in custom experimental platforms easier.

Obviously, performing simultaneously the control of single-cell in real time using optogenetics allows not only to drive one or a few cells in a target state, but to control a whole population of cells. Such experimental system would allow hypothesis-driven experimentation of processes subjected to population effects.

## 2. Cellular variability and context

***Cellular economics***

As it was exposed throughout this thesis, many cellular processes can be assumed to depend upon a certain cellular context. Elements of context may be chemical, physical of even geometric features of a cell which are relevant to biological processes. Cellular context can also be affected by the history of a cell. From a given process perspective, the cellular context also includes effects coming from physical and economical limits. Cellular resources present in limited supply needs to be shared between all processes.

Performing direct quantification of available resources would be very informative but is quite challenging. An important aspect in cellular resource allocation concerns the difference between global fluxes, turnover rates and intermediary stocks. In fact, as it was mentioned, most pools of key metabolites are renewed extremely rapidly (BNID 109701). Quantifying the level of ATP in a given cell (or population of cell) would not be very indicative of the actual power which could be generated. If a reporter system was to draw too much of the resource itself, the measuring probe would significantly affect the measured value. Conversely, if this hypothetical probe did only withdraw slight amounts of resource, it would only allow titrating the resource carrier and not the available power.

We studied the impact of repeated osmotic stress on cell cycle and growth. Using periodic osmotic stress, we found that proliferation is affected by stress frequency in a glucose availability manner. Repeated stress accumulates transient metabolic costs and makes it possible to measure its impact. Assuming that single cells maximize their growth rate, the drop in growth due to repeated stress is therefore directly related to the burden we apply. Overall, this allows measuring what a fluctuation cost is.

Given the magnitude with which osmotic stress impacts cellular activity, we wonder if part of the stress response could be related to economic considerations, in particular carbon and NADH sudden depletion related to glycerol production.

Some studies have been able to demonstrate precise resource limitations for specific processes in specific contexts (7, 150). Such hypothesis driven investigation allow identifying conservation laws operating in precise conditions. In order to capture the more general constrains imposed by the cellular context upon biological processes and conversely, the participation of processes to defining the cellular context, parsimonious whole cell models are instrumental. As many different cellular elements are in limited supply, we expect different shared resources to become limiting depending on the moment and the process at play. From specific conservation and economical mechanisms, whole cell models may help identifying more general conservation laws (5).

A particular consequence of cellular economics is that sharing resources somehow creates a flow of reciprocal information between processes which would otherwise be independent. To what extent this reciprocal information is exploited by processes in place of direct coordination mechanism? Can such *market-based* information be selected by evolution? What are the systematic characteristics of such indirect regulations compared to direct mechanisms? Can cells evolve new limiting elements mainly as a mean to impose a new regulation? What would be the consequence of artificially relaxing several of these constrains with exogenous supplies?

***Variability and cellular identity***

Variability in the outcome of biological processes can be broken down into an intrinsic and an extrinsic component. While the intrinsic component is fairly well characterized and is related to random events in chemical reaction, the extrinsic part is defined by its difference with intrinsic one and actually aggregates many different aspects of cells being different.

From a given biological process's perspective, different cells constitute different context possibly leading to different dynamics besides eventual intrinsic fluctuations. Accordingly, the cellular context is accounted as extrinsic variability. Here, we studied long lasting differences between cells as we followed cohorts of cells for several divisions. Doing so, not only we average out intrinsic variability, but we also average any extrinsic variability which would have fluctuated faster, typically at the scale of the cell cycle.

Such long lasting extrinsic variability in gene dynamics was surprisingly correlated to both a feature of the micro-environment as well as elements of the cellular physiology and was inherited from mother to daughter. Given these properties, we wonder to what extent such variability reflects some form of cellular identity which may arises from genealogy, physiology and micro-environment.

Using a mixed effect modelling approach, we can properly relate single cell individuality as defined by single cell parameter values with the overall population. The overall population is represented by a log-normal, multidimensional distribution of parameters. Besides rather technical investigations of the effect of using other distribution shapes, we wonder if any pertinent biological interpretation can be proposed on such distribution. Given the importance of covariance terms to obtain plausible simulated single-cell dynamics, it appears that population distributions capture some constrains on plausible cellular parameters. It would be interesting to perform similar experiment in

different environments and observe how the population distribution's shape might be affected (for example changing the amount of glucose to render adaptation more expensive and slow down growth or supplying a limited amount of amino acids so as to increase the price of protein production).

Since stable extrinsic variability is partially inherited, we wonder to what extent fluctuating osmotic stress may provide an interesting system to study *soft inheritance*. In fact, in particularly intense stress, we often observe a transient adaptation which lasts between one to two cell-cycles. A more thorough single-cell analysis will be performed to understand the mechanism behind such adaptation (do all cells adapt? Some cells are already resistant and proliferate?). Given the frequency of environmental perturbations and the availability of glucose we wonder what strategy is more successful between active adaptation (fast glycerol production) and a more "wait and see" strategy where minimal cellular adaptation allows an efficient energy management (minimal response). Testing such strategies could be done using a cell lacking Gpd1 and hence cannot adapt. Maybe, under specific conditions there can also be some form of phenotypical equivalence where two strategies are equally successful and the overall population distribution is subject to *phenotypic drift*.

***Embracing variability***

The introduction of variability as a genuine biological feature (and not a mere *noise* against which biological entity must act) has profound consequences on how we think of living organisms. Considering cells as machines leads to representing them as extremely complex and dynamic systems harboring so many interactions that any complete mapping or mathematical description is beyond our reach. Acknowledging that variability is a biological constituent cannot be conceived in a vision were living organisms are machines. Machines may be very complex, they may share features with biological entities like the presence of feedbacks; machines nevertheless always work against any variability in both their components and their processes. Machines rely on normalization and standardization where biological entities foster and harvest diversity and variability.

Accepting the fact that cells actually do not work like machines does not, however, leaves us helpless in our endeavor of understanding precisely and quantitatively biology. From a theoretical point of view a system-level approach can be broadened enough to include all biological features. Systems biology will need to find the proper abstraction in order to address the specific type of complexity present in biological entities and reach tractable mathematical descriptions. Mathematical tools currently used in systems biology are not really adapted to handle variability of biological systems. Some probabilistic frameworks can be used to represent it in a satisfactory manner, as well as to perform simulations, but we lack a dedicated analytical framework including variability in a more natural and general way.

Facing the limitations of the classic mechanistic reductionist approach will lead to forging new conceptual frameworks which are more adapted to the reality of living organisms. The well-known assertion of Theodosius Dobzhansky that "Nothing in Biology makes sense except in the light of evolution" is still hardly audible in quantitative molecular biology or even in systems biology. A more biology-centered framework for understanding cells should be constructed around the concepts of variability, natural selection and evolution rather than upon noisy machines. Although several

authors have praised such an approach (151) going beyond words and global concepts to the actual construction of such a new way of thinking and analyzing biological entities is yet to be done.

# List of abbreviations

| | |
|---|---|
| GNR | Gene Regulatory Network |
| ATP | Adenosine triphosphate |
| ADP | Adenosine diphosphate |
| NAD | Nicotinamide adenine dinucleotide (oxidized form) |
| NADH | Nicotinamide adenine dinucleotide (reduced form) |
| YFP | Yellow Fluorescent Protein |
| CFP | Cyan Fluorescent Protein |
| GFP | Green Fluorescent Protein |
| CCD | Charge-Coupled Device |
| ESR | Environmental Stress Response |
| FISH | Fluorescence In Situ Hybridization |
| CWI | Cell Wall Integrity |
| TOR | Target Of Rapamycin |
| PKA | Protein Kinase A |
| HOG | High Osmolarity Glycerol |
| PDMS | Polydimethylsiloxane |
| FP | Fluorescent protein |
| 5' UTR | 5' UnTranslated Region |
| 3' UTR | 3' UnTranslated Region |
| ODE | Ordinary Differential Equation |
| CME | Chemical Master Equation |
| FP | Fluorescent Protein |
| AU | Arbitrary Units |
| S&T | Segmentation and Tracking (image analysis) |
| SC | Synthetic Complete (culture medium) |
| TF | Transcription Factor |
| GRTM | Generalized Random Telegraph Model |
| ME | Mixed Effects |
| SAEM | Stochastic Approximation of Expectation Maximization |
| MAP | Maximum A Posteriori |
| PCA | Principal Component Analysis |
| PC | Principal Component |
| fft | Fast Fourier Transform |
| TCA cycle | Tricarboxylic acid cycle, *a.k.a.* Citric acid cycle or Krebs cycle |
| DHAP | Dihydroxyacetone phosphate |

# References


1. Legras JL, Merdinoglu D, Cornuet JM, Karst F (2007) Bread, beer and wine: Saccharomyces cerevisiae diversity reflects human history. *Mol Ecol* 16(10):2091–2102.

2. Botstein D, Fink GR (2011) Yeast: An experimental organism for 21st century biology. *Genetics* 189(3):695–704.

3. Milo R, Phillips R (2014) *Cell Biology by the Numbers* (Garland Science) Available at: http://book.bionumbers.org/.

4. Crick F (1970) Central Dogma of Molecular Biology. *Nature* 227(5258):561–563.

5. Weiße AY, Oyarzún D a., Danos V, Swain PS (2015) Mechanistic links between cellular trade-offs, gene expression, and growth. *Proc Natl Acad Sci* 112(9):E1038–E1047.

6. Oyarzun DA, Stan G-B V (2012) Synthetic gene circuits for metabolic control: design trade-offs and constraints. *J R Soc Interface* 10(78):20120671–20120671.

7. Gyorgy A, et al. (2015) Isocost Lines Describe the Cellular Economy of Genetic Circuits. *Biophys J* 109(3):639–46.

8. Mishra D, Rivera PM, Lin A, Del Vecchio D, Weiss R (2014) A load driver device for engineering modularity in biological networks. *Nat Biotechnol* 32(12):1268–1275.

9. Dahl RH, et al. (2013) Engineering dynamic pathway regulation using stress-response promoters. *Nat Biotechnol* 31(11):1039–1046.

10. Molenaar D, van Berlo R, de Ridder D, Teusink B (2009) Shifts in growth strategies reflect tradeoffs in cellular economics. *Mol Syst Biol* 5(323):1–10.

11. Shahrezaei V, Marguerat S (2015) Connecting growth with gene expression: of noise and numbers. *Curr Opin Microbiol* 25:127–35.

12. Gerosa L, Kochanowski K, Heinemann M, Sauer U (2013) Dissecting specific and global transcriptional regulation of bacterial gene expression. *Mol Syst Biol* 9(658):658.

13. Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic Gene Expression in a Single Cell. *Science (80- )* 297. doi:10.1126/science.1070919.

14. Raser JM, O'Shea EK (2004) Control of stochasticity in eukaryotic gene expression. *Science* 304(5678):1811–4.

15. Golding I, Paulsson J, Zawilski SM, Cox EC (2005) Real-time kinetics of gene activity in individual bacteria. *Cell* 123(6):1025–36.

16. Suter DM, et al. (2011) Mammalian genes are transcribed with widely different bursting kinetics. *Science* 332(6028):472–4.

17. Sanchez A, Golding I (2013) Genetic determinants and cellular constraints in noisy gene expression. *Science (80- )* 342(6163):1188–93.

18. Raj A, van Oudenaarden A (2008) Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell* 135(2):216–226.

19. Elkon R, Zlotorynski E, Zeller KI, Agami R (2010) Major role for mRNA stability in shaping the kinetics of gene induction. *BMC Genomics* 11:259.

20. Wang Y, et al. (2002) Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci U S A* 2002.

21. Alon U (2007) An Introduction to Systems Biology: Design Principles of Biological Circuits. *Chapman HallCRC Math Comput Biol Ser* 10(10):301.

22. Christiano R, Nagaraj N, Fröhlich F, Walther TC (2014) Global Proteome Turnover Analyses of the Yeasts S. cerevisiae and S. pombe. *Cell Rep* 9(5):1959–1965.

23. Belle A, Tanay A, Bitincka L, Shamir R, Shea EKO (2006) Quantification of protein half-lives in the budding yeast proteome. *Proc Natl Acad Sci U S A*.

24. Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB (2004) Noise Minimization in Eukaryotic Gene Expression. *PLoS Biol* 2(6):e137.

25. Tang F, Lao K, Surani MA (2011) Development and applications of single cell transcriptome analysis. *Nat Methods* 8 (4 Suppl:S6–11.

26. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA (2015) The Technology and Biology of Single-Cell RNA Sequencing. *Mol Cell* 58(4):610–620.

27. Buettner F, et al. (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* 33(2):155–60.

28. Trcek T, et al. (2012) Single-mRNA counting using fluorescent in situ hybridization in budding yeast. *Nat Protoc* 7(2):408–19.

29. Schwabe A, Bruggeman FJ (2014) Single yeast cells vary in transcription activity not in delay time after a metabolic shift. *Nat Commun* 5:4798.

30. Walther TC, Olsen J V., Mann M (2010) Yeast expression proteomics by high-resolution mass spectrometry. *Methods Enzymol* 470(C):259–280.

31. Steen H, Mann M (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* 5(9):699–711.

32. de Godoy LMF, et al. (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* 455(7217):1251–1254.

33. Shrestha B, Vertes A (2009) In situ metabolic profiling of single cells by laser ablation electrospray ionization mass spectrometry. *Anal Chem* 81(20):8265–8271.

34. Ibáñez AJ, et al. (2013) Mass spectrometry-based metabolomics of single yeast cells. *Proc Natl Acad Sci U S A* 110(22):8790–4.

35. Giesen C, et al. (2014) Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat Methods* 11(4):417–22.

36. Angelo M, et al. (2014) Multiplexed ion beam imaging of human breast tumors. *Nat Med* 20(4):436–42.

37. Zuleta I a, Aranda-Díaz A, Li H, El-Samad H (2014) Dynamic characterization of growth and gene expression using high-throughput automated flow cytometry. *Nat Methods* 11(4):443–8.

38. Chong YT, et al. (2015) Yeast Proteome Dynamics from Single Cell Imaging and Automated Analysis. *Cell* 161(6):1413–24.

39. Breker M, Gymrek M, Schuldiner M (2013) A novel single-cell screening platform reveals proteome plasticity during yeast stress responses. *J Cell Biol* 200(6):839–850.

40. Mazo-Vargas A, Park H, Aydin M, Buchler NE (2014) Measuring fast gene dynamics in single cells with time-lapse luminescence microscopy. *Mol Biol Cell* 25(22):3699–708.

41. Bertrand E, et al. (1998) Localization of ASH1 mRNA particles in living yeast. *Mol Cell* 2(4):437–445.

42. Beach DL, Salmon ED, Bloom K (1999) Localization and anchoring of mRNA in budding yeast. *Curr Biol* 9(11):569–578.

43. Paige JS, Wu KY, Jaffrey SR (2011) RNA mimics of green fluorescent protein. *Science (80- )* 333(6042):642–646.

44. Guet D, et al. (2015) Combining Spinach-tagged RNA and gene localization to image gene expression in live yeast. *Nat Commun* 6:8882.

45. Garcia JF, Parker R (2015) MS2 coat proteins bound to yeast mRNAs block 5’ to 3' degradation and trap mRNA decay products: implications for the localization of mRNAs by MS2-MCP system. *RNA* 21(8):1393–5.

46. Shin I, et al. (2014) Live-cell imaging of Pol II promoter activity to monitor gene expression with RNA IMAGEtag reporters. *Nucleic Acids Res* 42(11):1–9.

47. Pothoulakis G, Ellis T (2015) Using Spinach aptamer to correlate mRNA and protein levels in Escherichia coli. *Methods Enzymol* 550:173–85.

48. Von Bertalanffy L (1968) *General System Theory* (Goerge Braziller Inc).

49. Kitano H (2002) Systems biology: a brief overview. *Science* 295(5560):1662–4.

50. Muzzey D, Gómez-Uribe C a, Mettetal JT, van Oudenaarden A (2009) A systems-level analysis of perfect adaptation in yeast osmoregulation. *Cell* 138(1):160–71.

51. Zakeri B, Carr PA (2014) The limits of synthetic biology. *Trends Biotechnol*:1–2.

52. Ellis T, Wang X, Collins JJ (2009) Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat Biotechnol* 27(5):465–71.

53. Bennett MR, Hasty J (2010) Microfluidic devices for measuring gene network dynamics in single cells. *Nat Rev Genet* 10(9):628–638.

54. Mettetal JT, Muzzey D, Gómez-Uribe C, van Oudenaarden A (2008) The frequency dependence of osmo-adaptation in Saccharomyces cerevisiae. *Science* 319(5862):482–4.

55. Hersen P, McClean MN, Mahadevan L, Ramanathan S (2008) Signal processing by the HOG MAP kinase pathway. *Proc Natl Acad Sci U S A* 105(20):7165–70.

56. Olson EJ, Hartsough L a, Landry BP, Shroff R, Tabor JJ (2014) Characterizing bacterial gene circuit dynamics with optically programmed gene expression signals. *Nat Methods* (August

2013):1–11.

57. Villaverde AF, Banga JR (2014) Reverse engineering and identification in systems biology: strategies, perspectives and challenges. *J R Soc Interface* 11(91):20130505.

58. Kramer EM, Myers DR (2012) Five popular misconceptions about osmosis. *Am J Phys* 80(8):694.

59. Miermont A (2013) Severe osmotic compression of the yeast Saccharomyces cerevisiae.

60. Atwell BJ, Kriedermann PE, Turnbull CGN (1999) *Plants in Action: Adaptation in Nature, Performance in Cultivation* (Macmillan Education AU).

61. Hohmann S (2002) Osmotic Stress Signaling and Osmoadaptation in Yeasts. *Microbiol Mol Biol Rev* 66(2). doi:10.1128/MMBR.66.2.300.

62. Saito H, Posas F (2012) Response to Hyperosmotic Stress. *Genetics* 192(2):289–318.

63. Miermont A, Uhlendorf J, McClean M, Hersen P (2011) The Dynamical Systems Properties of the HOG Signaling Cascade. *J Signal Transduct* 2011:930940.

64. Reed RH, Chudek JA, Foster ROY, Gadd GM (1987) Osmotic Significance of Glycerol Accumulation in Exponentially Growing Yeasts. *Appl Environ Microbiol* 53(9):2119–2123.

65. Tao W, Deschenes RJ, Fassler JS (1999) Intracellular glycerol levels modulate the activity of Sln1p, a Saccharomyces cerevisiae two-component regulator. *J Biol Chem* 274(1):360–367.

66. Lam MHY, et al. (2015) A Comprehensive Membrane Interactome Mapping of Sho1p Reveals Fps1p as a Novel Key Player in the Regulation of the HOG Pathway in S. cerevisiae. *J Mol Biol* 427(11):2088–2103.

67. Smits GJ, Kapteyn JC, Van den Ende H, Klis FM (1999) Cell wall dynamics in yeast. *Curr Opin Microbiol* 2(4):348–352.

68. Levin DE (2005) Cell Wall Integrity Signaling in Saccharomyces cerevisiae Cell Wall Integrity Signaling in Saccharomyces cerevisiae. *Microbiol Mol Biol Rev* 69(2):262–291.

69. Levin DE (2011) Regulation of Cell Wall Biogenesis in Saccharomyces cerevisiae: The Cell Wall Integrity Signaling Pathway. *Genetics* 189(4):1145–1175.

70. Batiza AF, Schulz T, Masson PH (1996) Yeast respond to hypotonic shock with a calcium pulse. *J Biol Chem* 271(38):23357–23362.

71. Gasch AP, et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* 11(12):4241–57.

72. Futcher B (2000) Microarrays and cell cycle transcription in yeast. *Curr Opin Cell Biol* 12(6):710–715.

73. Liebermeister W, et al. (2014) Visual account of protein investment in cellular functions. *Proc Natl Acad Sci* 111(23):8488–8493.

74. Ruis H, Schüller C (1995) Stress signaling in yeast. *BioEssays* 17(11):959–965.

75. Gasch AP (2003) The environmental stress response : a common yeast response to diverse environmental stresses. *Yeast Stress Responses*, eds Hohmann S, Mager WH doi:10.1007/3-540-45611-2.

76. Vaga S, et al. (2014) Phosphoproteomic analyses reveal novel cross-modulation mechanisms between two signaling pathways in yeast. *Mol Syst Biol*:1–21.

77. Romero-santacreu L, Moreno J, Perez-Ortin JE, Alepuz P (2009) Specific and global regulation of mRNA stability during osmotic stress in Saccharomyces cerevisiae. *RNA*:1110–1120.

78. Warringer J, Hult M, Regot S, Posas F, Sunnerhagen P (2010) The HOG Pathway Dictates the Short-Term Translational Response after Hyperosmotic Shock. *Mol Biol Cell* 21(17):3080–3092.

79. De Nadal E, et al. (2004) The MAPK Hog1 recruits Rpd3 histone deacetylase to activate osmoresponsive genes. *Nature* 427(6972):370–4.

80. Nadal-Ribelles M, et al. (2012) Hog1 bypasses stress-mediated down-regulation of transcription by RNA polymerase II redistribution and chromatin remodeling. *Genome Biol* 13(11):R106.

81. Slaughter B, Li R (2006) Toward a molecular interpretation of the surface stress theory for yeast morphogenesis. *Curr Opin Cell Biol* 18(1):47–53.

82. Miermont A, et al. (2013) Severe osmotic compression triggers a slowdown of intracellular signaling , which can be explained by molecular crowding. *Proc Natl Acad Sci U S A* 110(33):5725–5730.

83. Hao N, et al. (2007) A systems-biology analysis of feedback inhibition in the Sho1 osmotic-stress-response pathway. *Curr Biol* 17(8):659–67.

84. Klipp E, Nordlander B, Krüger R, Gennemark P, Hohmann S (2005) Integrative model of the response of yeast to osmotic shock. *Nat Biotechnol* 23(8):975–82.

85. Petelenz-Kurdziel E, et al. (2013) Quantitative analysis of glycerol accumulation, glycolysis and growth under hyper osmotic stress. *PLoS Comput Biol* 9(6):e1003084.

86. Uhlendorf J, et al. (2012) Long-term model predictive control of gene expression at the population and single-cell levels. *Proc Natl Acad Sci U S A* 109(35):14271–6.

87. Zechner C, et al. (2012) Moment-based inference predicts bimodality in transient gene expression. *Proc Natl Acad Sci U S A* 109(21):8340–5.

88. Pelet S, et al. (2011) Transient activation of the HOG MAPK pathway regulates bimodal gene expression. *Science* 332(6030):732–735.

89. Neuert G, et al. (2013) Systematic identification of signal-activated stochastic gene regulation. *Science* 339(6119):584–7.

90. Adrover M a., et al. (2011) Time-Dependent Quantitative Multicomponent Control of the G1-S Network by the Stress-Activated Protein Kinase Hog1 upon Osmostress. *Sci Signal* 4(192):ra63–ra63.

91. Zi Z, Liebermeister W, Klipp E (2010) A quantitative study of the Hog1 MAPK Response to fluctuating osmotic stress in saccharomyces cerevisiae. *PLoS One* 5(3):1–13.

92. Gonzalez AM, Uhlendorf J, Cinquemani E, Batt G, Ferrari-trecate G (2013) Identification of biological models from single-cell data : a comparison between mixed-effects and moment-based inference.

93. Whitesides GM (2006) The origins and the future of microfluidics. *Nature* 442(7101):368–73.

94. Probst C, Grünberger A, Wiechert W, Kohlheyer D (2013) Microfluidic growth chambers with optical tweezers for full spatial single-cell control and analysis of evolving microbes. *J Microbiol Methods*. doi:10.1016/j.mimet.2013.09.002.

95. Ferry MS, Razinkov IA, Hasty J (2011) *Microfluidics for Synthetic Biology* (Elsevier Inc.). 1st Ed. doi:10.1016/B978-0-12-385075-1.00014-7.

96. Lee C, Chang C, Wang Y, Fu L (2011) Microfluidic Mixing : A Review. *Int J Molecuar Sci*:3263–3287.

97. Hansen AS, Hao N, Shea EKO (2015) High-throughput microfluidics to control and measure signaling dynamics in single yeast cells. *Nat Protoc* 10(8):1181–1197.

98. Shaner NC, Steinbach PA, Tsien RY (2005) A guide to choosing fluorescent proteins. *Nat Methods* 2(12):905–909.

99. Day RN, Davidson MW (2009) The fluorescent protein palette: tools for cellular imaging. *Chem Soc Rev* 38(10):2887–2921.

100. Chudakov DM, Matz M V, Lukyanov S, Lukyanov KA (2010) Fluorescent Proteins and Their Applications in Imaging Living Cells and Tissues. *Physiol Rev* 90(3):1103–1163.

101. Khmelinskii A, et al. (2012) Tandem fluorescent protein timers for in vivo analysis of protein dynamics. *Nat Biotechnol* 30(7):708–14.

102. Sheff M a, Thorn KS (2004) Optimized cassettes for fluorescent protein tagging in Saccharomyces cerevisiae. *Yeast* 21(8):661–70.

103. Garneau NL, Wilusz J, Wilusz CJ (2007) The highways and byways of mRNA decay. *Nat Rev Mol Cell Biol* 8(2):113–26.

104. Rechsteiner M, Rogers SW (1996) PEST sequences and regulation by proteolysis. *Trends Biochem Sci* 21(7):267–71.

105. Grilly C, Stricker J, Pang WL, Bennett MR, Hasty J (2007) A synthetic gene network for tuning protein degradation in Saccharomyces cerevisiae. *Mol Syst Biol* 3(127):127.

106. Ettinger A, Wittmann T (2014) Fluorescence live cell imaging. *Methods Cell Biol* 123(6):77–94.

107. Rodrigues I, Sanches J (2010) Photoblinking/photobleaching differential equation model for intensity decay of fluorescence microscopy images. *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (IEEE), pp 1265–1268.

108. Versari C, et al. (2016) Robust Long Term Single Cell Tracking from Brightfield Microscopy Images of Budding Yeast. *(In Prep*.

109. Snijder B, Pelkmans L (2011) Origins of regulated cell-to-cell variability. *Nat Rev Mol Cell Biol* 12(2):119–125.

110. Capaldi AP, et al. (2008) Structure and function of a transcriptional network activated by the MAPK Hog1. *Nat Genet* 40(11):1300–6.

111. de Nadal E, Casadomé L, Posas F (2003) Targeting the MEF2-like transcription factor Smp1 by the stress-activated Hog1 mitogen-activated protein kinase. *Mol Cell Biol* 23(1):229–237.

112. Bouwman J, et al. (2011) Metabolic regulation rather than de novo enzyme synthesis dominates the osmo-adaptation of yeast. *Yeast* 28(1):43–53.

113. Påhlman AK, Granath K, Ansell R, Hohmann S, Adler L (2001) The Yeast Glycerol 3-Phosphatases Gpp1p and Gpp2p Are Required for Glycerol Biosynthesis and Differentially Involved in the Cellular Responses to Osmotic, Anaerobic, and Oxidative Stress. *J Biol Chem* 276(5):3555–3563.

114. Huang S (2009) Non-genetic heterogeneity of cells in development : more than just noise. *Development* 136:3853–3862.

115. Tsuchyia M, et al. (2007) Gene expression waves: Cell cycle independent collective dynamics in cultured cells. *FEBS J* 274(11):2878–2886.

116. Zechner C, Unger M, Pelet S, Peter M, Koeppl H (2014) Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nat Methods* 11(2):197–202.

117. Gordon A, et al. (2007) Single-cell quantification of molecules and rates using open-source microscope-based cytometry. *Nat Methods* 4(2):175–181.

118. Delyon B, Lavielle M, Moulines E (1999) Convergence of a stochastic approximation version of the EM algorithm. *Ann Stat* 27(1):94–128.

119. Kuhn E, Lavielle M (2005) Maximum likelihood estimation in nonlinear mixed effects models. *Comput Stat Data Anal* 49(4):1020–1038.

120. Chan PLS, Jacqmin P, Lavielle M, McFadyen L, Weatherley B (2011) The use of the SAEM algorithm in MONOLIX software for estimation of population pharmacokinetic-pharmacodynamic-viral dynamics parameters of maraviroc in asymptomatic HIV subjects. *J Pharmacokinet Pharmacodyn* 38(1):41–61.

121. Lavielle M (2014) *Mixed Effects Models for the Population Approach.* (CRC Press).

122. Moazed D (2001) Enzymatic activities of Sir2 and chromatin silencing. *Curr Opin Cell Biol*:232–238.

123. Yang J, et al. (2015) Systematic analysis of asymmetric partitioning of yeast proteome between mother and daughter cells reveals “aging factors” and mechanism of lifespan asymmetry. *Proc Natl Acad Sci U S A* 112(38):11977–82.

124. Gonzalez AM, et al. (2013) Identification of biological models from single-cell data : a comparison between mixed-effects and moment-based inference. *Control Conference (ECC), 2013 European* (IEEE), pp 3652–3657.

125. Stoeger T, Battich N, Herrmann MD, Yakimovich Y, Pelkmans L (2015) Computer vision for image-based transcriptomics. *Methods* 85:44–53.

126. Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB (2005) Gene regulation at the single-cell level. *Science* 307(5717):1962–1965.

127. Turner JJ, Ewald JC, Skotheim JM (2012) Cell size control in yeast. *Curr Biol* 22(9):R350–9.

128. Alberghina L, et al. (2012) Cell growth and cell cycle in Saccharomyces cerevisiae : Basic regulatory design and protein – protein interaction network. *Biotechnol Adv* 30(1):52–72.

129. Marion RM, et al. (2004) Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. *Proc Natl Acad Sci U S A* 101(40):14315–22.

130. Porro D, Vai M, Vanoni M, Alberghina L, Hatzis C (2009) Analysis and modeling of growing budding yeast populations at the single cell level. *Cytometry A* 75(2):114–20.

131. Duch A, et al. (2013) Coordinated control of replication and transcription by a SAPK protects genomic integrity. *Nature* 493(7430):116–9.

132. Bellí G, Garí E, Aldea M, Herrero E (2001) Osmotic stress causes a G1 cell cycle delay and downregulation of Cln3/Cdc28 activity in Saccharomyces cerevisiae. *Mol Microbiol* 39(4):1022–35.

133. Escoté X, Zapater M, Clotet J, Posas F (2004) Hog1 mediates cell-cycle arrest in G1 phase by the dual targeting of Sic1. *Nat Cell Biol* 6(10):997–1002.

134. Duch A, et al. (2013) Coordinated control of replication and transcription by a SAPK protects genomic integrity. *Nature* 493(7430):116–9.

135. Clotet J, et al. (2006) Phosphorylation of Hsl1 by Hog1 leads to a G2 arrest essential for cell survival at high osmolarity. *EMBO J* 25(11):2338–46.

136. Reiser V, D'Aquino KE, Ee L-S, Amon A (2006) The stress-activated mitogen-activated protein kinase signaling cascade promotes exit from mitosis. *Mol Biol Cell* 17(7):3136–3146.

137. Robertson AM, Hagan IM (2008) Stress-regulated kinase pathways in the recovery of tip growth and microtubule dynamics following osmotic stress in S. pombe. *J Cell Sci* 121(Pt

24):4055–68.

138. Chen G, Bradford WD, Seidel CW, Li R (2012) Hsp90 stress potentiates rapid cellular adaptation through induction of aneuploidy. *Nature* 482(7384):246–50.

139. Dhar R, Sägesser R, Weikert C, Yuan J, Wagner a (2011) Adaptation of Saccharomyces cerevisiae to saline stress through laboratory evolution. *J Evol Biol* 24(5):1135–53.

140. Brewer BJ, Chlebowicz-Sledziewska E, Fangman WL (1984) Cell cycle phases in the unequal mother/daughter cell cycles of Saccharomyces cerevisiae. *Mol Cell Biol* 4(11):2529–2531.

141. Ölz R, Larsson K, Adler L, Gustafsson L (1993) Energy flux and osmoregulation of Saccharomyces cerevisiae grown in chemostats under NaCl stress. *J Bacteriol* 175(8):2205–2213.

142. Karr JR, et al. (2012) A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell* 150(2):389–401.

143. Pothoulakis G, Ceroni F, Reeve B, Ellis T (2013) The Spinach RNA Aptamer as a Characterization Tool for Synthetic Biology. *ACS Synth Biol*. doi:10.1021/sb400089c.

144. Bacchus W, Fussenegger M (2012) The use of light for engineered control and reprogramming of cellular functions. *Curr Opin Biotechnol* 23(5):695–702.

145. Yang X, Jost AP-T, Weiner OD, Tang C (2013) A light-inducible organelle-targeting system for dynamically activating and inactivating signaling in budding yeast. *Mol Biol Cell* 24(15):2419–30.

146. Levskaya A, Weiner OD, Lim WA, Voigt CA (2009) Spatiotemporal control of cell signalling using a light-switchable protein interaction. *Nature* 461(7266):997–1001.

147. Llamosi A, Mezine A, D'Alché-Buc F, Letort V, Sebag M (2014) Experimental Design in Dynamical System Identification: A Bandit-Based Active Learning Approach. *Exp Des* 8725:306–321.

148. Ruess J, Milias-Argeitis A, Lygeros J (2013) Designing experiments to understand the variability in biochemical reaction networks. *J R Soc Interface* 10(88):20130588.

149. Milias-Argeitis A, et al. (2011) In silico feedback for in vivo regulation of a gene expression circuit. *Nat Biotechnol* 29(12):1114–6.

150. Ceroni F, Algar R, Stan G-B, Ellis T (2015) Quantifying cellular capacity identifies gene

expression designs with reduced burden. *Nat Methods* 12(5):415–418.

151. Kupiec J-J, Sonigo P (2000) *"Ni Dieu ni gène." Pour une autre théorie de l'hérédité.* (Seuil, Paris).

152. Sherman F (2002) Getting Started with Yeast. *Methods Enzym* 41:350.

153. Aguilaniu H, Gustafsson L, Rigoulet M, Nyström T (2003) Asymmetric inheritance of oxidatively damaged proteins during cytokinesis. *Science* 299(5613):1751–1753.

154. Rujano M a., et al. (2006) Polarised asymmetric inheritance of accumulated protein damage in higher eukaryotes. *PLoS Biol* 4(12):2325–2335.

155. Grunstein M, Gasser SM (2013) Epigenetics in Saccharomyces cerevisiae. *Cold Spring Harb Perspect Biol* 5(7):a017491–a017491.

156. Giaever G, et al. (2002) Functional profiling of the Saccharomyces cerevisiae genome. *Nature* 418(6896):387–391.

157. Costanzo M, et al. (2010) The genetic landscape of a cell. *Science* 327(5964):425–431.

158. Huh W-K, et al. (2003) Global analysis of protein localization in budding yeast. *Nature* 425(6959):686–91.

159. Zopf CJ, Quinn K, Zeidman J, Maheshri N (2013) Cell-Cycle Dependence of Transcription Dominates Noise in Gene Expression. *PLoS Comput Biol* 9(7):e1003161.

160. Hughes TR, de Boer CG (2013) Mapping Yeast Transcriptional Networks. *Genetics* 195(1):9–36.

161. Lee TI, et al. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science (80- )* 298(5594):799–804.

162. Schwarz G, Mueller L, Beck S, Linscheid MW (2014) DOTA based metal labels for protein quantification: a review. *J Anal At Spectrom* 29(2):221–233.

163. Schwarz G, Beck S, Benda D, Linscheid MW (2013) MeCAT--comparing relative quantification of alpha lactalbumin using both molecular and elemental mass spectrometry. *Analyst* 138(8):2449–55.

164. Domitrovic T, Fernandes CM, Boy-Marcotte E, Kurtenbach E (2006) High hydrostatic pressure activates gene expression through Msn2/4 stress transcription factors which are involved in the acquired tolerance by mild pressure precondition in Saccharomyces cerevisiae. *FEBS Lett*

580(26):6033–6038.

165. Stark C, et al. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 34(Database issue):D535–D539.

166. Raue A, et al. (2009) Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* 25(15):1923–1929.

167. Gillespie DT (1992) A rigorous derivation of the chemical master equation. *Phys A Stat Mech its Appl* 188(1-3):404–425.

168. Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J Comput Phys* 22(4):403–434.

169. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81(25):2340–2361.

170. Lee CH, Kim K-H, Kim P (2009) A moment closure method for stochastic reaction networks. *J Chem Phys* 130(13):134107.

171. Tong AHY, et al. (2004) Global mapping of the yeast genetic interaction network. *Science* 303(5659):808–813.

# Appendix

## 1. List of Strains

| Strain ID | Genome | Comments |
|---|---|---|
| yPH_142 | (pSTL1-STL1)Δ0 | Construction strain to move away from telomere |
| yPH_171 | STL1::CFP-cln2Pest-KanMX | Active Degron for STL1 system |
| yPH_172 | HIS3::CFP-cln2Pest-KanMX | Active Degron control with constitutive promoter |
| yPH_173 | STL1::CFP-KanMX | Control for yPH_171 characterization |
| yPH_174 | HOG1-GFP-HIS3<br>HTB2-mCherry-URA3<br>HIS3::CFP-KanMX | Control for yPH_178 |
| yPH_175 | STL1::ECFP-HIS3 | Alternative selection to yPH_173 |
| yPH_177 | HOG1-GFP-HIS3<br>HTB2-mCherry-URA3<br>STL1::CFP-cln2Pest-KanMX | Strain for STL1 system with improved kinetics |
| yPH_178 | Hog1-GFP-HIS3<br>HTB2-mCHerry-URA3,<br>HIS3::CFP-cln2Pest-KanMX | Monitoring HOG1 load on constitutive promoter |
| yPH_146 | stl1-ECFP-HIS3MX6<br>HTB2::mCherry-kanMX | Made with S. Jaramillio for lineage extraction of STL1 system |
| yPH_147 | GDP1::ECFP-HIS3MX6<br>HTB2::mCherry-kanMX | Made with S. Jaramillio for ineage extraction of HOG1 impact |
| yPH_143 | pCup1-CplXP<br>pGal1-GFP-ssrA | Given by Hasty lab – Cup Inducible degron for Gal inducible GFP |
| yPH_144 | HIS3::pGal1-yEGFP-ssrA<br>TRP1::pADH1i-ClpP<br>URA3::pADH1i-yClpX<br>LEU2::pADH1-mLacI | Given by Hasty Lab – LacI Inducible degron for Gal inducible reporter |
| yPH_156 | LEU::PhyB-mCHerry-CAAX-LEU<br>HIS::Pif6-mCitrine-HIS | Given by Weiner Lab – Optogenetic membrane recruitment |
| yPH_157 | LEU::PhyB-mCHerry-CAAX-LEU<br>TRP::pGal1-mCerrulean-TRP<br>Gal80::Gla80-mCitrine-Pif6-NAT | Given by Weiner Lab – Optogenetic control of gene expression |

Table 1 – List of strains constructed or received during the project

## 2. The use of S. cerevisiae

In the table reproduced from Milo in Figure 79 are reported some characteristic orders of magnitudes for different common types of cells used for research. Despite apparent quantitative differences between these cell types, it is important to understand that a large portion of these reflect direct size scaling. For example, if we compare the DNA content it seems that the mammalian cell has way more DNA than *E. coli* or *S. cerevisiae*. But if we renormalize the number of bases given in Figure 79 with the cellular volume which is relevant for DNA (being the whole volume for *E. coli* and the nuclear volumes for *S. cerevisiae* and *H. Sapiens* as given in BNID 104709 and 104716) we find concentrations of 4.6, 4.3 and 4.6 Kb.$\mu m^{-3}$ respectively. As we see, the DNA concentration is quite similar. The same observation can be done for the proteins content: the actual total concentration of proteins in the cell is actually 1.0 1.0 and 3.2 mM for *E. coli*, *S. cerevisiae* and *H. sapiens* respectively and these have similar average sizes (Figure 79). Most of the time chemical reactions depend upon concentrations rather than absolute number of molecules. As a consequence, a large part of cellular biochemistry will be similar from bacteria to mammalian cells.

| property | *E. coli* | budding yeast | mammalian (HeLa line) |
|---|---|---|---|
| cell volume | 0.3–3 $\mu m^3$ | 30–100 $\mu m^3$ | 1,000–10,000 $\mu m^3$ |
| proteins per $\mu m^3$ cell volume | | $2–4\times10^6$ | |
| mRNA per cell | $10^3$-$10^4$ | $10^4$–$10^5$ | $10^5$–$10^6$ |
| proteins per cell | ~$10^6$ | ~$10^8$ | ~$10^{10}$ |
| mean diameter of protein | | 4–5 nm | |
| genome size | 4.6 Mbp | 12 Mbp | 3.2 Gbp |
| number protein coding genes | 4300 | 6600 | 21,000 |
| regulator binding site length | | 10–20 bp | |
| promoter length | ~100 bp | ~1000 bp | ~$10^4$–$10^5$ bp |
| gene length | ~1000 bp | ~1000 bp | ~$10^4$–$10^6$ bp (with introns) |
| concentration of one protein per cell | ~1 nM | ~10 pM | ~0.1–1 pM |
| diffusion time of protein across cell ($D \approx 10\ \mu m^2/s$) | ~0.01 s | ~0.2 s | ~1–10 s |
| diffusion time of small molecule across cell ($D \approx 100\ \mu m^2/s$) | ~0.001 s | ~0.03 s | ~0.1–1 s |

**Figure 79 - Table of typical values and dimensions for several cell types. Reproduced from** (3)

Although *S. cerevisiae* is a microscopic unicellular organism, it is eukaryote. Despites the billion year or so (no precise figure is available) of separated evolution between fungi and animals, it still shares many features with plants or even humans, both in terms of genetic material, and in terms of cellular processes. Baker's yeast is a central organism for the study of genetics as reflected by the fact that, *S. cerevisiae* was the first sequenced eukaryote (in 1996). Its genome is composed of 16 chromosomes (for haploids) which represents overall a bit more than 12 Mb (Millions of nucleobases). Baker's yeast has 6600 genes, whose sequences represent 72% of the total genome, such compactness contrasting sharply with most eukaryotes (152). Most strains also harbor a plasmid (called *2-µm*) although it is apparently useless to the cell (152).

Gene expression processes and gene regulation are identical or very similar to many animals. Therefore, from a system level point of view, yeast displays an equivalent level of complexity and of interaction between its elements to be representative of many other living organisms at the gene expression and regulation levels. Yet having one of the shortest genomes among eukaryotes, using *S. cerevisiae* for our study allows working on a *relatively* small system[121] while keeping most of the complexity.

*S. cerevisiae* can live both as a haploid and a diploid which is useful for many studies or genome manipulation. It mostly reproduces by cellular division where a cell gives rise to two isogenic cells (which allows easily obtaining large population of isogenic cells). An important and quite peculiar aspect of S. cerevisiae division is that it is asymmetric. This pattern, termed *budding* makes it possible to distinguish a *mother* and a *daughter* cell which will share the same DNA sequence but will differ in size, replicative age and also surprisingly in protein content because of asymmetric protein repartition upon mitosis (101, 153). This asymmetry between the two cells resulting from mitosis is also true for other cells whose division is symmetric (154), but is much easier to study in S. cerevisiae because the two resulting cells are visually distinguishable. Therefore budding yeast, as it is also called, proved to be instrumental in the study of cellular aging and inheritability. Budding yeast can also display more complex life cycles including sexual conjugation and sporulation which are beyond the scope of this work but definitely constitute some perks of this organism as research subject as much as genetic manipulation method.

*S. cerevisiae* is also subject to *epigenetic* processes which imply (depending on the definition of epigenetic) some form of inheritance and regulation of gene expression that is not related to the DNA sequence itself (and therefore is not sequence-specific as classic gene regulation by transcription factors). Epigenetic processes are related to the organization of the DNA and to the molecules in proximity like histones. Recent advances in epigenetics are imposing a new vision of cellular adaptability and evolution dynamics and again, the yeast *S. cerevisiae* has been instrumental in the understanding of many epigenetic phenomena. Some more precise aspects of yeast epigenetics will be detailed later in this document and the reader can refer to (155) for a presentation of some of the main epigenetic processes in *S. cerevisiae*.

If budding yeast has been studied so much it is for a large part because of its ease of use compared to other cellular types. Although the most common laboratory strain (S288c) is not believed any longer to be fairly representative of wild yeast, it has exceptional qualities for lab work (152). It is by far the most documented strain of yeast and for the system level approach we employ, the fact that it is a domesticated hybrid rather than a wild organism will only impose caution in making conclusion about yeast in its natural environment. Yeast cells divide pretty fast (with a division time of roughly 100 min depending on the strain and conditions) which allow faster

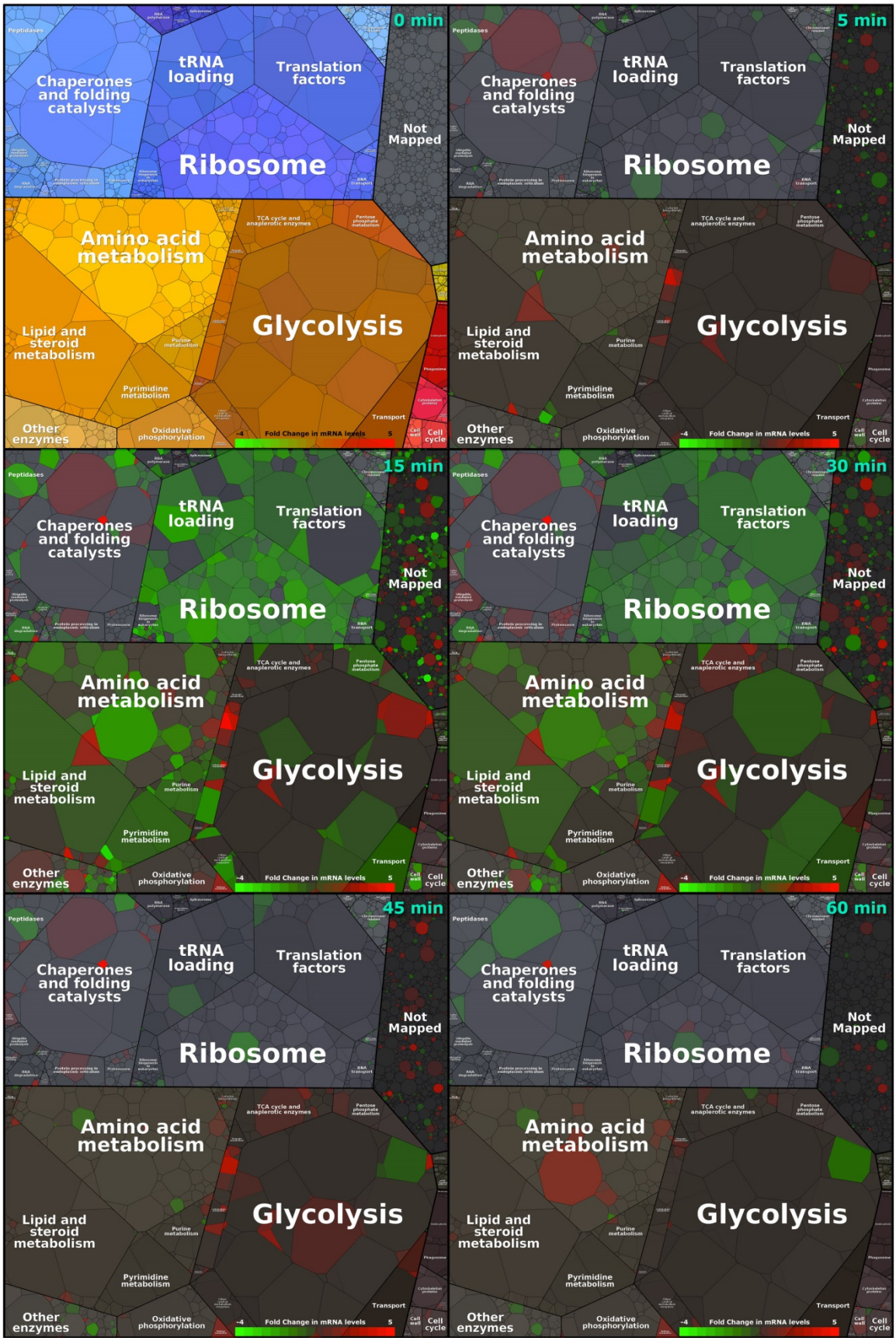
[121] The most up-to-date database of yeast genes interaction (through physical or genetic means) has 334 500 interactions entries out of which 209 538 are genetic. This means than on average every of the 6600 genes of *S. cerevisiae* is connected to 50 other genes (although using the mean degree of such a network makes no sense as the degree distribution of such networks is in the form of a power law also called *small world* or *scale free*). Experimental attempts to map with a single procedure such interaction map (171) predicted around 100 000 genetic interactions with an average of 34 interactions ranging from 1 to 146.

experiments than most eukaryote cells. Most strains do not aggregate and therefore are easier to observe under the microscope. Their round shape makes it easier, relatively to other cells, to automatically segment microscopy images; which as it will be detailed in II.2, is a fundamental prerequisite for high throughput single cell longitudinal studies.

Concerning genetic manipulation, *S. cerevisiae* is an exceptional subject. Like bacterium, it can transcribe plasmids and replicate them (although using a different origin of replication than bacteria). But if plasmids are easy to use and give rise to more efficient transformation, it is the possibilities of genome integration that make this yeast a great organism for gene related studies. Like nearly all cells, this organism can perform homologous recombination (which acts to repair double strand breaks). But for some reason, it is much easier to use this natural process in yeast to perform gene targeting (*i.e*. modifying precise endogenous DNA sequences) than in most other cell types. Using transformation techniques based on homologous recombination, it is therefore possible to integrate, alter or delete nearly any portion of the genome in a reliable and stable manner. The possibility to use auxotrophic markers in addition to the common antibiotic resistance selection markers allows creating multiple mutants more easily. At last, the possibility to perform counter selection (where we select a loss of a marker rather than the integration of one) is also convenient to create highly modified strains.

This ease of use led to courageous initiatives like the *Saccharomyces* gene deletion project (156) where around 20 000 yeast strains were constructed by deleting one open reading frame (ORF) at a time. This generated a library of mutants covering 90% of all ORF (some ORF being so redundant that achieving specific deletion proved too difficult). The existence of such library also explains that yeast's genome is much better annotated than most sequenced organisms. It was followed by a similar initiative using double mutants to study interaction between genes at the genome level (157). In a similar way, but so as to study gene expression, the GFP collection consists of more than 4000 strains where GFP was fused to the C-terminal of protein producing genes. This collection covers roughly 75% of yeast's proteome (158).

## 3. Transcriptome time course in response to hyperosmotic stress




Montage representing the time evolution of mRNA abundance following a hyperosmotic stress (1M sorbitol, data from (71)). To visualize the functional implications this data is overlaid over yeast regular proteome (mass spectrometry data from (32)) visualized as a Voronoi diagram (73) where surface represents protein abundance and proteins interactions are represented by spatial proximity.

## 4. Custom microfluidic chips fabrication method

***Making microfluidic chips***

Microfluidic chips are most of the time made out of PDMS[122] which is bonded to a microscope coverslip. PDMS is a silicone polymer which is crosslinked in order to form a solid resin which has numerous advantages: It is non-toxic, inert, very transparent and porous to gas but not to liquids. PDMS presents itself as a viscous liquid which is mixed with a cross-linking agent (curing agent) in order to for a silicon resin. In addition, PDMS is relatively cheap and its mechanical properties can be adjusted by adapting curing agent ratio or polymerization temperature. Microfluidic chips in PDMS are created from a wafer (or mold) which is produced by photolithography. The overall process of wafer and chip fabrication is summarized In Figure 80, reproduced from Ferry *et al* (95).

Microfluidic chip fabrication protocol contains the following steps:

1. Mix base and curing agent in the proper ratio (here 1:10 curing agent to base ratio) by manual mixing with a pipette.
2. Remove bubbles trapped in the mix placing the viscous mix in a vacuum chamber (~15 min).
3. Pour the mix on top of the master mold. Eventually remove dust or new bubbles with a pipette (Figure 80 G)
4. Cure at 65°C for at least 12h
5. Once PDMS has been cured on the wafer, the chip is cut of with a scalpel and peeled off the wafer.
6. Holes are punched[123] which will allow tubing to be connected to the chip.
7. Chips are cleaned with office tape (Magic tape 3M) and using a plasma cleaner
8. Finally, the chip is bonded to a coverslip[124] (Figure 80 H).

[122] PDMS stands for Polydimethylsiloxane, which is a specific type of silicone. In our experiments, Sylgard 184 (Dow chemical) was used.
[123] Punchers and connectors are usually homemade from syringes needles using a Dremel tool.
[124] In order to bind PDMS to glass coverslips we activate both surfaces in a plasma cleaner (for 1 min 10) and simply put activated surfaces one against the other. This leads to the formation of hydrogen bonds which ensure a water-tight sticking between glass and PDMS.

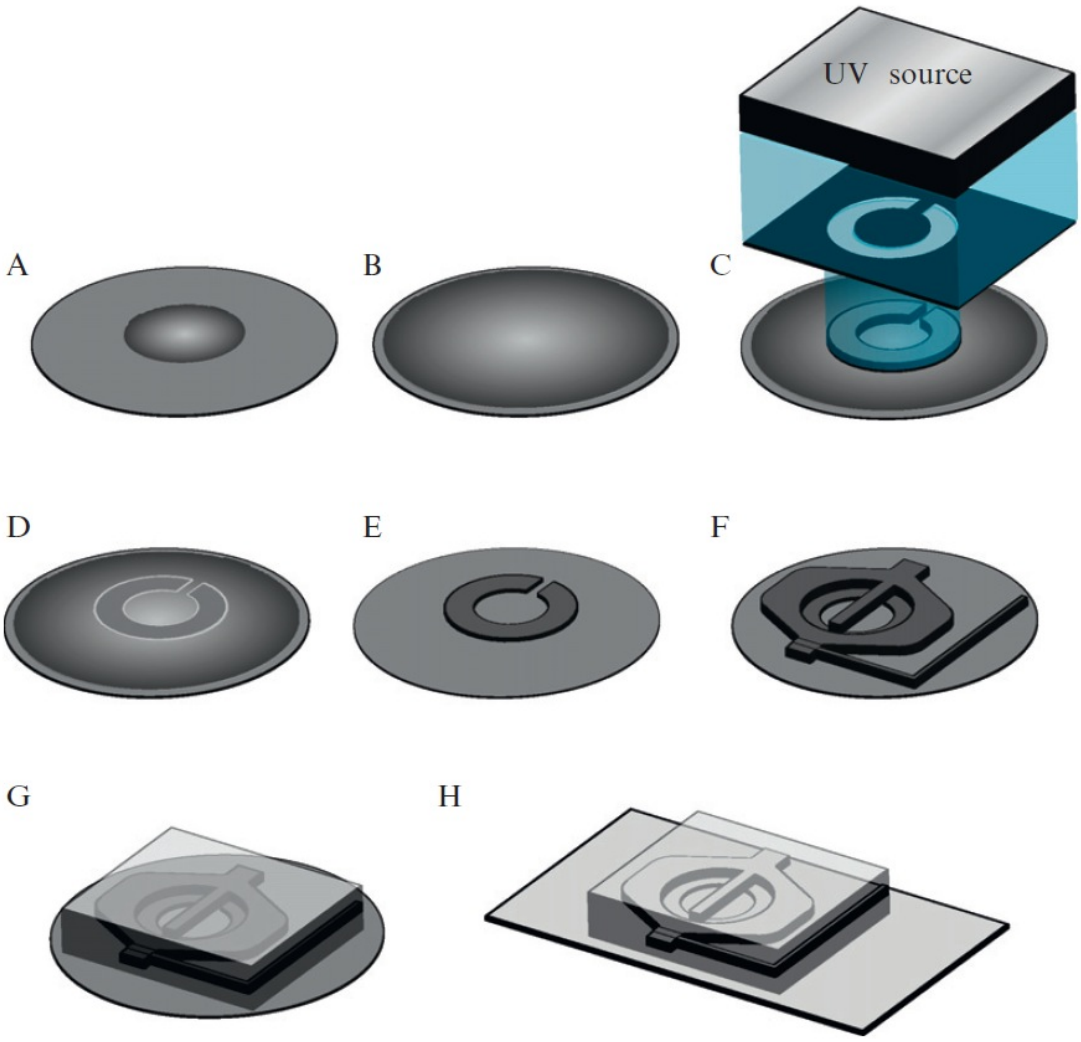



**Figure 80 - Overview of the fabrication process: Photolithography (A–F), soft lithography (G), and PDMS processing (H). (A) Photoresist deposition. (B) Spin-coating: the deposited photoresist is spun at a specific speed to create a uniformly thick layer. (C) UV exposure cross-links the photoresist creating a pattern identical to the photomask. (D) Postexposure baking joins the silicon wafer and the cross-linked photoresist. (E) Developing removes the uncross-linked photoresist, revealing the features. (F) Repeating steps A–E creates additional features. (G) Pouring and curing PDMS over the patterned wafer creates a mold. (H) Bonding the PDMS mold to a glass coverslip finishes a microfludic chip. Figure from** (95).

***Designing microfluidic chips***

While many microfluidic chips are now commercially available, we continue to hand make them along with designing and producing master molds. Masters are usually made out of more solid resins and are fabricated in a clean room to avoid any dust or floating particle to get included in the mold. Molds are made by soft-lithography which is similar to argentic photographical printing. For an extended protocol, see (95).

The essential steps are the following:

1. Thin layers of UV sensitive resin[125] are deposited on a silica wafer by spin-coating (Figure 80 A-B).

2. Using custom designed masks, we expose parts of this resin layer to UV (Figure 80 C).

3. This layer is then *developed* in a solvent which dissolves the resin which was not exposed to UV (Figure 80 E).

4. If several different heights are needed, operations 1 to 3 are repeated using different resin thickness and different masks, starting by the lower pattern(Figure 80 F). This requires sub-micrometer alignment of already developed patterns with the mask for the next pattern. This delicate operation requires a specific microscope coupled to a UV source which is called an aligner[126].

5. When a mask is finished, its surface quality along with its actual dimensions is measured using a profilometer (dektak).

6. Prior to the first molding of PDMS, wafers are *silanized i.e*. a thin layer of cross linking agent such as (3-Mercaptopropyl)trimethoxysilane (Sigma Aldrich) is. This allows passivating Si groups from the silica wafer so they will not for Si-Si bonds with the PDMS which is molded. This is performed by placing the wafer in a vacuum chamber with a beaker of silane so silane vapor can react with the wafer.

The high resolution masks required for soft lithography are printed by external contractor (either in chrome deposited on glass or high resolution inkjet printing on transparent plastic sheets) and designed with CAD software like L-Edit (Tanner), AutoCAD (Autodesk) for very high resolution or Illustrator (adobe) for micron scale resolution.

[125] We usually use the photoresists SU8 (2000.5 to 2050 from microchem) which are epoxy-based, UV sensitive resins available in several densities which yield different film thickness. To make 3 to 4µm high chambers, I make my own SU8 by mixing SU8 2005 and a SU-8 thinner available from microchem.
[126] In Paris Diderot facility we use for wafer fabrication, the aligner is a MJB4 aligner (SUSS MicroTec).

## 5. Glucose diffusion and consumption in microfluidic chambers

# 1 Assumptions and scope

**Scope** In this simple computation we try to estimate diffusion of glucose in a microfluidic chamber where cells grow as a monolayer. We try to model the diffusion limited glucose transport along with growth related glucose consumption. Typical time scale for this equilibrium is small compared to cell division so cellular proliferation itself is not taken into account. Rather, a constant density of cells (defined in a surfacic manner as we consider monolayers) is assumed.

**Defining equations** We use a unidimensional version of the reaction diffusion equation which relates $\frac{dc(x,t)}{dt}$ to glucose consumption and diffusion. Diffusion is represented using Fick's second law where the concentration change due to diffusion in a fixed volume is proportionnal to $\frac{d^2c(x,t)}{dx^2}$

Glucose uptake by cells is consider to follow Hill kinetics and is of the form $v_{max} \cdot \frac{c(x,t)}{K_m + c(x,t)}$.

**PDE resolution scheme** Given glucose consumtion is non linear, explicit resolution of the reaction-diffusion PDE is difficult. We therefore implemented a simple finite differences scheme.

In a H-shaped device, chambers are connected by both ends to flow channels. The growth chamber is parallepipedic and subdivided in $N$ slices (see Figure 1).

**Boundary conditions** Given the speed of flow in flow channels, a constant concentration of glucose can be assumed which imposes Dirichlet boundary conditions. Because of the symmetry at half the chamber, we can consider that it acts as a wall (since diffusion from both sides will be exactly the same, the number of molecules diffusing in each direction will be the same and no flux can be assumed). This leads to Neumann boundary conditions. Since the second Fick law can be derived from Ficks first law (which accounts

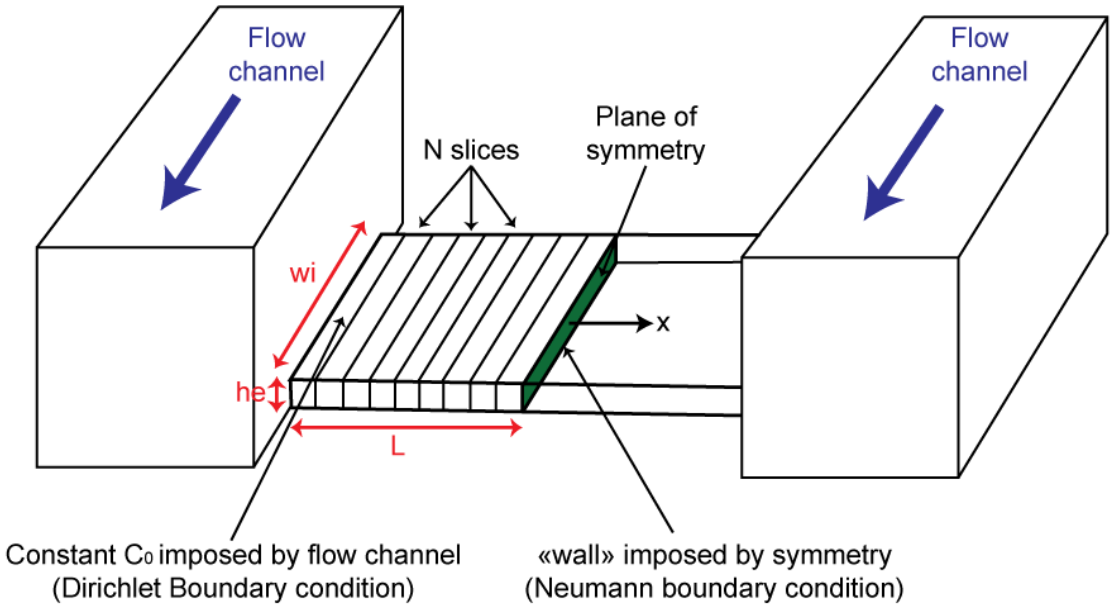



Figure 1: Sketch of the system's geometry and boundary conditions

for the transport of molecules in absolute numbers), we used considerations of absolute numbers to derive equations concerning concentrations for the first and last slice.

# 2 Notations and parameters

## 2.1 Geometric parameters

Half-chamber's dimensions using the notations of Figure 1 are:

$$L = 200\mu m$$

$$he = 3.7\mu m$$

$$wi = 3\mu m$$

The chamber is divided into $N$ equal slices (this corresponds to the mesh step of our finite differences scheme) having an elementary volume equal to:

$$evol = \frac{L.he.wi}{N}$$

with length $l = \frac{L}{N}$

A portion of this volume is occupied by cells. In this simulation, our free parameter is the surfacic cell density (*i.e.* the portion of surface seen from above which is occupied

by cells). This is called the packing ratio and in the case of circles of equal radius, a higher bound is approximately 0.9. Two illustrations are given in Figure 2 A.

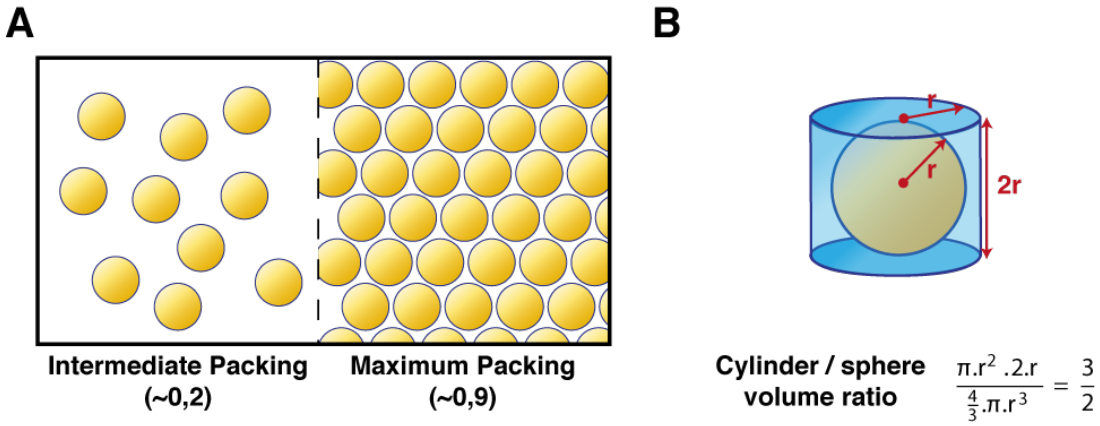



Figure 2: Sketch representing A. Different surfacic cell packing and B. Explaining the prefactor in free volume computation.

We assume cells have a radius of $cr = \frac{he}{2}$ and are evenly distributed in space. As a consequence the total number of cells in this half chamber is:

$$TotNcells = \frac{wi.L.cellpacking}{\pi.cr^2}$$

In terms of volume, the portion of chamber volume which is no occupied by cells is:

$$VolPack = \frac{1}{3}.cellpacking + (1 - cellpacking)$$

where $\frac{1}{3}.cellpacking$ corresponds to available volume for surface occupied by cells (cylinder to sphere ratio, see Figure 2. B) and $(1 - cellpacking)$ is the volumic packing contribution of surface free of cells.

## 2.2 Biological and physical parameters

Cellular glucose uptake is modeled with a Michaelis Menten rate, *i.e.* the molar uptake rate is in the form:

$$u_{rate} = v_{max}.\frac{c(x,t)}{c(x,t) + K_m}$$

with $c(x,t)$ the glucose concentration at distance x from the flow channel and at time t. Using BNID 110954 and 106225 and adapting units we find the following parameters

$$v_{max} = 55.10^{-9} nmol.cell^{-1}.s^{-1}$$

$$K_m = 76.10^{-3} M$$

As a consequence, the total glucose uptake for a given slice is:

$$u_{slice} = v_{max}.\frac{TotNcells}{N}$$

in $nmol.s^{-1}$ which impacts the concentration of glucose in a given slice as:

$$c_{up} = \frac{u_{slice}}{VolPack.evol}.10^6$$

where the $10^6$ factor allows this quantity to be in $mol.L^1.s^{-1}$.
The constant glucose concentration from the flow channel corresponding to 2
Overall, our ODE system has 5 parameters:

$$[c_{up}, K_m, D, C_0, l]$$

# 3 ODEs definition

With all the previous notations, and using a centered approximation of the spatial second derivative along direction $x$, we have the following ODE for a slice $i$ accounting for the time evolution of the glucose concentration in $mol.L^{-1}$:

$$\frac{dc(i,t)}{dt} = -c_{up}.\frac{c(i,t)}{K_m + c(i,t)} + D.\frac{c(i-1,t) - 2.c(i,t) + c(i+1,t)}{l^2}$$

For the frist and last slices, this equation must be adapted accordingly:

$$\frac{dc(1,t)}{dt} = -c_{up}.\frac{c(1,t)}{K_m + c(1,t)} + D.\frac{C_0 - 2.c(1,t) + c(2,t)}{l^2}$$

$$\frac{dc(N,t)}{dt} = -c_{up}.\frac{c(N,t)}{K_m + c(N,t)} + D.\frac{-c(N,t) + c(N-1,t)}{l^2}$$

# 4 Simulation results

Simulation with $N = 20$ slices, starting with a chamber fully packed (0.9) with cells, with no glucose inside and over the course of 3 min yields the following results (Figure 3):

As we can see, equilibrium is reached rapidly at all length. We can visualize the same data both in space and time as in Figure 4:

As we can see, when the chamber is fully packed with cells, the actual glucose concentration in the middle of the chamber can be significantly impacted. This highly depends on the cellular density. This dependance is shown in Figure 5 depicting the steady state distribution of glucose concentration for different densities of cells. As we can see, the

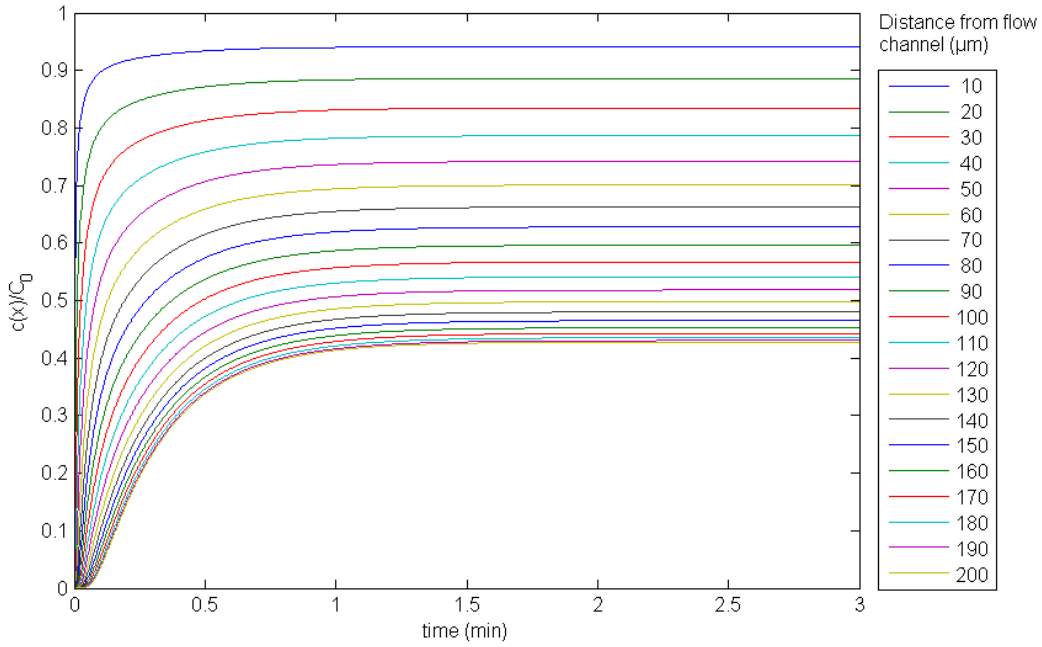



Figure 3: Time evolution of glucose concentration in different slices. Regular H-shaped device

concentration drop can can affect cells at high densities since it drops below $K_m$ meaning effective uptake will be less than half its maximal value.

This situation has been improved by using square chambers as in the parallel H-shaped device type C which has $wi = 400\mu m$ and $L = 200\mu m$ as we can see in figure 6

At last, using chambers of the parallel H-shaped device type A which have a shape of a hourglass with $wi_{flowchambertochannel} = 400\mu m$ $wi_{midcahmber} = 200\mu m$ and $L = 200\mu m$ further improves this situation as we can see in figure 7. (nb: this required to adapt the proposed equations yet, these are very similar).

As a conclusion, we see that shape matters in order to ensure homogeneous conditions in microfluidic devices. Nevertheless, these calculations are too crude to consider using these results for correcting data. Rather, these are used as design guidelines and in order to determine the conditions in which the homogeneous assumption is possible.

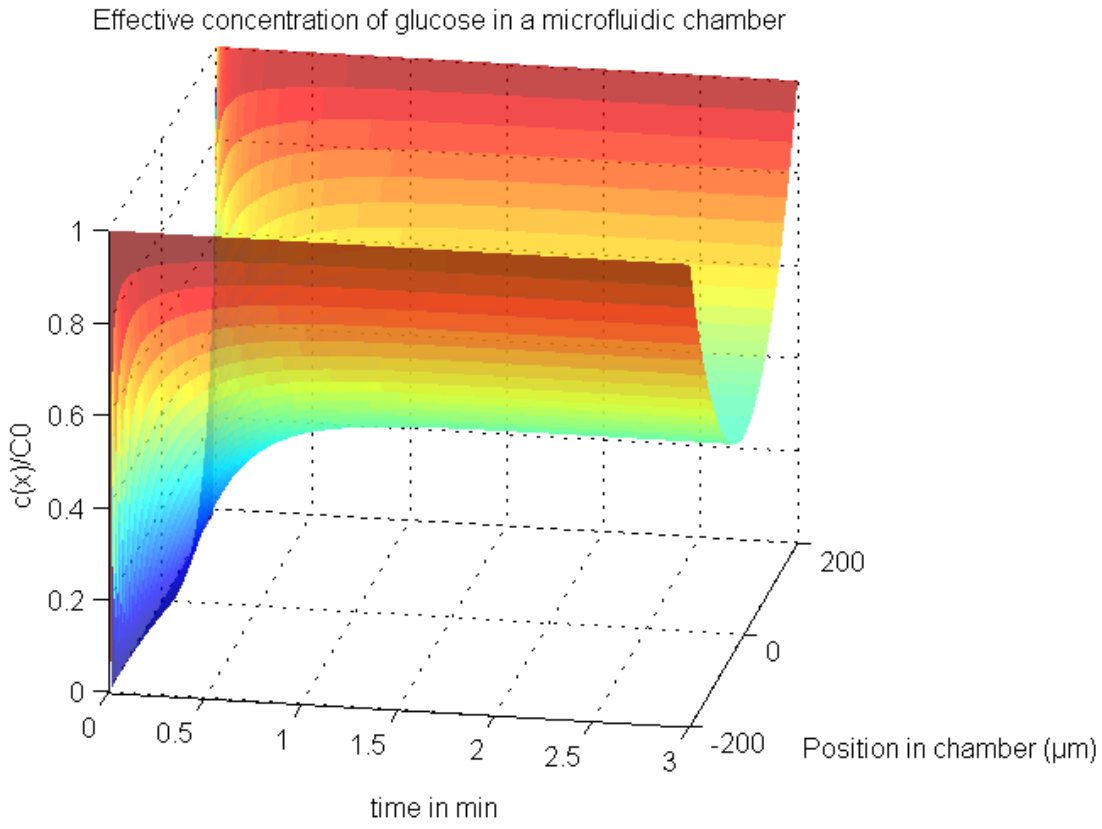



Figure 4: Time and spatial evolution of glucose concentration in the overall microfluidic chamber. For this graph, $x = 0$ corresponds to the center of the chamber. Regular H-shaped device.

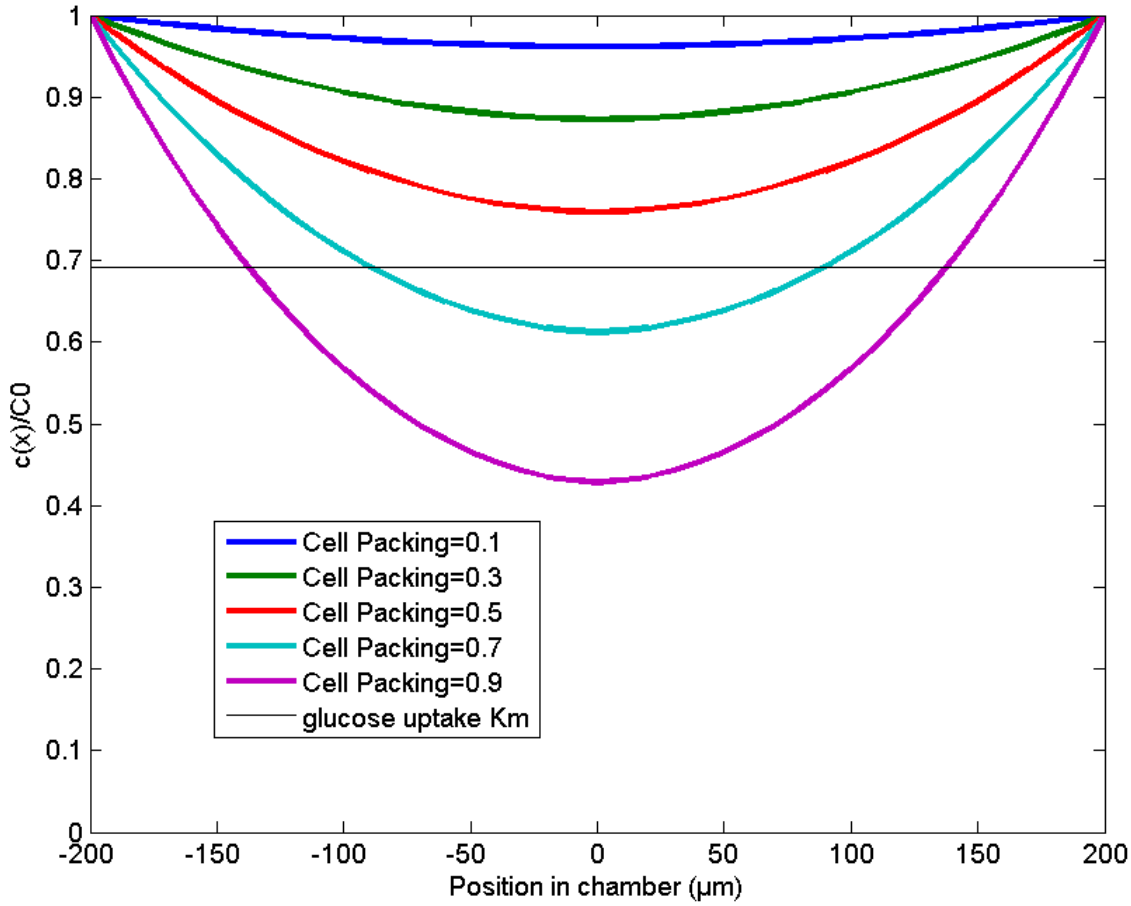



Figure 5: Steady state glucose distribution for various cellular densities. Regular H-shaped device

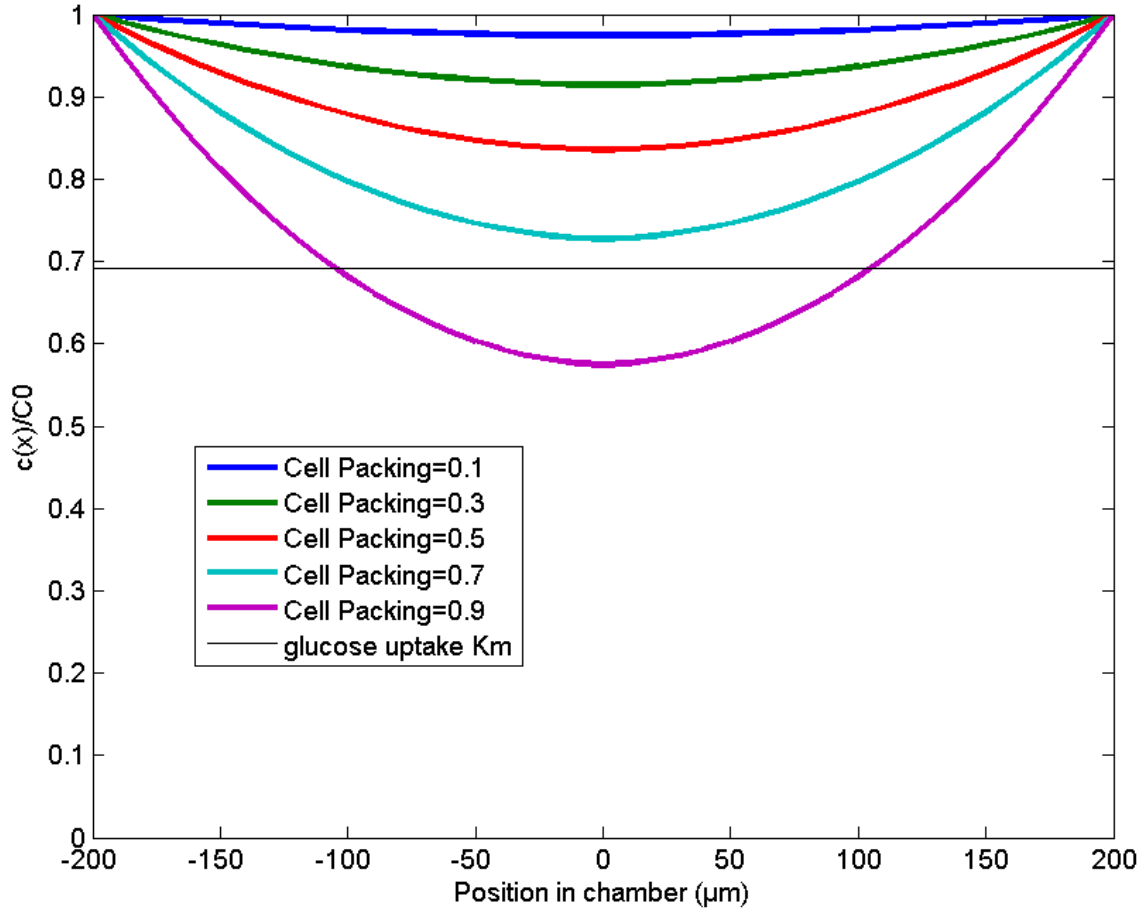



Figure 6: Steady state glucose distribution for various cellular densities for the parallel H-shaped device type C.

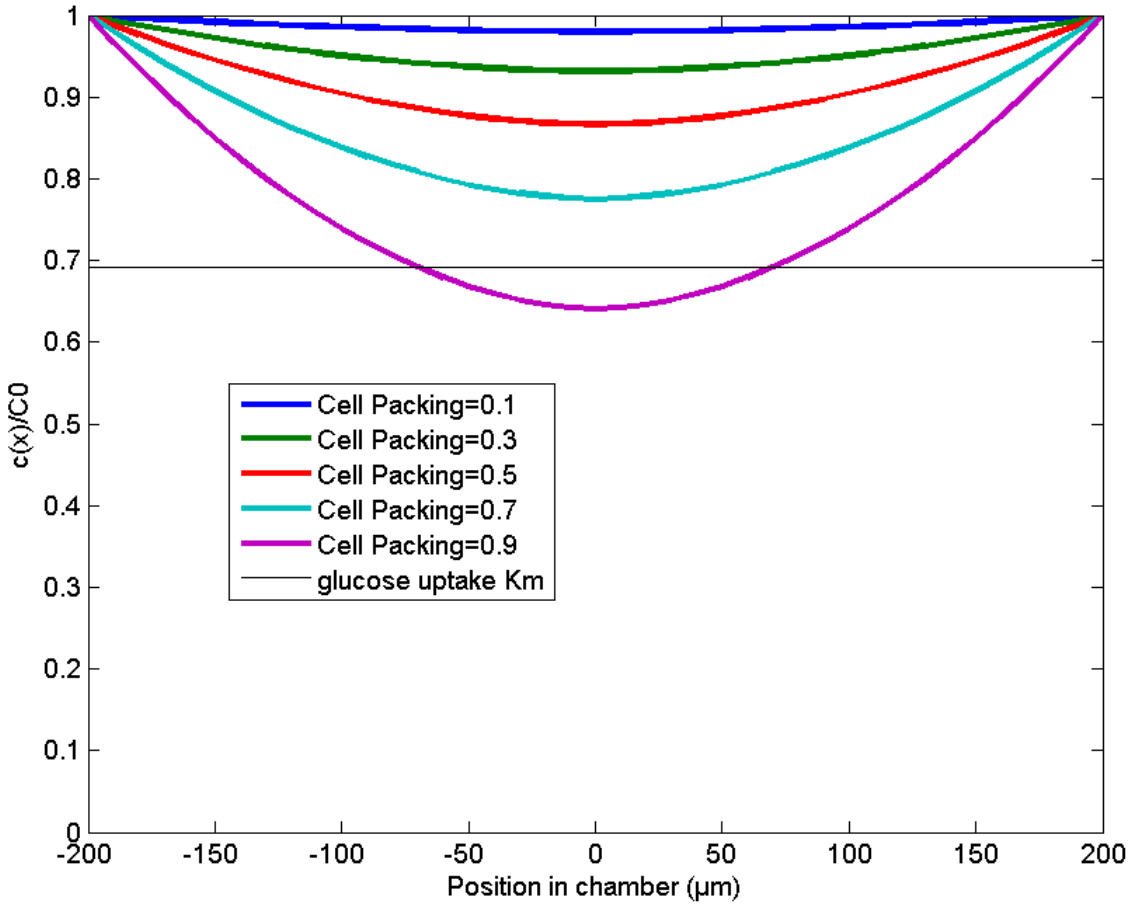



Figure 7: Steady state glucose distribution for various cellular densities for the parallel H-shaped device type A.

## 6. Single-cell parameter estimation of models of gene expression (article, submitted version)

# What Population Reveals about Individual Cell Identity: Single-cell Parameter Estimation of Models of Gene Expression in Yeast

Artémis Llamosi[1,2,†], Andres M. Gonzalez-Vargas[3,†◊], Cristian Versari[4], Eugenio Cinquemani[5], Giancarlo Ferrari-Trecate[3,‡], Pascal Hersen[2,‡], and Gregory Batt[1,‡]

[1] *INRIA Paris-Rocquencourt, 78153 Le Chesnay, France*
[2] *Laboratoire Matière et Systèmes Complexes, UMR 7057, Université Paris Diderot & CNRS, 10 Rue Alice Domon, 75013, Paris, France.*
[3] *Dipartimento di Ingegneria Industriale e dell'Informazione, Università degli Studi di Pavia, via Ferrata 3, 27100 Pavia, Italy*
[4] *Laboratoire d'Informatique Fondamentale de Lille, UMR 8022, Université de Lille 1 & CNRS, 59655 Villeneuve d'Ascq Cedex, France*
[5] *INRIA Grenoble – Rhône-Alpes, 38330 Montbonnot, France*

[†] *These authors contributed equally to this work*
[‡] *These authors contributed equally to this work*
◊ *Current address: Universidad Santiago de Cali, Calle 5 # 62-00, Cali, Colombia*

**Corresponding authors:** Giancarlo Ferrari-Trecate - giancarlo.ferrari@unipv.it, Pascal Hersen - pascal.hersen@univ-paris-diderot.fr, Gregory Batt - gregory.batt@inria.fr

**Key words:** Cell-to-cell variability | Parameter estimation | Gene expression | Mixed-effects model | Cellular identity

**Short title:** Parameter estimation at the single-cell level

# Abstract

Significant cell-to-cell heterogeneity is ubiquitously observed in isogenic cell populations. Consequently, parameters of models of intracellular processes, usually fitted to population-averaged data, should rather be fitted to individual cells to obtain a population of models of similar but non-identical individuals. Here, we propose a quantitative modeling framework that attributes specific parameter values to single cells for a standard model of gene expression. We combine high quality single-cell measurements of the response of yeast cells to repeated hyperosmotic shocks and state-of-the-art statistical inference approaches for mixed-effects models to infer multidimensional parameter distributions describing the population, and then derive specific parameters for individual cells. The analysis of single-cell parameters shows that single-cell identity (*e.g.* gene expression dynamics, cell size, growth rate, mother-daughter relationships) is, at least partially, captured by the parameter values of gene expression models (*e.g.* rates of transcription, translation and degradation). Our approach shows how to use the rich information contained into longitudinal single-cell data to infer parameters that can faithfully represent single-cell identity.

# Summary

Because of non-genetic variability, cells in an isogenic population respond differently to a same stimulation. Therefore, the mean behavior of a cell population does not generally correspond to the behavior of the mean cell, and more generally, neglecting cell-to-cell differences biases our quantitative representation and understanding of the functioning of cellular systems. Here we introduce a statistical inference approach allowing for the calibration of (a population of) single cell models, differing by their parameter values. It enables to view time-lapse microscopy data as many experiments performed on one cell rather than one experiment performed on many cells. By harnessing existing cell-to-cell differences, one can then learn how environmental cues affect (non-observed) intracellular processes. Our approach is generic and enables to exploit in unprecedented manner the high informative content of single-cell longitudinal data.

# Introduction

It is well-recognized that cellular heterogeneities exist in a population of isogenic cells [1–3]. Indeed, cellular processes are noisy and generate cell-to-cell differences. Microfluidics and time-lapse fluorescence microscopy combined with cell-tracking algorithms make it possible to follow the behavior of populations of cells at the single-cell level over long time and to apply stimulations homogeneously [4,5]. Therefore, cell-cell variability in the expression of a gene of interest can be observed over extended time scales. The origins of the variability of biological processes and phenotypes are multifarious. Indeed, the observed heterogeneity of cell responses to a common stimulus is believed to originate partly from differences in cell phenotypes (age, cell size, ribosome and transcription factor concentrations, etc...), from spatio-temporal variations of the cell environments and from the intrinsic randomness of biochemical reactions. A proper assessment and modelling of such heterogeneity is therefore a challenging task since not only it has several sources but also those sources are inter-dependent and act with different strengths and on different time-scales [6].

Regarding dynamical models of gene expression, the most widely-accepted approach to take into account cell-cell variability so far relies on modelling transcription as a stochastic process [7]. Yet, these approaches only give a partial representation of cellular heterogeneity as they assume that all the measured variability originates only from the noisy expression of the modelled genes. The level of expression of other genes and their products, along with the cell's phenotype that emerges from it, are considered as fixed in time and equal for all cells. That is, the standard modeling approach considers all gene expression noise to be intrinsic. Yet, it is known from seminal works on noise in gene expression  that the overall noise breaks down into intrinsic and extrinsic components [8,9]. Although both are always present, intrinsic noise contribution is generally dominant only on short time scales and for unstable or weakly expressed proteins.

Therefore, a purely stochastic representation of cellular heterogeneity is not appropriate for a large proportion of genes and biological processes. Witnessing that validating a model encompassing both types of variability against data is still very difficult given current experimental possibilities [10], we propose to explore a different approach in which variability is represented only as stable differences among cells. This simplifying assumption is a necessary first step towards a congruent representation of the total variability in gene expression, and can be readily applied to other biological processes in which extrinsic variability dominates or when the focus lies on cellular identity.

Here we analyzed the temporal evolution of the level of expression of an inducible fluorescent reporter in a population of yeast cells growing in a microfluidic device. By selecting a strong inducible promoter and using a stable reporter, we placed ourselves in experimental conditions where extrinsic variability is dominant over the neglected intrinsic component. In addition we assess directly how the inferred individuality in gene expression can be related to measurable features of cell's phenotype and physiology and therefore related to typical biological measures of cellular identity. We use a modeling approach in which, for a standard model of gene expression in yeast, each single cell is given specific parameter values while the cell population is described by a multidimensional parameter distribution (Figure 1). This leads to a challenging inference task compared to a classic situation where all cells are described by the same "mean-cell" model and parameters. Indeed the problem is shifted from obtaining a single value per parameter to obtaining parameter values for

each observed cell, as well as a multidimensional distribution representing parameter values in large cell populations. This problem not only involves determining the distribution within a population for each parameter but also their mutual relationships, or more formally, their joint distribution. In order to do so, we used state-of-the-art statistical methods [11,12] that allow inferring parameters distribution across the population that are congruent with parameters attributed to each single cell. We motivate the use of such demanding statistical tools by showing why a simpler and more straightforward method is inappropriate for our current objective of representing populations by a distribution of parameters.

We propose several validations of the inference results and we analyze the obtained parameter distributions representing cell populations. Then we focus on single cells and analyze the correlation across parameters or between parameters and other single-cell features related to phenotypic and physiological variability. At last, the inheritability of the parameters of gene expression is assessed. Taken together, our results demonstrate that using the proposed framework, biologically-relevant model parameters can be attributed to individual cells and related to single-cell features, while the population of cells is represented in a concise manner. As such, this work is an important step towards identifying the major determinants of extrinsic cell-cell variability, as well as introducing quantitatively the concept of single-cell identity.

# Results

**Gene expression in response to repeated osmotic stress shows a high level of variability between cells.**

Using microfluidics and time-lapse microscopy we acquired longitudinal data of the response of individual yeast cells subjected to repeated hyperosmotic shocks (see Material and Methods) [13,14]. Cells were bearing a stable fluorescent reporter driven by the STL1 promoter which is strongly activated by hyperosmotic stress [15,16]. We extracted fluorescence values for large numbers of single yeast cells (typically 300) over a long period of time (typically 8-10 hours). Markedly-different behaviors were observed between individual cells (Fig. 1 and S1). As extrinsic variability is arguably the dominant component of phenotypic heterogeneity in gene expression in eukaryotic cells [17,18], these differences are expected to depend at least in part on variations in the rates of transcription, translation and degradation/dilution from one cell to another. Parameters of a model of our reporter gene expression should therefore be different from one cell to another to account for extrinsic variability. By using short but pronounced and repeated inductions of gene expression with a stable reporter protein, we limited both the impact of intrinsic noise in our experiments and the deleterious effects of hyperosmotic shocks (see Experimental Design in Text S1).

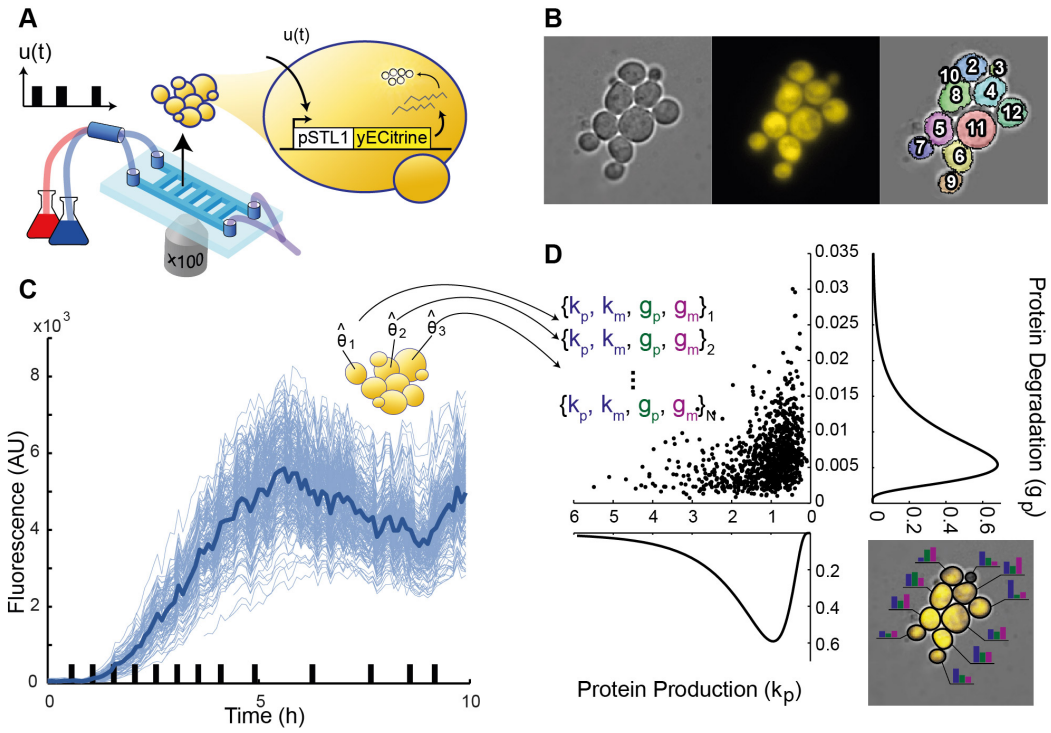



Figure 1: Experimental setup and principle of single-cell parameter estimation. A. **Microfluidic device enabling the growth and imaging of yeast cells over extended durations while applying repeated hyperosmotic shocks by rapidly switching their environment between normal and hyperosmotic media. Using a reporter gene that drives the transcription of the yellow fluorescent protein yECitrine under control of the osmoresponsive promoter pSTL1, one can track the transcriptional response of cells to repeated osmotic shocks.** B. **Thanks to segmentation and tracking algorithms, the response of single-cells can be measured over several generations.** C. **As a result we obtain single-cell trajectories (thin blue lines) that show the variability in cells response to hyperosmotic stress. The thick blue line represents the median behavior. Black bars on the x-axis represent the hyperosmotic shocks applied to cells.** D. **From these trajectories, our goal is to extract the parameters of a standard model of gene expression (see text) for each cell, and therefore a multidimensional distribution describing the cell-to-cell variability. As an illustration, the right inset shows that different cells will be modeled with different parameter values to account for their own specific behavior.**

**Mixed-effects model is an ideal framework for representing extrinsic variability.**

Mixed-effects (ME) models are a class of statistical models introduced to describe the response of different individuals within a population to known stimuli. Here, we used a ME model where the response of individual cells was described in terms of a simple dynamical model of gene expression. Denoting with $m$ and $p$ the cellular level of mRNA and fluorescent protein, respectively, we have

$$\begin{cases} \dot{m}(t) = k_m u(t) - g_m m(t) \\ \dot{p}(t) = k_p m(t) - g_p p(t) \end{cases}$$

where $u(t)$ represents the activity of transcription factors – in our case, the phosphorylation and nuclear import of the kinase Hog1p – and is a function of the osmolarity of the cell environment (see Material and Methods and Text S1). The production and decay rates are denoted $k_m$ and $g_m$ for the mRNA, and $k_p$ and $g_p$ for the protein, respectively. To relate fluorescence measurements to actual protein concentrations, we accounted for protein folding time using a delay $\tau$. We also assumed the presence of multiplicative and additive white Gaussian measurement noise whose strength is the same for all cells (see Text S1 and Table S1 for details). Importantly, in the ME framework, it is considered that $k_m$, $g_m$, $k_p$, and $g_p$ vary within the population. Differences in parameter values may

typically originate from differences in the level of key components of the gene expression machinery (*e.g.* RNA polymerase and ribosomes) or in environmental or physiological parameters (*e.g.* cell growth rate). We assumed that these parameters were log-normally distributed across the population: $\theta = (k_m, g_m, k_p, g_p)$ with $\ln(\theta) \sim \mathcal{N}(\mu, \Sigma)$, where $\mu$ and $\Sigma$ correspond to a vector of means and a covariance matrix, respectively. This assumption ensures the population is represented in a much more concise and general manner than what would be possible by only representing a population by the dynamics of every cell observed in an experiment.

Here, we are looking for a multidimensional distribution defined by its center of mass (*i.e.* a vector of mean values) and its spread (*i.e.*, a covariance matrix) across the population. A simple, intuitive manner to tackle this problem is to search for the different parameter values that best describe each individual cell, and then compute the statistics (mean and covariance) of the underlying distribution from the set of parameter estimates. We refer to this method as the '*naive approach*' since it is the natural starting point, bearing limitations that are not apparent until a proper analysis is performed. The proposed alternative is to use state-of-the-art approaches for the identification of ME models, such as Stochastic Approximation Expectation Maximization (SAEM). SAEM is a stochastic approximation version of the well-known Expectation–Maximization algorithm and has been developed for the inference of population models in presence of limited available information [11,19]. Notably SAEM is the reference approach in pharmacokinetics/pharmacodynamics studies [12,20]. However, it has not yet been applied to time-lapse single-cell data. The SAEM algorithm directly searches for multivariate distributions by alternating (i) an estimation of (an approximation of) the likelihood of the population parameters and individual observations given the current best estimate of the parameter distribution in the population and (ii) an update of the current estimate of the parameter distribution. In a second step, *a posteriori* estimates of the individual cell parameters are obtained from the inferred parameter distribution and individual data (maximum *a posteriori* estimate, MAP). This way, the fact that all parameters share (hidden) traits of the common population is explicitly taken into account. The naive and SAEM approaches are graphically represented in Fig. S2.

**The SAEM approach provides relevant and robust single-cell parameter distributions.**

Both the *naive approach* and the SAEM estimation method were applied to an experimental data set comprising more than 300 cells observed during several hours. Despite the significant diversity in the behavior of individual cells (Fig. 2A), both the *naive approach* and the SAEM estimation method were able to find single-cell parameters that fitted well the set of observed single-cell behaviors (Fig. 2B,C). For the naïve approach, one can observe that the envelope of the fitted trajectories is slightly larger than the data at the early time points (Fig. 2C). This simply results from the absence of data to constrain the fits at the early times for cells born during the experiment. Indeed, the average relative absolute difference between single-cell predictions and data are nearly identical in the two approaches (naïve approach: 8.7%; SAEM approach: 8.3%).

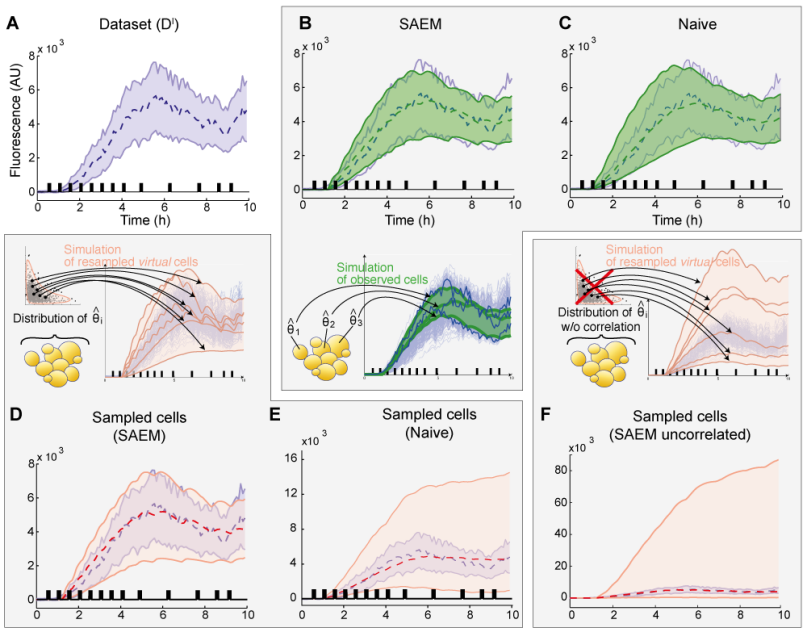



Figure 2: The SAEM approach provides parameter distributions that capture the population behavior because of cross-correlations between parameters**. A. Representation of the experimental dataset. B. Simulated behavior obtained when using the parameters of each observed cell in the dataset (325 cells) inferred with the SAEM approach. C. Simulated behavior obtained when using the parameters of each observed cell in the dataset (325 cells) inferred with the naive approach. D. Simulated behavior of 10000 cells when resampling the population joint distribution inferred with SAEM, (pink). E. Simulated behavior of 10000 cells when resampling the population joint distribution inferred with the naive approach. F. As an illustration we show the simulated behavior of 10000 cells when resampling the population parameter distribution as in D but without preserving the covariance between parameters (*i.e.,* using marginal distributions). For E and F, note that the y-axis has been scaled differently. Shaded areas represent the fluorescence values of 95% of the population and the dashed lines represent the median. Experimental data is represented in blue. Black bars indicate the presence of osmotic shocks. Note that unlike actual cells, all simulated cells are represented during the whole experiment (i.e. from 0 to 10hrs).**

We then evaluated the capability of the obtained parameter *distributions* to actually describe the behavior of the cell population (mean and spread). To do so, the parameter distributions obtained using the *naive* and the SAEM approaches were randomly sampled, thus creating two different virtual '*cell populations*', and the two corresponding sets of behaviors were computed from our model of gene expression. The SAEM-inferred parameter distribution accurately reproduced the observed behavior of the real cell population (Fig. 2D), whereas the *naive approach* failed to do so (Fig. 2E). Therefore, although both approaches were able to identify a set of single-cell parameters that reproduce well the behaviors of the set of observed cells, only SAEM was able to infer a parameter distribution at the population level consistent with the observed heterogeneity in gene expression.

To investigate the causes of the marked differences between the predictive power of the ME models inferred using either the naive approach or the SAEM algorithm, we compared the corresponding parameter distributions. In both cases, the mean values of the parameters were comparable and within the expected ranges (see Table S1 for parameter values and Text S1 for literature values). However, the distribution obtained with the SAEM algorithm was significantly more compact (*i.e.* it had a smaller volume in the parameter space) and was more structured (*i.e.* it had higher cross-correlations on average; Fig. S3). This strongly suggested that capturing the structure of the parameter distribution is essential in order to explain the population behavior. Both the individual statistics of each parameter, and their covariance, describing mutual relationships, contain essential

information to properly account for the cell-cell variability observed in the dataset. And indeed, when using a parameter distribution with the same individual parameter statistics (mean and variance) as the distribution inferred using SAEM but with null cross-correlations (*i.e.* using the marginal distributions), the model lost its capability to predict the behavior of the population (compare Fig. 2D and 2F). Our understanding is that in the *naive approach*, all cells are fitted individually and are subsequently *casted* into a multidimensional distribution. In contrast, SAEM allows finding equally good single-cell parameters while favoring a concise multidimensional representation of the overall population. The difference in performance between these two approaches is rooted in the fact that even with a simple model of gene expression the information contained in a single trajectory is too small to constrain the inferred parameter values in a satisfactory way. Using SAEM, we actually allow each single-cell fit to use information about the overall population, which ensures coherence between the representation of the population by distributions and of the single cells by specific parameter values. Having demonstrated that the SAEM-based identification approach captures the behavior of the cell population, from here on we focus only on the results obtained using this method.

We then tested the robustness of the inference approach which is an essential property for learning algorithms. Interestingly, the performance of the SAEM inference method degraded gracefully as the number of available single-cell trajectories for identification was decreased to as few as 32 cells (Fig. 3A and Text S2), and also as the experimental time period used for learning was reduced (Fig. 3B and Text S2). Lastly, ME models with SAEM-inferred parameter distributions were still able to give good predictions when tested on a different data set (Fig. 3C, see also Text S3).

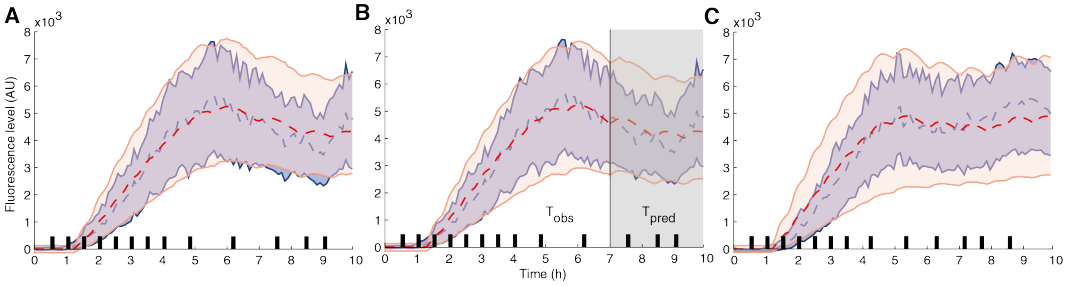



Figure 3 Robustness of the SAEM approach and validation of model predictive power**. A Predictions obtained for a ME model having parameter distributions estimated on only 32 randomly-chosen cell trajectories (see also Text S2). B Predictions obtained for an ME model having parameter distributions estimated using only the first 7 h of the experimental data (see also Text S2). C Prediction obtained for the validation dataset for a ME model with parameter distributions estimated using the identification data set. Different temporal patterns of osmotic shocks were applied.**

**Parameters of the gene expression model only make sense at the single-cell level.**

At this point, we have showed how to efficiently and robustly extract the distributions of parameters of a standard model of gene expression from a collection of longitudinal single-cell data, and a set of parameters for each cell in the population. While we are here mostly interested in the details of the parameter distribution, we can also extract the average value for each parameter of the model. Importantly, they are different from the parameters that are obtained by fitting directly our model of gene expression to the population-averaged behavior. This is illustrated on Figure 4 where the '*average cell*' trajectory (whose parameters are the average of single-cell parameters) is different from the average trajectory (obtained by directly averaging the single-cell trajectories). As mentioned in the introduction, this expected result reminds us that parameters of a model of a biological

process estimated from average behaviors, as done in the vast majority of quantitative studies, may poorly represent the actual process.

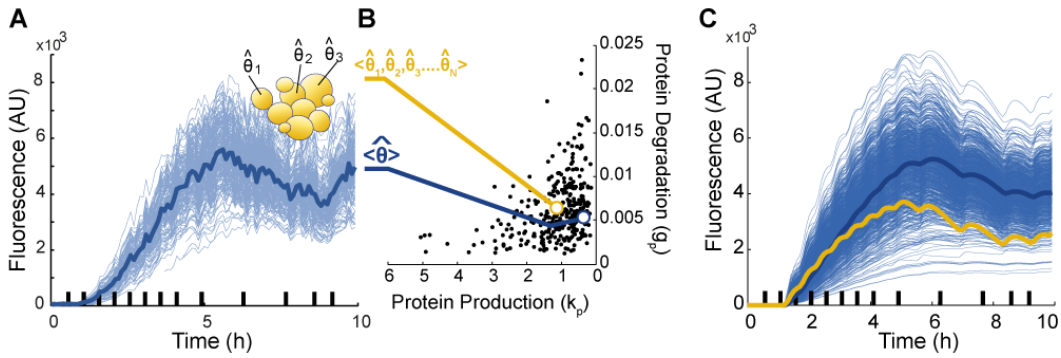



Figure 4: Parameters only make sense at the single-cell level. **A-B. Starting from an experimental dataset (A), one can either extract the parameters that describe the average behavior (in blue), or use our framework to extract the entire collection of single-cell parameters (black dots in B) and compute the average parameters (in yellow). B-C. The average parameters do not match the parameters that best describes the average behavior. C. Visualization of 1000 simulated single-cell behaviors (blue thin lines) based on the parameters distributions shown (partially) in B. The solid blue line is a (good) simulation of the average behavior (also shown in blue in panel A). The yellow solid line is the behavior corresponding to the "average cell", which has for parameters, the average parameters of the parameters distributions. The "average cell" behavior is clearly different from the averaged behavior.**

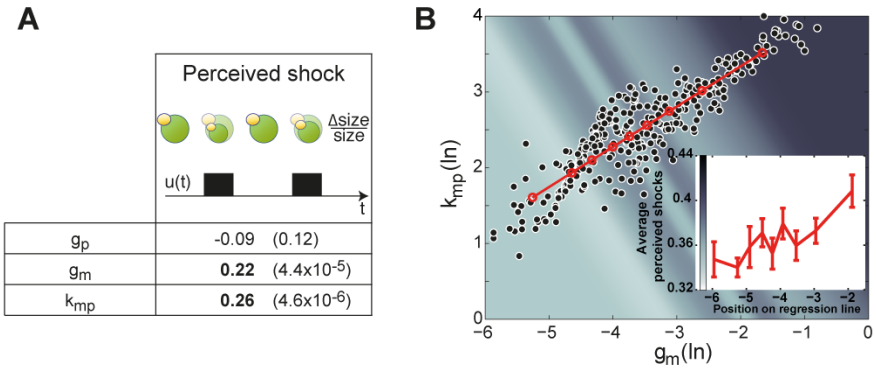
**Analysis of parameter correlations may reveal non-identifiability relations.**

Non-identifiability arises when the information contained in data along with a model structure does not allow for the proper estimation of parameter values: several parameter values (or more usually combinations of parameter values) yield equally-good results given the available data. In our framework, very high correlations between parameter values may indicate the existence of non-identifiability relations among parameters. The first application of the SAEM algorithm showed that $k_m$ and $k_p$ were highly correlated, and, indeed, checking single-cell values suggested that the rates of transcription and translation could hardly, if at all, be quantified independently. A detailed identifiability analysis showed that, at the level of individual cells, these two parameters are structurally non-identifiable; only their product can be quantified (Text S4). However, in population approaches, partial information about the second-order statistics of individual parameters can be inferred from the population statistics even if these parameters are non-identifiable at the single-cell level (Text S5). Consequently, to address identifiability issues while preserving maximal information, we fixed the mean value for $k_p$ when inferring parameter distributions using SAEM, and introduced the protein production rate $k_{mp}$, defined as the product of $k_m$ and $k_p$, for the single-cell models. With these changes, shrinkage was then found to be negligible (Text S4).

**Single-cell parameters correlate with the intensity of shocks perceived by single cells.**

Having identified single-cell parameter values, one may wonder whether they can be used to retrieve known facts or discover new ones on the physiology of the cell response to hyperosmotic shocks. In our model, hyperosmotic shocks affect all cells identically. However, in the microfluidic device, the intensity of the shock perceived by different cells varied, as evidenced by differences in the reduction of cellular volume following shocks. Therefore, one should find that protein production parameters inferred for the most severely impacted cells are statistically higher than average. We thus estimated the perceived shock intensities as the time-averaged reduction of cellular volume following shocks, and compared for all the cells the inferred parameter values and the perceived shock intensities. We

found a strong correlation between protein production rates and shock intensities in agreement with our hypothesis. Moreover an equally-strong correlation was also found with mRNA degradation rates (Fig. 5A). This second feature, obtained by our framework without any additional measurements or hypothesis, is consistent with the known global destabilization of mRNAs observed after hyperosmotic shocks [21]. Lastly, the simultaneous increase of protein production rates and mRNA degradation rates strongly correlates with the increase of the perceived shock (Fig. 5B) indicating that these two processes are jointly regulated in response to hyperosmotic shocks. Note that the direct experimental identification of such co-variations would be very challenging. This shows the interest of extracting and analyzing distributions of model parameters for the identification of joint regulations.




| | |
|---|---|
| $g_p$ | -0.09 (0.12) |
| $g_m$ | **0.22** ($4.4\times10^{-5}$) |
| $k_{mp}$ | **0.26** ($4.6\times10^{-6}$) |

Figure 81 : Effects of hyperosmotic shocks on intracellular processes involved in gene expression. **A. Correlations between the perceived intensity of hyperosmotic shocks and single-cell parameter estimates are provided with their corresponding *p*-values (Text S1). B. Estimated values for protein synthesis rates $k_{mp}$ and mRNA degradation rates $g_m$ for each individual cell. Their strong correlation (Spearman coefficient: 0.88; *p*-value<$10^{-15}$) together with their mutual increase with perceived shocks intensity indicates that these two processes are jointly regulated in response to hyperosmotic shocks. Insert plot and colored background represent perceived shock intensity for 9 groups of 35 cells along the regression line.**

**Single-cell parameters correlate with single-cell physiological features.**

In addition to hyperosmotic shocks, several features related to the cell physiology or local environment are also expected to relate to gene expression [22]. Such features notably include cell division rate, cell size, cell age, and local cell density. Since these features can be measured or estimated for each single-cell based on bright-field time-lapse imaging, one can again harness cell-to-cell variability and search for relations between these features and the parameters that describe intracellular processes involved in gene expression. Firstly, we searched for a correlation between the protein decay parameter, $g_p$, and the cell division rate. Indeed, as the fluorescent reporter we used has a long half-life and photobleaching is negligible (see Initial parameters values Text S1), one should expect that its observed decay comes mostly from dilution due to cellular growth. Therefore, we quantified for each cell its division rate, averaged over the observation period (Text S1) and, as expected, found a significant positive correlation between the measured average single-cell division rate and the protein decay parameter $g_p$ (Fig. 6). Stated differently, using exclusively the fluorescence profile of individual cells and the inferred parameter distribution for the cell population as an *a priori*, the inference approach attributed statistically higher dilution rates to cells that grow faster. Several other highly significant correlations between single-cell parameters and the above-mentioned single-cell measured features were observed (Fig. 6). Note that all measured features were averaged across time to allow the comparison with the time-invariant model parameters (Text

S1). Although it is difficult to attribute in a systematic manner a direct and unambiguous biological interpretation of the observed correlations between coarse-grained model parameters and cell features, one can nevertheless observe (i) that cell density appears to have a pronounced influence on the protein production rate, suggesting that - even in microfluidic growth chambers - the environment of the cells should not be assumed to be perfectly homogeneous, and (ii) that the correlations of the protein production rates and mRNA degradation rates with every measured feature always have the same sign, corroborating the presence of mechanisms for the joint regulation of these processes in our system.

More generally, one wonders how the different measured cell features relate to the overall (multivariate) parameter variability. We conducted a principal component analysis (PCA) of the set of inferred single-cell parameter values. This yielded a new parameterization of the model (new parameters being called principal components PC1, PC2 and PC3) that is particularly relevant to investigate variability as, unlike natural parameters, each principal component is uncorrelated to the others. The analysis showed that the first two components PC1 and PC2 represented 87% and 12%, respectively, of the overall variance in single-cell parameter values, and that these principal components correlated very significantly with measured cell features. We then ranked the various features based on their correlation with the variability captured by the inferred ME model. For a given feature, this is defined as the weighted average correlation with the different PCs, with weights equal to the importance (*i.e.*, variance) of every PC. It appeared that local cell density was the most important factor (average correlation: 0.23), followed by cell size (0.21) and the division rate (0.2). To our knowledge, there is no established direct connection between local cell density and gene expression in yeast. It would be interesting to investigate this connection at the molecular level. Quite surprisingly, from our data, age was not associated with a significant variability in parameter values. Taken together, our results show that, for quantitative studies, features other than colony growth rate should be taken into account. A natural extension of this study would be to investigate how the inclusion of these features in the model, seen as covariates, could improve single-cell predictions.

| | Mean Density | Mean Division rate | Mean Size | Mean Age |
|---|---|---|---|---|
| $g_p$ | -0.03 (0.61) | **0.20** ($2.1\times10^{-4}$) | 0.04 (0.44) | -0.10 (0.16) |
| $g_m$ | **0.21** ($3.5\times10^{-4}$) | **-0.23** ($2.6\times10^{-5}$) | **-0.21** ($1.3\times10^{-4}$) | -0.06 (0.39) |
| $k_{mp}$ | **0.35** ($4.4\times10^{-11}$) | **-0.19** ($7.4\times10^{-4}$) | **-0.30** ($2.4\times10^{-8}$) | **-0.19** ($2.7\times10^{-3}$) |
| PC1 (87%) | **0.22** ($7.2\times10^{-5}$) | **-0.23** ($4.2\times10^{-5}$) | **-0.21** ($1.5\times10^{-4}$) | -0.05 (0.46) |
| PC2 (12%) | **0.34** ($4.5\times10^{-10}$) | 0.05 (0.35) | **-0.24** ($1.4\times10^{-5}$) | **-0.24** ($1.5\times10^{-4}$) |
| PC3 (<1%) | **-0.13** (0.02) | **-0.12** (0.03) | -0.02 (0.76) | 0.00 (0.98) |

Figure 6: Harnessing cell-to-cell variability reveals correlations between parameter values and independently-measured cellular features. **Local cellular density, division rate, size and age were quantified with single-cell resolution (Text S1). Correlations between these single-cell features and the single-cell parameter estimates and their principal components are provided with their corresponding *p*-values. Note the expected correlation between protein degradation/dilution rate $g_p$ and the cell division rate. The proportion of variance accounted for by each principal component is indicated in parenthesis.**

**Single-cell parameters are partly inherited from mother to daughter.**

Finally, we investigated inheritance of single-cell parameters. Statistical tests showed that the parameters of mother and daughter cells were significantly closer to each other than the parameters of random cell pairs (Text S1 and Fig. S4). However, this comparison does not exclusively test the effect of lineage. The fact that mother and daughter cells share a similar environment may also explain this result. To study the specific influence of lineage, we compared the parameter values between pairs of cells that either were mother and daughter (related mother/daughter pairs) or were a mother and the unrelated daughter of another mother cell (non-related mother/daughter pairs), with all cells growing in the same microfluidic chamber so as to limit environmental bias. As shown in Fig. 7, the parameter values of individual cells were statistically closer to the parameters of their own mother cell than to the parameters of another mother cell. It appears that parameter values are 16% (resp. 14%, 10%) closer in genuine mother/daughter pairs for $g_p$ (resp. $g_m$, $k_{mp}$). Although mild in absolute terms, bootstrapping testing showed the presence of a statistically strong inheritance effect (*p*-values < $10^{-15}$ for all parameters, Text S1). Importantly, we verified using a more restrictive notion of nMD pairs that the observed inheritance effect was not due to the fact that mother and daughter cells have more similar mean densities on average than nMD cells since the former share the same environment. Interestingly, we also found that daughter cells are on average 14% more sensitive than their mothers and that the intensity of the perceived shocks is anti-inherited: the most resistant mothers have the most sensitive daughters, and conversely.

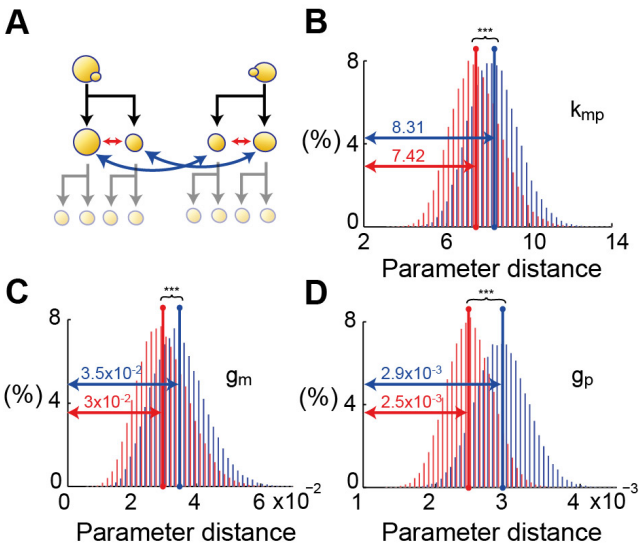



Figure 7: Parameter values of individual cells are statistically closer to the parameters of their own mother than to the parameters of another mother cell. **(A) The distance between parameters of related mother and daughter cells (MD) and non-related mother and daughter cells (nMD) were compared. (B-D) Distribution for each parameter of the average distance between 40 pairs of MD (red) and nMD (blue) for 50000 combinations obtained by bootstrapping (Text S1). All parameters are closer between mothers and daughters than on average (*** p-values < $10^{-15}$).**

# Discussion

In this work, we proposed an approach for capturing the biological variability observed in single-cell time-lapse microscopy experiments by *distributions* of parameters. By doing so, we address a fundamental issue encountered in the vast majority of quantitative studies where parameters of deterministic or stochastic models of intracellular processes make sense at the single-cell level but are estimated for a virtual 'mean cell'. The analysis was based on the mixed-effects (ME) modeling framework and two inference approaches were evaluated. The relevance of the ME framework for

modeling biological processes has been recently recognized [23,24]. The use of advanced statistical methods, like SAEM, was essential to properly capture the variability of the biological parameters across the population in a simple manner, including most notably the correlation among them. In addition, we showed that the SAEM method scales to real-life problems and provides robust results. With this approach, the information on each and every cell is jointly used to calibrate the model parameter distribution, alleviating the problem of limited observability and noisy observations encountered at the individual-cell level. We then demonstrated the biological relevance of the inferred cell-specific parameters, as they were partly inherited from mother to daughter cells and correlated with independently-measured single-cell features.

Our approach is adapted to calibrate models explicitly accounting for extrinsic variability. From a mechanistic viewpoint, two components of biological variability, termed intrinsic and extrinsic noise, have been proposed. For a given cellular process, intrinsic variability is mostly related to fast fluctuations coming from stochasticity in molecular reactions while extrinsic variability includes more stable cell-to-cell differences in intracellular and extracellular environments [25,8,17]. Thanks to recent methodological developments, such as finite state truncation methods, significant progress have been made in the identification of intrinsic noise models, in particular for the study of gene expression [26]. Such models assume that the different observations arise from different realizations of the same stochastic process and, therefore, are still based on the notion of a virtual mean - although noisy - cell. In comparison, and despite recent methodological developments [27,28], few attempts have been made to infer extrinsic noise models from data, see [4,10,23,29,30] and our previous work [32]. We refer the reader to Karlsson et al. [24] for a detailed discussion of these works. This is surprising, given the fact that extrinsic noise has been shown to be the dominating component in many biological systems [17,18,31] and that application of cell population models has proven extremely useful, notably to explain cell decision processes [3]. Moreover, with the notable exceptions of Zechner *et al* [10] and Gonzalez *et al* [32], no method that exploits single-cell time-lapse data for the identification of cell population models has been able to predict population behaviors. Interestingly, Zechner *et al* [10] proposed a very general framework capturing intrinsic and extrinsic variability by using a stochastic model based on the chemical master equation with parameter distributions. They investigated whether this modeling framework was able to capture both noise components appropriately, all of the extrinsic variability being aggregated into a unique cell-dependent parameter. Here, we pursued a different objective. We focused on extrinsic noise and investigated whether multidimensional parameter distributions provide an accurate description thereof and can be inferred from the available experimental data, whether the inferred single-cell parameter values are biologically-relevant, and how extrinsic noise is distributed across different cellular processes. Given the identifiability issues encountered already on relatively simple ME models, one might wonder whether more complex models combining the use of a stochastic interpretation of the reactions and of distributions for all (or most of) the parameters can be accurately identified based on available experimental data. Another attractive possible extension of the mixed-effect framework is to replace the purely static description of cell-to-cell differences obtained by using different, time-invariant parameter values by a more dynamical representation using reaction parameters that slowly fluctuate in time. This can typically be done by accounting for the stochastic turnover of the proteins underlying the various reactions involved in the processes of interest [33].

The possibility of identifying single-cell models opens new perspectives. Indeed, our results support the approach advocated by Pelkmans and coworkers *(18)* in which "*studying cell-to-cell variability [...] will increase our understanding of how cellular activities are embedded in the physiology of a cell.*" Following what we have shown here, one could dissect the variability of the different cellular processes involved in a particular phenotypic response and search for correlations with different cellular processes and with environmental factors. Such rich information on the integrated functioning of cells is otherwise barely accessible. More fundamentally, single-cell modeling provides a quantitative tool to study the notion of *cell identity*, as it offers a quantitative description of cell-to-cell differences. Lastly, to which extent this increased knowledge can be used to improve our ability to predict and ultimately control single-cell behavior is a question of interest for both the systems and synthetic biology communities [14,34–36].

# Materials and methods

**Yeast strain and microscopy.** All experiments were performed using a STL1::yECitrine-HIS5, Hog1-mCherry-hph yeast strain derived from the S288C background [14]. Cells were cultured overnight in synthetic complete (SC) medium at 30°C, in a shaking incubator at 250 rpm, and then the cultures were diluted in SC so as to reach an optical density of ∼0.2 in 4h. Exponentially-growing cells were injected into a home-made microfluidic device [14]. Liquid medium was flowed using a peristaltic pump (IPC-N, Ismatec) placed after the microfluidic device (flow rate: 120µL/min). A computer-controlled three-way valve (LFA series; The Lee Company) was used to select between normal medium (SC) or the same medium supplemented with 1M sorbitol. The microfluidic chip was made by casting polydimethylsiloxane (PDMS; Sylgard 184 kit; Dow Corning) on a master wafer (made by soft lithography), curing it at 65 °C overnight, pealing it off, and bonding it to a glass coverslip after plasma activation. The device has 5 chambers of 200x400x3.6 µm where cells are imaged. These chambers are connected to larger channels where medium flows such that the environment of the imaging chamber is changed by diffusion only (see [14]). After having loaded cells in the device, we leave them to rest with SC flowing for 30 min before starting the experiment. A switch of the valve state did not lead to an instantaneous change of the cells' environment inside the microfluidic device: ~2 min were needed for the fluid to pass from the valve to the channels and the imaging chamber.

The cells were imaged using an automated inverted microscope (IX81; Olympus) equipped with an X-Cite 120PC fluorescent illumination system (EXFO) and a QuantEM 512 SC camera (Roper Scientific). The temperature of the microscope chamber, which also contains the media reservoirs, was constantly held at 30°C by a temperature control system (Life Imaging Services). All of these components were driven by the open-source software µManager which was interfaced with Matlab. Images were taken using a 100× oil immersion objective (PlanApo 1.4 NA; Olympus). The fluorescence exposure time was 200 ms, with fluorescence illumination intensity set to 50% of maximal power. The fluorescence exposure time was chosen such that the fluorescent illumination did not cause noticeable effects on cellular growth over extended periods of time. Importantly, illumination, exposure time, and camera gain were not changed between experiments, and besides background and auto-fluorescence subtraction (defined as the minimum intensity in the first frame), no data renormalization or processing was done. Imaging was performed at a frequency of one

frame every 3 min for bright-field and one frame every 6 min for fluorescence measurements. The duration of the experiments was 10 hours.

**Measurements of gene expression and physiological features at the single-cell level.** Single-cell gene expression profiles were obtained in two experiments: one for identification ($\mathcal{D}^I$; 325 single-cell trajectories) and one for validation ($\mathcal{D}^V$; 166 single-cell trajectories). The randomly-generated profiles of hyperosmotic stresses differed in each experiment. Image analysis was performed using a home-made segmentation and tracking tool, CellStar. After observing that newly-detected cells usually corresponded to buds still attached to their mother for a long period of time after detection and might present fluorescence quantification artifacts (due to their small size and variable focus), we discarded the information obtained during the first two hours for new cells. Only cells imaged for more than 5 h were selected for identification and validation. The average size of a cell corresponds to its size measured at each time point in bright-field images and averaged over all time points. Average cell age and density were defined analogously. The density of the environment of a single cell was defined as the area occupied by neighbor cells relative to the area of the neighborhood of the cell. The neighborhood was defined as a disk with a radius corresponding to five times the radius of a typical cell. The relative changes in the size of the cells caused by budding events were used to estimate single-cell division times from bright-field images and compute the average cell specific division rate. After automated segmentation and tracking, lineage was manually extracted from the microscopy images. More details are provided in Text S1.

**Single-cell models and ME population models.** We assumed that the transcription factor activity, $u(t)$, depends on the osmolarity effectively sensed by the cells inside the microfluidic chambers, $u_c(t)$, which itself depends on the valve status, $u_v(t)$ (Text S1). To relate fluorescence measurements to actual protein concentrations, we accounted for protein maturation time using a delay $\tau$ and assumed the presence of multiplicative and additive measurement noises that are white and Gaussian (Text S1). A mixed-effects population model is then obtained from single-cell models by assuming that the parameters of the population of cells follow log-normal distributions. More details on the modeling assumptions are provided in Text S1.

**Inference of single-cell and ME population models.** Two methods were proposed to infer ME population models: a naive approach and SAEM. The naive approach used the local optimization algorithm fminsearch from Matlab to maximize the (log-)likelihood of the parameters tested, given the observed data for the considered cell. The parameter distribution for the ME model is then defined based on the set of single-cell parameters. The SAEM approach aims directly at maximizing the likelihood of the population (high-level) parameters describing the distributions of the model parameters, given all the single-cell data. We used the SAEM implementation of Monolix software. Lastly, having inferred a distribution for the model parameters of a population of cells, one could estimate the most likely parameter values for each single cell (ME single-cell models). We used the local optimization tool fminsearch from Matlab to implement a maximum *a posteriori* approach. For more details on the parameter inference approach see Text S1.

**Relating the specific intracellular processes involved in gene expression with other, non-modeled cellular properties.** The analysis of the correlations between the perceived shocks or the single-cell measured features and the estimated parameters was performed using the Spearman coefficient of correlation. The significance of the correlations (*p*-values) was assessed using the standard two-tailed

test implemented in the Matlab statistics toolbox. To test whether parameters of mother and daughter cells were statistically closer than on average, we constructed pairs of cells that differed solely by whether they were direct relatives (mother/daughter pairs, MD pairs) or not (non-related mother/daughter pairs, nMD pairs). The comparison of the mean distance between MD pairs and nMD pairs was performed by bootstrapping (Text S1).

# Supplementary Information

Supplementary information contains details on data analysis and modeling framework (Text S1), an analysis of the robustness of population prediction that extends the results shown in figure 3 (Text S2), a discussion on the validation of population predictions (Text S3), an identifiability analysis (Text S4), a discussion on learning the statistics of non-identifiable parameters (Text S5), four figures to support the results presented in the main text, and one table that regroups the different parameter distributions obtained in this work.

# Acknowledgements

The authors acknowledge Jean-Marc Di Meglio, Benoit Sorre, and Hidde de Jong for insightful discussions. A.L. is grateful to the doctoral program Frontiers of Living Systems.

# Funding

This work was supported by the Agence Nationale de la Recherche (ICEBERG-ANR-10-BINF-06-01), the Who am I? Laboratory of Excellence (ANR-11-LABX-0071 and ANR-11-IDEX-0005-01), the Interdisciplinary Program of Sorbonne Paris Cité (SPC), the C'Nano program from the region Ile de France, and the EU Seventh Framework Program (FP7/2007-2013) under grant agreement n° 257462 HYCON2 Network of excellence.

# Author Contributions

AL performed experiments, image processing, and inheritance and correlation analysis. AG performed parameter search and model validations. CV contributed software for automated image analysis. EC performed identifiability analysis. GFT, PH and GB coordinated the research. AL, AG, EC, GFT, PH and GB analyzed the results and wrote the manuscript.

# Conflict of Interest

All authors declare that they have no conflict of interest.

# References

1. Balázsi G, van Oudenaarden A, Collins JJ. Cellular decision making and biological noise: from microbes to mammals. Cell. 2011;144: 910–25.

2. Raj A, van Oudenaarden A. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. Cell. 2008;135: 216–226.

3. Spencer SL, Gaudet S, Albeck JG, Burke JM, Sorger PK. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. Nature. 2009;459: 428–32.

4. Spiller DG, Wood CD, Rand DA, White MRH. Measurement of single-cell dynamics. Nature. 2010;465: 736–45. doi:10.1038/nature09232

5. Locke JCW, Elowitz MB. Using movies to analyse gene circuit dynamics in single cells. Nat Rev Microbiol. 2009;7: 383–92. doi:10.1038/nrmicro2056

6. Huang S. Non-genetic heterogeneity of cells in development: more than just noise. Development. 2009;136: 3853–3862. doi:10.1242/dev.035139

7. Paulsson J. Models of stochastic gene expression. Phys Life Rev. 2005;2: 157–175. doi:10.1016/j.plrev.2005.03.003

8. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science. 2002;297: 1183–1186.

9. Raser JM, O'Shea EK. Noise in gene expression: origins, consequences, and control. Science. 2005;309: 2010–3. doi:10.1126/science.1105891

10. Zechner C, Unger M, Pelet S, Peter M, Koeppl H. Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. Nat Methods. 2014;11: 197–202. doi:10.1038/nmeth.2794

11. Kuhn E, Lavielle M. Maximum likelihood estimation in nonlinear mixed effects models. Comput Stat Data Anal. 2005;49: 1020–1038. doi:10.1016/j.csda.2004.07.002

12. Lavielle M. Mixed Effects Models for the Population Approach. CRC Press; 2014.

13. Hersen P, McClean MN, Mahadevan L, Ramanathan S. Signal processing by the HOG MAP kinase pathway. Proc Natl Acad Sci U S A. 2008;105: 7165–70. doi:10.1073/pnas.0710770105

14. Uhlendorf J, Miermont A, Delaveau T, Charvin G, Fages F, Bottani S, et al. Long-term model predictive control of gene expression at the population and single-cell levels. Proc Natl Acad Sci U S A. 2012;109: 14271–14276. doi:10.1073/pnas.1206810109

15. O'Rourke SM, Herskowitz I. Unique and redundant roles for HOG MAPK pathway components as revealed by whole-genome expression analysis. Mol Biol Cell. 2004;15: 532–542. doi:10.1091/mbc.E03

16. Ferreira C, van Voorst F, Martins A, Neves L, Oliveira R, Kielland-Brandt MC, et al. A member of the sugar transporter family, Stl1p is the glycerol/H+ symporter in Saccharomyces cerevisiae. Mol Biol Cell. 2005;16: 2068–2076. doi:10.1091/mbc.E04

17. Raser JM, O'Shea EK. Control of stochasticity in eukaryotic gene expression. Science. 2004;304: 1811–1814.

18. Pedraza JM, van Oudenaarden A. Noise Propagation in Gene Networks. Science. 2005;307: 1965–1969.

19. Delyon B, Lavielle M, Moulines E. Convergence of a stochastic approximation version of the EM algorithm. Ann Stat. 1999;27: 94–128.

20. Chan PLS, Jacqmin P, Lavielle M, McFadyen L, Weatherley B. The use of the SAEM algorithm in MONOLIX software for estimation of population pharmacokinetic-pharmacodynamic-viral dynamics

parameters of maraviroc in asymptomatic HIV subjects. J Pharmacokinet Pharmacodyn. 2011;38: 41–61. doi:10.1007/s10928-010-9175-z

21. Romero-Santacreu L, Moreno J, Pérez-Ortín JE, Alepuz P. Specific and global regulation of mRNA stability during osmotic stress in Saccharomyces cerevisiae. RNA. 2009;15: 1110–20.

22. Snijder B, Pelkmans L. Origins of regulated cell-to-cell variability. Nat Rev Mol cell Biol. 2011;12: 119–25. doi:10.1038/nrm3044

23. Almquist J, Bendrioua L, Adiels CB, Goksör M, Hohmann S, Jirstrand M. A Nonlinear Mixed Effects Approach for Modeling the Cell-To-Cell Variability of Mig1 Dynamics in Yeast. PLoS One. 2015;10: e0124050. doi:10.1371/journal.pone.0124050

24. Karlsson M, Janzén DLI, Durrieu L, Colman-Lerner A, Kjellsson MC, Cedersund G. Nonlinear mixed-effects modelling for single cell estimation: when, why, and how to use it. BMC Syst Biol. BMC Systems Biology; 2015;9: 52. doi:10.1186/s12918-015-0203-x

25. Hilfinger A, Paulsson J. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. Proc Natl Acad Sci U S A. 2011;108: 12167–72. doi:10.1073/pnas.1018832108

26. Neuert G, Munsky B, Tan RZ, Teytelman L, Khammash M, van Oudenaarden A. Systematic identification of signal-activated stochastic gene regulation. Science. 2013;339: 584–7. doi:10.1126/science.1231456

27. Hasenauer J, Waldherr S, Doszczak M, Radde N, Scheurich P, Allgöwer F. Identification of models of heterogeneous cell populations from population snapshot data. BMC Bioinformatics. BioMed Central Ltd; 2011;12: 125. doi:10.1186/1471-2105-12-125

28. Hasenauer J, Hasenauer C, Hucho T, Theis FJ. ODE constrained mixture modelling: a method for unraveling subpopulation structures and dynamics. PLoS Comput Biol. 2014;10: e1003686. doi:10.1371/journal.pcbi.1003686

29. Bonassi F V, You L, West M. Bayesian learning from marginal data in bionetwork models. Stat Appl Genet Mol Biol. 2011;10: 49. doi:10.2202/1544-6115.1684

30. Zechner C, Ruess J, Krenn P, Pelet S, Peter M, Lygeros J, et al. Moment-based inference predicts bimodality in transient gene expression. Proc Natl Acad Sci U S A. 2012;109: 8340–8345. doi:10.1073/pnas.1200161109

31. Colman-Lerner A, Gordon A, Serra E, Chin T, Resnekov O, Endy D, et al. Regulated cell-to-cell variation in a cell-fate decision system. Nature. 2005;437: 699–706. doi:10.1038/nature03998

32. Gonzalez AM, Uhlendorf J, Cinquemani E, Batt G, Ferrari-Trecate G. Identification of biological models from single-cell data: A comparison between mixed-effects and moment-based inference. Proc 12th IEEE Eur Control Conf. IEEE Press; 2013; 3652–3657.

33. Bertaux F, Stoma S, Drasdo D, Batt G. Modeling dynamics of cell-to-cell variability in TRAIL-induced apoptosis explains fractional killing and predicts reversible resistance. PLoS Comput Biol. 2014;10: e1003893. doi:10.1371/journal.pcbi.1003893

34. Toettcher JE, Gong D, Lim WA, Weiner OD. Light-based feedback for controlling intracellular signaling dynamics. Nat Methods. 2011;8: 837–839.

35. Milias-Argeitis A, Summers S, Stewart-Ornstein J, Zuleta I, Pincus D, El-Samad H, et al. In silico feedback for in vivo regulation of a gene expression circuit. Nat Biotechnol. 2011;29: 1114–1116.

36. Menolascina F, Fiore G, Orabona E, De Stefano L, Ferry M, Hasty J, et al. In-vivo real-time control of protein expression from endogenous and synthetic gene networks. PLoS Comput Biol. 2014;10: e1003625. doi:10.1371/journal.pcbi.1003625

# Supplementary Material:

# What Population Reveals about Individual Cell Identity: Single-cell Parameter Estimation of Models of Gene Expression in Yeast

Artémis Llamosi[1,2,†], Andres M. Gonzalez-Vargas[3,†], Cristian Versari[4], Eugenio Cinquemani[5], Giancarlo Ferrari-Trecate[3,‡], Pascal Hersen[2,‡], and Gregory Batt[1,‡]

[1] *INRIA Paris-Rocquencourt, 78153 Le Chesnay, France*

[2] *Laboratoire Matière et Systèmes Complexes, UMR 7057, Université Paris Diderot & CNRS, 10 Rue Alice Domon, 75013, Paris, France.*

[3] *Dipartimento di Ingegneria Industriale e dell'Informazione, Università degli Studi di Pavia, via Ferrata 3, 27100 Pavia, Italy*

[4] *Laboratoire d'Informatique Fondamentale de Lille, UMR 8022, Université de Lille 1 & CNRS, 59655 Villeneuve d'Ascq Cedex, France*

[5] *INRIA Grenoble – Rhône-Alpes, 38330 Montbonnot, France*

Appendix

Contents


- Text S1. Supplementary Methods (p 3)
  - Experimental Design (p3)
  - Data analysis (p 3)
  - Model of osmostress-induced gene expression (p 5)
  - Parameter inference (p 7)
  - Simulation of population behavior (p 10)
  - Correlation with quantitative single cell measurements (p 10)
  - Heritability analysis (p 10)
- Text S2. Robustness of population predictions: influence of the cell number and of the learning time horizon (p 12)
  - Influence of cell number on the robustness of population predictions (p 12)
  - Influence of the learning time horizon on the robustness of population predictions (p 13)
- Text S3. Validation of population predictions: predicting population behavior on two validation datasets (p 14)
- Text S4. Identifiability analysis (p 15)
- Text S5. On learning the statistics of non-identifiable parameters (p 16)
- Figure S1. pSTL1 expression in response to repeated osmotic stresses shows a high level of variability between cells (p 19)
- Figure S2 – Statistical inference methods for single-cell and population parameter estimation (p 20)
- Figure S3. The distribution that better describes the entire population is more compact and more structured (p 21)
- Figure S4. Average parameter distance of Mother-Daughter pairs against random pairs from the same experiment (p 22)
- Table S1. Parameter estimates for the mixed-effects model using the naive inference approach, using SAEM on the identification dataset, and using SAEM on the validation dataset (p 23)

# Text S1. Supplementary Materials and Methods

## Experimental Design

In this study, we aim at characterizing the extrinsic variability in gene expression. When designing experiments, we aimed at *(i)* obtaining relatively high signal-to-noise ratio, *(ii)* obtaining highly-informative single cell traces, and *(iii)* minimizing the effect of intrinsic variability on the level of our fluorescent protein.

To obtain high signal-to-noise ratio, we used the STL1 promoter, one of the strongest osmoresponsive promoter (1), and applied hyperosmotic shocks in a repeated manner, starting experiments with a short series of 7 strong shocks. Because of cell adaptation to hyperosmotic environments, significantly higher expression levels are reached in fluctuating environments, in which stresses are repeatedly applied, than in sustained hyperosmotic environments (2).

Regarding single-cell gene expression dynamics, experiments in which periods dominated by protein production or by protein degradation are both present are in principle more informative than experiments in which only one behavior is observed. We therefore applied after the first seven shocks randomized sequences of shocks, instead of more the more conventional periodic shock profiles.

To minimize the effects of intrinsic variability on the level of our reporter protein, we used a stable fluorescent protein (yECitrine) and non-lethal but relatively strong shocks (1M sorbitol). Firstly, the use of long-lived reporters averages fast fluctuations in time. Secondly, although intrinsic variability is important for mild stresses (eg 0.1M NaCl), this effect was found to be significantly attenuated for more pronounced shocks (3). Note that care must be taken when comparing the two systems since our fluorescent reporter is integrated at the endogenous STL1 locus whereas in (3) it was inserted in the LEU2 and/or HIS3 loci and noise properties are context dependent. In particular, the endogenous locus of STL1 is very close to the telomeres of chromosome IV, a region expected to be more subjected to epigenetic effects, a potential source of extrinsic variability in gene-expression.

Lastly, we note that although the experimental design has to comply with the specific constraints of our biological system, the proposed inference approach is general.

## Data analysis

### Datasets of single-cell gene expression measurements

The cell expression profiles were generated in three experiments: one for identification ($\mathcal{D}^I$) and two for validation ($\mathcal{D}^V$ and $\mathcal{D}^P$). In each experiment, the cells were first subjected to a series of seven hyperosmotic shocks of eight minutes every 30 minutes to obtain fluorescence measurements with good signal-to-noise ratio. After this, several osmotic shocks of eight minutes were applied with randomly-selected time intervals between shocks ($\mathcal{D}^I$ and $\mathcal{D}^V$) or with periodic shocks ($\mathcal{D}^P$). Image analysis was performed with a home-made segmentation and tracking tool, called CellStar. Raw data coming from image analysis were processed as follows. First, the data gathered from all imaging chambers were pooled together. Second, a manual review of the images and their analysis with CellStar was carried out to look for tracking problems and other possible sources of error. Missing data during the lifespan of a cell was replaced via linear interpolation when no more than one

sample was missing. Subsequently, we discarded the information for the first two hours of newly detected cells after observing that usually, these cells correspond to buds that remain attached to their mother a long time after their detection. Also because of their very small size, fluorescence quantification artifacts (very dark, very steep increase) are encountered. Finally, only cells whose lifespan extended for more than 5 hours were selected for identification and validation. The motivation for this being that cells whose lifespan is too short may not contain enough information on the dynamics of the system and may generate unreliable parameter estimates.

The datasets $\mathcal{D}^I$, $\mathcal{D}^V$, and $\mathcal{D}^P$ contain 325, 166, and 285 single-cell trajectories, respectively (see Figure S1). Among them, 63 and 39 trajectories start from time zero and remain in the device during all the experiment in $\mathcal{D}^I$ and $\mathcal{D}^V$, respectively. They are called "initial cells" and will be denoted with $\mathcal{D}^{I0}$ and $\mathcal{D}^{V0}$. Unless otherwise noted, all results are given for the identification dataset $\mathcal{D}^I$.

## Estimation of single cell quantitative features

**Average perceived shock intensity.** Hyperosmotic shocks cause reductions of the cellular volume. The cell volume returns approximatively to its pre-shock value upon restoration of the normal growth conditions. For each cell and each shock, the perceived shock intensity is defined as the relative change in volume, with the minimal and maximal volumes estimated over an 18 min time window around the considered shock. The duration of the time window was set so that it includes data before, during and after the shock. Volumes are estimated from apparent sizes in $px^2$ under the assumption that cells are spherical. Then, for each cell, the average perceived shock intensity is defined as the perceived shock intensity averaged over all shocks.

**Average cell size.** The size of each cell is computed at each time instant in bright-field images and then averaged over all time instants. Because hyperosmotic shocks lead to marked reductions of the cell volume and steep changes in the cell fluorescence, images taken less than 12 minutes after a shock are removed for this analysis.

**Average age**. The time of birth of a cell is defined as the time of its detection by the image analysis tool. It is to be noted that the cell may not yet have detached from its mother at that time. The cell's mean age is simply the average of the cell's age at every time frame. This feature cannot be defined for cells present at the beginning of the experiment (initial cells).

**Average density.** The density of the environment of a single cell is defined as the area occupied by neighbor cells relative to the area of the cell's neighborhood. The neighborhood is defined as a disk of radius of 75 px, corresponding approximatively to 5 typical cell's radii. If the centroid of another cell is inside this neighborhood, then it is considered a neighbor cell. The mean density is the averaged cell's density over all time frames.

**Average division rate.** In order to automatically estimate single cell division times from bright-field images, we use the relative changes in size of cells (buds volume are not accounted for). As visible in Figure 1, these relate with budding. Since osmotic shocks greatly impact the size of cells (Figure 1), the first 12 min following shocks are discarded, yielding the dark blue curve in Figure 1 A. Two signals are defined: the first one, $s_2(t)$ (green), is obtained by smoothing twice the size curve with an 11-frame (33 min) window average; the second one, the general tendency $s_{10}(t)$ (yellow), is obtained by iteratively smoothing 10 times with an 11-frame window.

We then define the relative cell size as the relative difference between the smoothed size and the general tendency $\frac{s_2(t)-s_{10}(t)}{s_{10}(t)}$ (Figure 1 B) and compute the Fourier power spectrum of this signal. The average division time is defined as the power averaged frequency for frequencies having a period in between 60 and 400 min, conservatively including possible doubling times for yeast. This approach has been manually validated on fifty cells yielding an average error (compared to manual bud appearance based doubling rate) in the mean doubling rate of 12%.

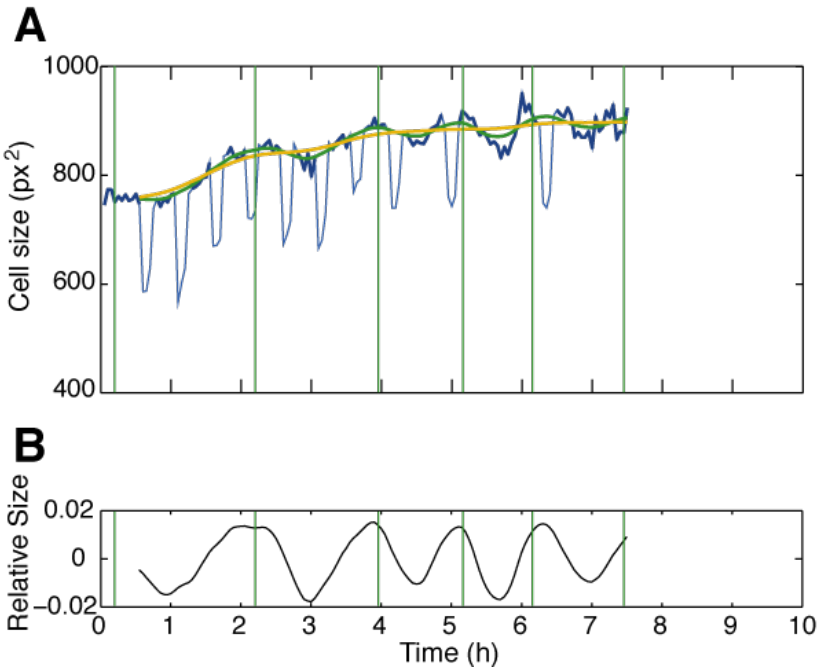



**Figure 1. Automated detection of cell budding times.** A. Plots representing the temporal evolution of the size of a cell together with several signals used for automatically extracting doubling rate. Here, the cell #6 in $\mathcal{D}^I$ was used. Light blue is the raw size. Dark blue is the raw size without shocks. Green is $s_2(t)$, the smoothed version of the size without shocks. Yellow is $s_{10}(t)$, the general tendency. Vertical green bars show manually detected bud appearance time (shown for validation purpose). B. Normalized cell size used for the Fourier analysis. Vertical green bar are the same as in A.

### Cell lineage reconstruction

After automated segmentation and tracking, lineage was manually extracted from microscopy images for the first imaging chamber of the identification experiment $\mathcal{D}^I$ in order to retrieve the complete lineage tree for the cells tracked in that chamber. Among the 86 cells in the chambers, 55 mother-daughter relationships are identified (the experiment starts with 26 cells in the fields of view and 5 additional cells come from the outside).

## Model of osmostress-induced gene expression

### Gene expression model

We use here the following model of gene expression:

$$\begin{cases} \dot{m}(t) = k_m u(t) - g_m m(t), \\ \dot{p}(t) = k_p m(t) - g_p p(t), \end{cases}$$

where $m$ and $p$ denote, respectively, the cellular concentration of the mRNA and of the fluorescent protein yECitrine. Synthesis and degradation rates for mRNA are represented by $k_m$ and $g_m$, whereas their respective counterparts for the protein are denoted with $k_p$ and $g_p$. At time zero, we consider the initial concentrations $m_0 = p_0 = 0$. The input function $u(t)$ represents the phosphorylation and nuclear import of the Hog1 protein, and like in Uhlendorf *et al* (2012) and Muzzey *et al* (2009), we assume that it depends on the osmolarity effectively sensed by the cells inside the microfluidic chambers $u_c(t)$ as follows:

$$\dot{u}(t) = k_h u_c(t) - g_h u(t).$$

In accordance with the observations made in Zechner *et al* (2012), we assume that in comparison to gene expression, signal transduction shows little variability. Therefore, we assume fixed values for $k_h$ and $g_h$: $k_h = 0.3968$ and $g_h = 0.9225$ (6). Lastly, as shown in Uhlendorf *et al* (2012), there is a known lag between the valve actuation $u_v(t)$ and the actual change in the osmolarity of the cellular environment in the imaging chamber $u_c(t)$. This relation can be simply represented by a piecewise linear function. The relations between the valve status, the chamber osmolarity, and the Hog1 activity are graphically represented in Figure 2 for an 8-minute shock.

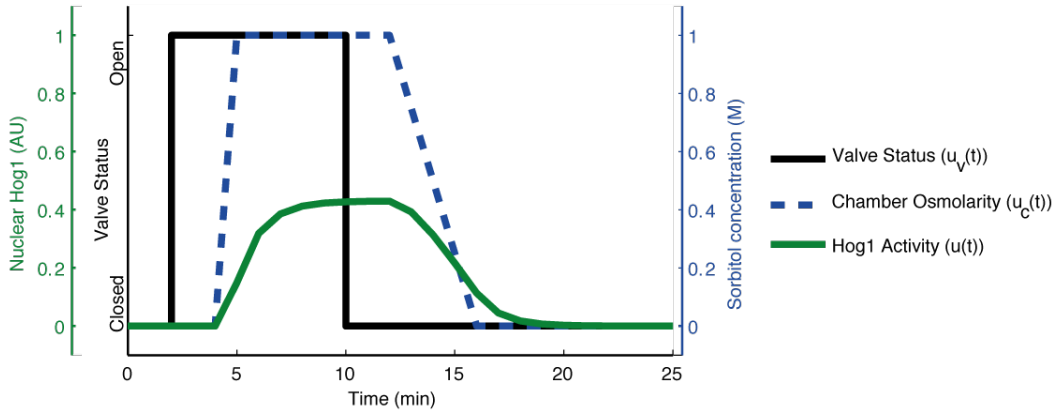



**Figure 2: Temporal evolution of the osmolarity of the cellular environment $u_c$, and of the Hog1 activity $u$, as a function of the position of the microfluidic valve $u_v$ (0/1: normal/hyper-osmotic medium).**

In order to account for fluorescent protein maturation time, we introduce a delay $\tau$. The measured concentration of mature protein depends on the total (mature or not) protein concentration $\tau$ instants before and on the dilution rate due to cell (exponential) growth (we neglect degradation and photobleaching, see section Initial parameter values). To establish these relations, consider a cell that grows at a rate $g_p$. Then, $p(t)$ is the total protein concentration at time $t$, and we denote $P(t)$, $F(t)$, $V(t)$, and $f(t)$, the total and mature protein amounts, the cell volume, and the mature protein concentration, respectively. Then it holds that $V(t) = V(t-\tau)e^{g_p\tau}$ and that $F(t) = P(t-\tau)$. So, we can describe the cell fluorescence as

$$f(t) = \frac{F(t)}{V(t)} = \frac{P(t-\tau)}{V(t-\tau)e^{g_p\tau}} = e^{-g_p\tau}p(t-\tau).$$

We assume $p(t) = 0$ for $t \leq 0$. The choice of representing protein maturation by a delay rather than with first or second order linear reaction comes from the observation that in our data there is a delay following a shock during which absolutely no increase in fluorescence is observed. Finally, we assume a Gaussian noise model for the observations noise, with an additive component and a multiplicative component, meaning that the measured cell fluorescence $y(t)$ follows

$$y(t) = f(t) + h(t)\eta(t),$$

with $h(t) = (\varepsilon_a + f(t)\varepsilon_b)$, and where $\eta(t)$ is white Gaussian noise with mean 0 and intensity 1, and ${\varepsilon_a}^2$, ${\varepsilon_b}^2$ define the intensity of the additive and multiplicative noise components. NB: experimental noise being considered iid, this formulation is equivalent to having two independent Gaussian white noises for the additive and multiplicative contributions by additivity of Gaussian random variables. Also in the SAEM inference these noise parameters are the same for all the cells. For the Naïve inference, it is not possible to infer noise parameters shared by all cells. These have either to be fixed or to be estimated separately for each cell. Nevertheless, estimating measurement noise parameters for each cell gives very close noise values to that found with SAEM. We verified that fixing the same measurement noise for the Naïve approach did not change the results.

## Mixed-effects model of gene expression

In this framework, we assume that cells share the structural model of gene expression described above but that their parameters are different. Let denote $S$ the number of molecular species, and $x(t) = [\,x_1(t), \dots, x_S(t)]^T$ the vector of their respective concentrations at time $t$. The velocity of the change in concentration for each species can be described as a differential equation of the form

$$\dot{x}_i(\mathrm{t}) = v(\mathrm{x(t)}, u(t), \psi_i)$$

where $i = 1, \dots, \mathrm{N}$ and N is the total number of cells and $\psi_i$ are cell-specific parameters. System quantities that are measured over time can be described via an output equation:

$$y_i(\mathrm{t}) = \tilde{v}(x_i) + h(\tilde{v}(x_i), \xi)\eta_i(t)$$

where the vector $y_i(\mathrm{t}) \in \mathrm{R^m}$ is the system output. Function $\tilde{v}(\cdot)$ allows us to select state variables that are observed over time (e.g. total concentration of protein). We assume output measurements are corrupted by additive and multiplicative noise, i.e.

$$h(\tilde{v}(x_i), \xi) = (\varepsilon_a + \tilde{v}(x_i)\varepsilon_b)$$

$\eta_i$ represents white Gaussian noise where $\eta_i(t) \sim N(0,1)$. Denoting the output transition map when $\eta_i(t) = 0$ as $f(t, \mathrm{u}, x_0, \psi_i)$, we have:

$$y_i(t) = f(t, u, x_0, \psi_i) + h(t, u, x_0, \psi_i, \xi)\eta_i(t)$$

For a given observation at time $t_j$ :

$$y_{ij} = f\big(t_j, \mathrm{u}, x_0, \psi_i\big) + h\big(t_j, \mathrm{u}, x_0, \psi_i, \xi\big)\eta_{ij}, \qquad i = 1, \dots, N \;\; j = 1, \dots, T$$

This equation is called the individual-level model.

For our particular system, $x_0 = x(t_0) = \{m_0, p_0\}$, $\psi_i = \{k_{mi}, g_{mi}, k_{pi}, g_{pi}, \tau\}$, $\xi = \{\varepsilon_a, \varepsilon_b\}$, and:

$$v(\mathrm{x(t)}, u(t), \psi_i) = \begin{cases} \dot{m}(t) = k_m u(t) - g_m m(t), \\ \dot{p}(t) = k_p m(t) - g_p p(t), \end{cases}$$

$$\tilde{v}(x_i) = e^{-g_p \tau} p(t - \tau)$$

Note that the noise parameters $\xi = \{\varepsilon_a, \varepsilon_b\}$ are constant over the whole population of cells. The single-cell parameters $\psi_i$ depend themselves on a model that defines their statistics. This model is called population-level model and it defined as

$$\psi_i = d(\mu, b_i); \; b_i \sim N(0, \Omega)$$

where $\mu$ is called the vector of fixed-effects and represents the typical parameters of the population, which are not to be confused with the mean-cell parameters. Vector $b_i$ denotes

the random effects, drawn from a multivariate Gaussian distribution with mean 0 and covariance matrix $\Omega$, which determine how different from $\mu$ the individual parameters are. Function $d(\cdot)$ performs a monotonic transformation on the parameters to have non-normal parameter distributions. Here we assumed that the parameters $\psi_i$ are log-normally distributed:

$$\psi_i = d(\mu, b_i) = e^{\mu + b_i}, \text{ with } \varphi_i = \mu + b_i \text{ and } \varphi_i \sim N(\mu, \Omega)$$

where for a vector $\varphi_i = [\varphi_{i_1}, \dots, \varphi_{i_P}]$, we define the element-wise operation $e^{\varphi_i} = [e^{\varphi_{i1}}, \dots, e^{\varphi_{ip}}]$.

The set of parameters to identify is $\theta = \{\mu, \Omega, \xi\}$. Note that the number of parameters to identify scales quadratically with the number of model parameters.

## Parameter inference

### Initial parameter values

Initial parameter values are estimated based on literature data as described in the table below.

| Parameter | Description | Unit | Reference value | Reference value (Log. scale) | Source |
|---|---|---|---|---|---|
| $k_m$ | Transcription rate | $min^{-1}$ | $1.00\ 10^{1}$ | 2.30 | (7) |
| $g_m$ | mRNA degradation rate | $min^{-1}$ | $2.94\ 10^{-1}$ | -1.22 | (7)[1] |
| $k_p$ | Translation rate | $min^{-1}$ | $9.47\ 10^{-1}$ | $-5.4\ 10^{-2}$ | Computations using (8) & (9) [2] |
| $g_p$ | Protein decay rate | $min^{-1}$ | $4.00\ 10^{-3}$ | -5.52 | This study [3] |
| $\tau$ | Protein maturation time | min | $3.00\ 10^{1}$ | 3.40 | This study [4] |

[1] Note that the value used as reference taken from (7) appeared later to be inappropriate because we express here an exogenous mRNA (pSTL1-yECitrine) which differs on the 3' end from that used in (7). A more conservative value would be to take the average mRNA degradation rate in yeast ($1{,}31\ 10^{-2}$) as measured in (10). Indeed, we found an estimated value closer to this average value and similar to that of native STL1 transcripts as measured in (11).

[2] Translation rate was estimated by using the rate of translation initiation and the rate of correct translation using models from (8, 9) calibrated for our specific DNA sequence.

[3] This term comes from three processes: degradation, photo-bleaching and dilution. Degradation was assumed to be negligible since yECitrine is very stable. The dilution rate was computed from the average doubling rates estimated on $\mathcal{D}^I$ and $\mathcal{D}^V$ (linear fit on the logarithmic number of cells). It yields an average dilution rate of $4\ 10^{-3}$ $min^{-1}$. Photo-bleaching was estimated by repeatedly imaging a cell population and extracting a bleaching rate per frame by fitting the decay curve obtained. This yields a bleaching rate of $3.5\ 10^{-4}$ $min^{-1}$.

[4] The value is based on the time required for the cells to respond to the first stimulus. Note that this value was larger for $\mathcal{D}^V$.

### Inference of mixed-effects model: the naive approach

One intuitive approach for estimating single-cell and population parameters in a mixed-effects framework consists in obtaining individual estimates of rate parameters $\psi_i = \left\{k_{m_i}, g_{m_i}, k_{p_i}, g_{p_i}, \tau_i\right\}$ and noise parameters $\xi_i = \left\{\varepsilon_{a_i}, \varepsilon_{b_i}\right\}$ by fitting one cell at a time and then to compute the population statistics directly from the set of obtained parameters. For each cell $i = 1, \dots, N$, the total set of parameters $\theta_i = \{\psi_i, \xi_i\}$ has been inferred using maximum likelihood estimation (MLE), yielding a total of $7 \times N$ inferred parameters. Given that the estimation of the single-cell parameters, $\theta_i$, is done cell by cell, one has to solve $N$ optimization problems, each involving 7 variables only. This is done as follows:

$$\mathcal{L}(\theta_i|Y_i) = p(Y_i|\theta_i)$$
$$Y_{ij} = f(t_j, \mathrm{u}, x_0, \psi_i) + \varepsilon_j \qquad \varepsilon_j \sim N\left(0, \left(h(t_j, \mathrm{u}, x_0, \psi_i, \xi_i)\right)^2\right)$$

So, for a given sample in cell $i$ at time $j$:

$$p(Y_{i,j}|\theta_i) = \frac{1}{\sqrt{2\pi}\cdot h(f(t_j, \mathrm{u}, x_0, \psi_i), \xi_i)} \exp\left(-\frac{1}{2}\left(\frac{Y_{ij} - f(t_j, \mathrm{u}, x_0, \psi_i)}{h(f(t_j, \mathrm{u}, x_0, \psi_i), \xi_i)}\right)^2\right)$$

$$\log[p(Y_{i,j}|\theta_i)] = \log\frac{1}{\sqrt{2\pi}} - \log\left(h(f(t_j, \mathrm{u}, x_0, \psi_i), \xi_i)\right) - \frac{1}{2}\left(\frac{Y_{ij} - f(t_j, \mathrm{u}, x_0, \psi_i)}{h(f(t_j, \mathrm{u}, x_0, \psi_i), \xi_i)}\right)^2$$

And, for the complete set of samples in cell $i$:

$$p(Y_i|\theta_i) = p(Y_{i,1}, Y_{i,2}, \dots, Y_{i,j}|\theta_i) = \prod_j p(Y_{i,j}|\theta_i)$$

Therefore:

$$\log[p(Y_i|\theta_i)] = -\sum_j \frac{1}{2}\left(\frac{Y_{ij} - f(t_j, \mathrm{u}, x_0, \psi_i)}{h(f(t_j, \mathrm{u}, x_0, \psi_i), \xi_i)}\right)^2 - \sum_j \log\left(h(f(t_j, \mathrm{u}, x_0, \psi_i), \xi_i)\right) + \sum_j \log\frac{1}{\sqrt{2\pi}}$$

The last term is a constant, thus can be removed from the equation. Now we can compute:

$$\hat{\theta}_{i_{ML}} = \underset{\psi_i, \xi_i}{\mathrm{Argmax}} \left[-\frac{1}{2}\sum_j \left(\frac{Y_{ij} - f(t_j, \mathrm{u}, x_0, \psi_i)}{h(f(t_j, \mathrm{u}, x_0, \psi_i), \xi_i)}\right)^2 - \sum_j \log\left(h(f(t_j, \mathrm{u}, x_0, \psi_i), \xi_i)\right)\right]$$

The maximization is performed using the *fminsearch* function from Matlab. For convenience, all the parameters are estimated in the logarithmic scale, which is equal to estimating the normally-distributed variables $\varphi_i$ as defined in the mixed-effects model section. Denoting $\varphi = \{\varphi_1, \varphi_2, \dots \varphi_N\}^T$, the population statistics $\mu, \Omega$ and $\xi$ are computed in the following way.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \varphi_i\,, \qquad \Omega = \frac{1}{N-1}\sum_{i=1}^{N} (\varphi_i - \mu)(\varphi_i - \mu)^T\,, \qquad \xi = \frac{1}{N}\sum_{i=1}^{N} \xi_i$$

**Inference of mixed-effects model: the SAEM approach**

Instead of the naive approach, one can choose a more inclusive approach, in which we directly estimate the population statistics by accounting simultaneously for the ensemble of all cell's observations. Single-cell fits can be computed a posteriori using the population distribution as prior knowledge (see below). The estimation of the population statistics can be done via population-likelihood maximization algorithms, such as the Stochastic Approximation Expectation Maximization (SAEM) (12). Here the set of parameters to identify is $\theta = \{\mu, \Omega, \xi\}$. At each iteration of the algorithm, the objective is to maximize the log-likelihood [Lixoft, Monolix Methodology v.4.3.3, 2014]

$$\log p(Y, \psi\,|\theta) = -\sum_{i,j} \log\left(h(x_{ij}, \psi_i, \xi)\right) - \frac{1}{2}\sum_{i,j}\left(\frac{y_{ij} - f(x_{ij}, \psi_i)}{h(x_{ij}, \psi_i, \xi)}\right)^2 - \frac{N}{2}\log(|\Omega|)$$
$$-\frac{1}{2}\sum_{i=1}^{N} (\varphi_i - \mu)'\Omega^{-1}(\varphi_i - \mu) - \frac{N_{tot} + N_d}{2}\log(2\pi),$$

where $N_{tot}$ is the total number of samples and $N_d$ is the number of degrees of freedom. We used the SAEM implementation available in the Monolix software (13).

**Inference of single-cell models from population models: a MAP approach**

Based on a population distribution with parameters $\theta = \{\mu, \Omega, \xi\}$, single cell estimates $\psi_i$ are obtained via maximum *a posteriori* estimation (MAP).

$$\hat{\psi}_{i_{MAP}} = \underset{\psi_i}{\text{Argmax}}[p(\psi_i|Y_i, \mu, \Omega, \xi)] = \underset{\psi_i}{\text{Argmax}}\left[\frac{p(Y_i|\psi_i, \xi) \cdot p(\psi_i|\mu, \Omega)}{p(Y_i|\mu, \Omega, \xi)}\right]$$

Note that, because of identifiability issues, $\psi_i$ might be reduced to 4 effective parameters only, where $k_{mp}$ captures the product of $k_m$ and $k_p$ (see main text and Text S4). By a slight abuse of notation, and because it should be clear from the context, we use $\psi_i$ to denote both vectors. As $p(Y_i|\mu, \Omega, \xi)$ does not depend on $\psi_i$, it can be removed from the equation. Then:

$$\hat{\psi}_{i_{MAP}} = \underset{\psi_i}{\text{Argmax}}\left[\log[p(Y_i|\psi_i, \xi)] + \log[p(\psi_i|\mu, \Omega)]\right]$$

The first term, $\mathbf{log}[\boldsymbol{p}(\boldsymbol{Y_i}|\boldsymbol{\psi_i}, \boldsymbol{\xi})]$, corresponds to the log-likelihood $\mathbf{log}\left[\boldsymbol{p}\left(\boldsymbol{Y_{i,j}}|\boldsymbol{\theta_i}\right)\right]$ explained previously summed over $\boldsymbol{j}$ (the only difference being that in the present case the noise parameters $\boldsymbol{\xi}$ are common to all cells. For the second term, $\mathbf{log}[\boldsymbol{p}(\boldsymbol{\psi_i}|\boldsymbol{\mu}, \boldsymbol{\Omega})]$, we have:

$$p(\psi_i|\mu, \Omega) = \frac{1}{\sqrt{2\pi^n|\Omega|}} \exp\left(-\frac{1}{2}(\psi_i - \mu)^T \Omega^{-1}(\psi_i - \mu)\right)$$

$$\log[p(\psi_i|\mu, \Omega)] = \log\left(\frac{1}{\sqrt{2\pi^n|\Omega|}}\right) - \frac{1}{2}(\psi_i - \mu)^T \Omega^{-1}(\psi_i - \mu)$$

The term $\log\left(\frac{1}{\sqrt{2\pi^n|\Omega|}}\right)$ being a constant, we obtain:

$$\hat{\boldsymbol{\psi}}_{\boldsymbol{i_{MAP}}} = \underset{\boldsymbol{\psi_i}}{\mathbf{Argmax}}\left[-\frac{\mathbf{1}}{\mathbf{2}}\sum_j \left(\frac{\boldsymbol{Y_{ij}} - \boldsymbol{f}(\boldsymbol{t_j}, \mathbf{u}, \boldsymbol{x_0}, \boldsymbol{\psi_i})}{\boldsymbol{h}(\boldsymbol{f}(\boldsymbol{t_j}, \mathbf{u}, \boldsymbol{x_0}, \boldsymbol{\psi_i}), \boldsymbol{\xi})}\right)^2 - \sum_j \mathbf{log}\left(\boldsymbol{h}(\boldsymbol{f}(\boldsymbol{t_j}, \mathbf{u}, \boldsymbol{x_0}, \boldsymbol{\psi_i}), \boldsymbol{\xi})\right) - \frac{\mathbf{1}}{\mathbf{2}}(\boldsymbol{\psi_i} - \boldsymbol{\mu})^T \boldsymbol{\Omega^{-1}}(\boldsymbol{\psi_i} - \boldsymbol{\mu})\right]$$

## Simulation of population behavior

Predictions of the behavior of cell populations using mixed-effects models are obtained by sampling 10000 parameter values from the distributions and performing the corresponding numerical simulations.

**Correlation with quantitative single cell measurements**

To compute correlations between single cell features and estimated parameters, we used the rank-based Spearman coefficient of variation. The standard two-tailed statistical test included in the Matlab statistics toolbox is used to test the significance of the correlations (p-values). Note that a few cells have been discarded because a given feature cannot always be determined for those cells. For instance, cells present at the beginning of the experiment cannot be assigned an age.

## Heritability analysis

We wanted to test whether single cell parameter values could capture some form to inheritance. We considered the average mother-daughter distance in parameter values using the Euclidian distance $d(q_i, q_j) = \sqrt{(q_j - q_i)^2}$. This distance, computed parameter by parameter, was then compared to the distribution of parameter distances for random pairs taken from the whole experiment population, and showed significant differences (Figure S4). Nevertheless, this comparison is subjected to several biases, mainly that mother and daughter parameters could be closer only because they share a more similar environment. To compensate for these biases, the average mother-daughter parameter value was compared to that of a more thoughtfully constructed control population. First, we only considered cells coming from the same chamber. Second, from this original set of mother-daughter pairs, we constructed a control set of pairs of cells where every pair is made of one mother cell and one daughter cell of a different mother. This allows comparing two set of pairs of cells (related mothers and daughters, termed MD and non-related mothers and daughters, termed nMD) which are made from exactly the same cells and only differ by the presence of direct lineage relationship, therefore minimizing the previously mentioned biases. The MD set is made of 55 pairs and the nMD set is made of 1870 pairs. We see that mean distances based on MD pairs are always smaller than mean distances based on nMD pairs. Nevertheless, we wanted to derive a p-value on this hypothesis. Because the distribution of distances between pairs of parameters is a priori of unknown shape (and in practice non-gaussian), bootstrapping, a standard method to derive confidence intervals for random variables with unknown distribution was used. *p*-values are based on the Welch two sided t-test for checking whether the mean of 40 cells in MD is smaller than the mean of 40 cells in nMD. The effective degree of freedom was computed using the Welch-Satterthwaite equation. In the main text example, we used 50 000 bootstrapped sets of 40 pairs. Similar results were obtained when we used a different number of bootstrapped sets, and smaller and bigger sets (*e.g.* bootstrapping sets of 5, 10, 20, 30, 40 and 50 cells).

## Bibliography

1. O'Rourke SM, Herskowitz I (2004) Unique and redundant roles for HOG MAPK pathway components as revealed by whole-genome expression analysis. *Mol Biol Cell* 15(February):532–542.

2. Uhlendorf J, et al. (2012) Long-term model predictive control of gene expression at the population and single-cell levels. *Proc Natl Acad Sci U S A* 109(35):14271–14276.

3. Pelet S, et al. (2011) Transient Activation of the HOG MAPK Pathway Regulates Bimodal Gene Expression. *Science* 332(6030):732–735.

4. Muzzey D, Gómez-Uribe C a, Mettetal JT, van Oudenaarden A (2009) A systems-level analysis of perfect adaptation in yeast osmoregulation. *Cell* 138(1):160–71.

5. Zechner C, et al. (2012) Moment-based inference predicts bimodality in transient gene expression. *Proc Natl Acad Sci U S A* 109(21):8340–45.

6. Uhlendorf J, Bottani S, Fages F, Hersen P, Batt G (2011) Towards real-time control of gene expression: controlling the HOG signaling cascade. *Pacific Symposium on Biocomputing*, pp 338–49.

7. Neuert G, et al. (2013) Systematic identification of signal-activated stochastic gene regulation. *Science* 339(6119):584–7.

8. Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB (2013) Rate-limiting steps in yeast protein translation. *Cell* 153(7):1589–601.

9. Gilchrist M a., Wagner A (2006) A model of protein translation including codon bias, nonsense errors, and ribosome recycling. *J Theor Biol* 239:417–434.

10. Wang Y, et al. (2002) Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci U S A* 99(9):5860–5865.

11. Romero-Santacreu L, Moreno J, Pérez-Ortín JE, Alepuz P (2009) Specific and global regulation of mRNA stability during osmotic stress in Saccharomyces cerevisiae. *RNA* 15(6):1110–20.

12. Kuhn E, Lavielle M (2005) Maximum likelihood estimation in nonlinear mixed effects models. *Comput Stat Data Anal* 49(4):1020–1038.

13. Chan PLS, Jacqmin P, Lavielle M, McFadyen L, Weatherley B (2011) The use of the SAEM algorithm in MONOLIX software for estimation of population pharmacokinetic-pharmacodynamic-viral dynamics parameters of maraviroc in asymptomatic HIV subjects. *J Pharmacokinet Pharmacodyn* 38(1):41–61.

## Text S2. Robustness of population predictions: influence of the cell number and of the learning time horizon

### Influence of cell number on the robustness of population predictions

What is the minimum number of cells we have to track in order to obtain reliable estimates? This is an important question to address when dealing with the identification of parameter distributions. We therefore tested the robustness of SAEM inference with respect to the number of cells available in the identification dataset for ME models. For this purpose we repeatedly estimated ME models from datasets containing only a few cells and quantified the variability of the corresponding predictions. More precisely, we extracted from $\mathcal{D}^{I0}$ 25 sub-sets $\mathcal{D}^{I0}_{c,n}$ with $c = 2,4,8,16,32$ and $n = 1,..,5$. Each subset has a number $c$ of cells which will be extracted randomly $n$ times from dataset $\mathcal{D}^{I0}$. We performed population parameter inference in each subset, and compared the quantiles of the predicted population to those of the observed population. The selected quantiles were q=0.5, q=0.025 and q=0.975. These values represent, respectively, the median of the population and the lower and higher bound of the 95% of the population. For the comparison, we used the root mean squared error, normalized by the difference between the maximal and minimal observed values (NRMSE). Table 1 shows the computed NRMSE values for each test and each quantile. The means and standard deviations of the NRMSE in the different tests are indicated in Table 1 and graphically represented in Figure 1. They give a measure of the accuracy and the uncertainty of the estimates. The uncertainty is large when there are only two cells (the quantile's predictions even overlap), but rapidly decreases and stabilizes above 16 cells.

**Table 1. Effect of the number of cell traces on the robustness of the predictions.** The deviation (NRMSE) between the predicted quantiles and the observed quantiles for 5 random subsets of cells is reported. The mean is an indicator of the accuracy of the prediction. The standard deviation is an indicator of the dispersion of these predictions; a low SD indicates that the predictions do not vary considerably when selecting different subsets with the given number of cells.

| Robustness of population predictions with respect to number of cells in $D^I$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | NRMSE(q) | | | | | | |
| # of Cells | Quantile | Test 1 | Test 2 | Test 3 | Test 4 | Test 5 | Mean | SD |
| 2 | q0.025 | 0.14 | 0.37 | 0.33 | 0.32 | 0.14 | **0.26** | **0.11** |
| | q0.5 | 0.11 | 0.13 | 0.22 | 0.08 | 0.07 | **0.12** | **0.06** |
| | q0.975 | 0.08 | 0.05 | 0.16 | 0.07 | 0.10 | **0.09** | **0.04** |
| 4 | q0.025 | 0.15 | 0.20 | 0.11 | 0.06 | 0.13 | **0.13** | **0.05** |
| | q0.5 | 0.06 | 0.06 | 0.06 | 0.08 | 0.07 | **0.07** | **0.01** |
| | q0.975 | 0.08 | 0.08 | 0.06 | 0.13 | 0.07 | **0.08** | **0.03** |
| 8 | q0.025 | 0.09 | 0.11 | 0.05 | 0.11 | 0.09 | **0.09** | **0.02** |
| | q0.5 | 0.06 | 0.06 | 0.07 | 0.06 | 0.05 | **0.06** | **0.01** |
| | q0.975 | 0.12 | 0.05 | 0.12 | 0.07 | 0.10 | **0.09** | **0.03** |
| 16 | q0.025 | 0.05 | 0.06 | 0.14 | 0.07 | 0.07 | **0.08** | **0.04** |
| | q0.5 | 0.06 | 0.06 | 0.05 | 0.06 | 0.07 | **0.06** | **0.01** |
| | q0.975 | 0.10 | 0.07 | 0.08 | 0.09 | 0.09 | **0.09** | **0.01** |
| 32 | q0.025 | 0.06 | 0.06 | 0.05 | 0.06 | 0.07 | **0.06** | **0.01** |
| | q0.5 | 0.06 | 0.05 | 0.06 | 0.06 | 0.06 | **0.06** | **0.00** |
| | q0.975 | 0.10 | 0.06 | 0.09 | 0.12 | 0.08 | **0.09** | **0.02** |

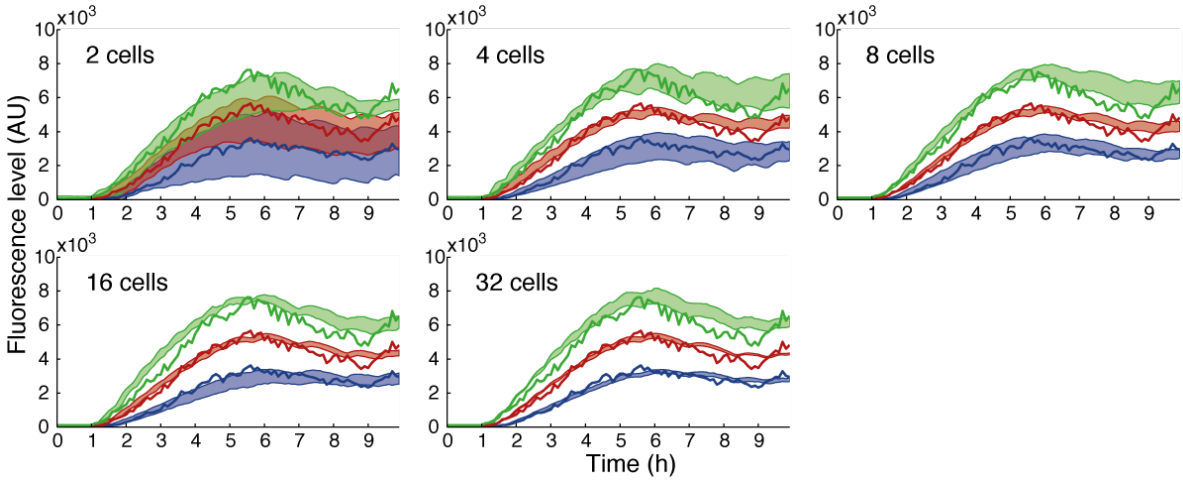



**Figure 1. Effect of the number of cell traces on the robustness of the predictions.** Green, red and blue solid lines denote, respectively the 0.025, 0.5 and 0.975 quantiles, corresponding to 95% of the observed population in $\boldsymbol{D^I}$ (325 cells). The shaded areas with the corresponding colors represent the maximum and minimum boundaries of the quantiles estimated with 5 randomly selected subsets of 2,4,8,16 and 32 cells. Thinner shaded areas indicate less variability in the predictions. After 16 cells the width of these areas has decreased considerably.

## Influence of the learning time horizon on the robustness of population predictions

We tested the robustness of SAEM inference with respect to the duration of the learning time horizon (observation time $T_{obs}$) by testing the prediction capabilities of the resulting mixed-effect models on the rest of the data (prediction time $T_{pred}$). We used the identification dataset $D^I$. ME models inferred on datasets with 5 or 6 hours of observations show bad prediction capabilities on the subsequent hours. After 7 hours the performance increases significantly (Figure 2). This suggests that an accurate inference of the model's parameter values in this experimental setup requires acquisition of data during extended time intervals.

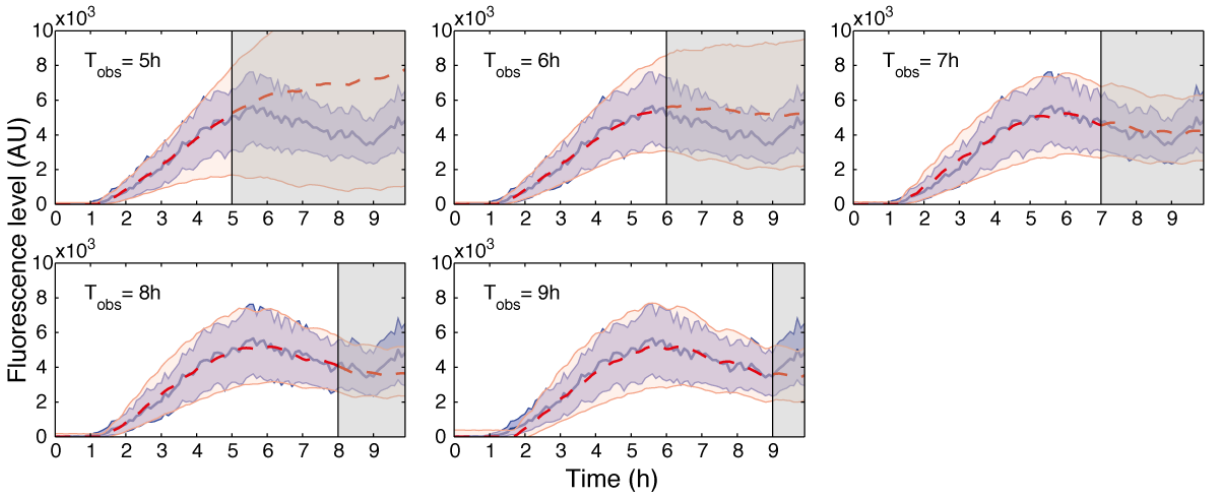



**Figure 2. Influence of the learning time horizon in population predictions.** The blue line and blue shaded area represent observed cell populations (median and 95% of the population). Red dashed line and pink shaded area represent the model predictions during observation time ($T_{obs}$) and prediction time ($T_{pred}$, gray shaded area) (again median and 95% of the population). All are given for experiment $\boldsymbol{D^I}$.

## Text S3. Validation of population predictions: predicting population behavior on two validation data sets

To test the capacity of our model and inferred parameter distribution to predict the behavior of cell populations, we generated two validation datasets. The first one, $\mathcal{D}^V$, uses temporal profile of hyperosmotic shock that is different from but close to the identification dataset $\mathcal{D}^I$. The second one, $\mathcal{D}^P$, uses periodic shocks (8 minutes shocks every half-hour) and is markedly different from the identification dataset.

In both cases, we simulated 1000 single-cell traces using the population distribution of parameters estimated on $\mathcal{D}^I$. Prediction results are represented on Figure 1. On $\mathcal{D}^V$, the prediction quality was acceptable (Fig 1A). However, a significant bias was observed for the validation dataset $\mathcal{D}^P$ (Fig 1B – see the difference between real and predicted median profiles). How can this be explained? One of the results of our study is that, in addition to hyperosmotic shocks, several factors are likely to influence gene expression, one of them being the cell division rate. And indeed, cells grow and divide in $\mathcal{D}^P$ significantly slower than in $\mathcal{D}^I$ (mean division rates are 3.5 $10^{-3}$ and 6.3 $10^{-3}$ $min^{-1}$, respectively). This growth rate difference can be explained by the significantly higher total amount of stress imposed to cells in $\mathcal{D}^P$.

Because protein degradation rate and photobleaching can be neglected in comparison to the dilution effect due to growth (see section Initial parameter values), one can correct the parameter $g_p$ in a systematic manner. In fact, when replacing the median value of $g_p$ from the population distribution of parameters estimated on $\mathcal{D}^I$ with the empirical population division rate, the prediction capability improves and the systematic bias is effectively corrected (Fig 1C). Yet, we observe that the predicted variability is still somewhat lower than the observed one. This might likely come from other differences in environmental or physiological conditions. Unlike the possible correction of $g_p$ which used a direct relationship between a measurable influence factor (division rate) and a parameter of our model, most other influence factors cannot be similarly corrected for prediction purposes. This is because either they cannot be measured, or because applying a correction would require a specific model of the relationship between these factors and single cell parameter values.

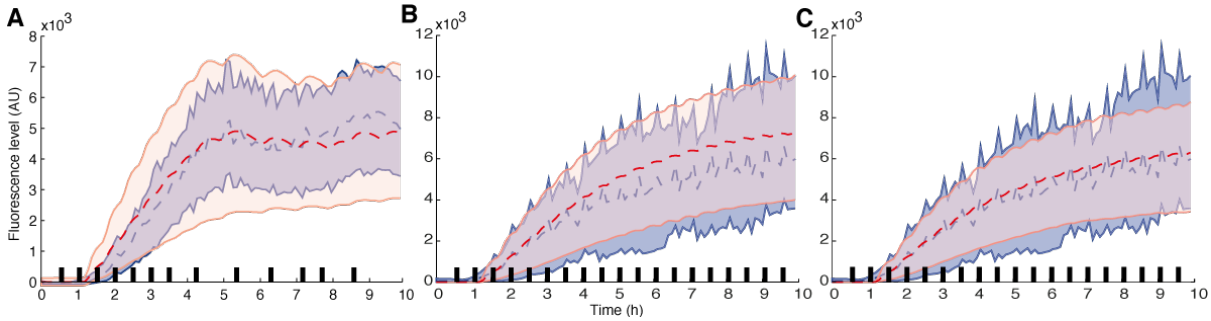



**Figure 1. Predicting population behavior on validation data sets** (A) Prediction for experiment $\mathcal{D}^V$ using a population model estimated on $\mathcal{D}^I$. (B) Prediction for experiment $\mathcal{D}^P$ using a population model estimated on $\mathcal{D}^I$. (C) Prediction for experiment $\mathcal{D}^P$ using a population model estimated on $\mathcal{D}^I$ for which the average dilution rate $g_p$ was set to the population division rate measured in $\mathcal{D}^P$. Shaded areas represent the fluorescence values of 95% of the population and the dashed lines represent the median. Experimental data is represented in blue and simulations of 1000 virtual cells are shown in pink. Black bars indicate the presence of osmotic shocks. Note the different scale for the y-axis.

## Text S4. Identifiability analysis

We show here that parameters $k_m$ and $k_p$ cannot be assigned values unambiguously no matter the quality and quantity of fluorescence measurements. Since the gene expression model giving mRNA ($m$) and protein ($p$) levels as a function of time is a linear dynamical system, and the cell fluorescence $f$ is assumed to be simply a rescaled and delayed version of $p$, we can easily verify this by looking at the transfer function of the system (1). This is given by

$$F(s) = e^{-\tau(g_p+s)}P(s),$$

$$P(s) = \frac{k_m\ k_p}{(s\ +\ g_m)(s\ +\ g_p)}U(s) + \frac{k_p}{(s\ +\ g_m)(s\ +\ g_p)}m(0) + \frac{1}{s\ +\ g_p}p(0),$$

where $U(s)$, $P(s)$ and $F(s)$ are the Laplace transforms of $u(t)$, $p(t)$ and $f(t)$, in the same order. For any fixed input $U(s)$, since $m(0) = p(0) = 0$, it is apparent that $F(s)$ depends on $k_m$ and $k_p$ only via their product. That is, all models with the same values of $g_m$, $g_p$ and $k_m \cdot k_p$ will respond identically to the same input no matter the specific values of $k_m$ and $k_p$. A similar issue would arise if $m(0)$ was different from zero but unknown. In this case, the term depending on $k_p$ only would actually depend on the product $k_p m(0)$, with both factors unknown. This issue is commonly referred to as "structural non-identifiability'' (of $k_m$ and $k_p$).

Structural non-identifiability of $k_m$ and $k_p$ generally results in issues in the identification of their population statistics as well. In order to ensure a well-posed mixed-effects identification problem, all identification results reported in this work were obtained with the mean of $k_p$ fixed to a default value. We stress the fact that, although related, single-cell model (non-)identifiability should not be confused with the (non-)identifiability of the parameter statistics in the mixed-effects approach. To what extent, if at all, identifiability of the statistics of non-identifiable single-cell parameters is ameliorated by a population approach (e.g. through their correlation with yet other parameters) is not obvious. While a full theoretical investigation of this issue would go beyond the scope of this paper, this point is illustrated on a simple example, analogous to our case study, in Text S5.

When using Mixed Effects models and SAEM, controlling shrinkage is also useful in order to detect potential identifiability-related issues. We speak of shrinkage when the empirical distribution of single-cell parameters (as estimated by MAP or maximum likelihood) is narrower than the population distribution. Obtaining shrinkage is indeed reminiscent of having single-cell parameters ill-defined (having a flat likelihood). In such a case, single-cell parameter estimates given by MAP will mostly represent the mode of the prior (i.e. the population distribution), resulting in a narrow distribution of single-cell parameter values (2). Subsequently to our non-identifiability analysis, we found that no substantial shrinkage was present. Indeed, computing the $\eta$-shrinkage as in (2) yielded 12%, 0% and 4% for parameters $k_{mp}$, $g_p$ and $g_m$ on $\mathcal{D}^I$, and 6%, 4% and 4% on $\mathcal{D}^V$, respectively.

**Bibliography**

1. Oppenheim AV., Willsky AS, Nawab SH (1996) Signals and Systems. Prentice Hall

2. Savic RM, Karlsson MO (2009) Importance of shrinkage in empirical Bayes estimates for diagnostics: problems and solutions. AAPS Journal, 11(3):558–569

3. Lavielle M (2014) Mixed Effects Models for the Population Approach. Models, Tasks, Methods & Tools. Chapman & Hall / CRC Press

## Text S5. On learning the statistics of non-identifiable parameters

We illustrate here that statistical properties of parameters that are not distinguishable at the single-cell level can nevertheless be constrained in a population approach. To this aim, we consider a simple model, linear in its parameters:

$$\begin{aligned} y_1 &= \theta_1 + \theta_2 \\ y_2 &= \theta_3 \end{aligned}$$

In this model, $\theta_1$ and $\theta_2$ cannot be distinguished from one observation of the outputs, i.e. they would be non-identifiable at the single-cell level. Note that the situation is analogous to the one encountered in our gene expression model considering log-transformed parameters. Indeed $\log(k_m)$ and $\log(k_p)$ are reflected in the model output only via their sum, $\log(k_{mp})$ (see Text S3 Identifiability analysis). Denoting the experimentally observable vector variable $y = [y_1, y_2]^T$, and the vector of parameters $\theta = [\theta_1 \dots \theta_3]^T$, we have

$$y = L\theta, \text{ with } L = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Suppose that, across different cells, $\theta$ is distributed with mean $\mu_\theta = [\mu_{\theta,1} .. \mu_{\theta,3}]^T$ and covariance matrix $\Sigma_\theta = (s_{r,c})_{r,c=1..3}$. Regardless of the distribution, by the linearity of the model one has that

$$\mu_y = L\mu_\theta, \quad \Sigma_y = L\,\Sigma_\theta L^T$$

For the sake of simplicity, we assume that the output statistics $\mu_y = [\mu_{y,1}\ \mu_{y,2}]^T$ and $\Sigma_y = (\sigma_{r,c})_{r,c=1,2}$ are known (while in practice, in the ME model inference framework, they would be estimated from population data).

**The statistics of non-identifiable parameters are constrained by output statistics**

In this section, we explore to what extent the different statistics of $\theta$, $\mu_\theta$ and $\Sigma_\theta$, are constrained by the output statistics $\mu_y$ and $\Sigma_y$.

**First-order moments.** Because $\mu_y = L\mu_\theta$, it holds that $\mu_{y,1} = \mu_{\theta,1} + \mu_{\theta,2}$ and $\mu_{y,2} = \mu_{\theta,3}$. Thus, knowledge of $\mu_y$ allows reconstruction of the mean of $\theta_3$, but the means of $\theta_1$ and $\theta_2$ are indistinguishable. Unless one is fixed, the other cannot be reconstructed.

**Second-order moments.** Because $\Sigma_y = L\,\Sigma_\theta L^T$, it must hold that

$$\sigma_{1,1} = s_{1,1} + 2s_{1,2} + s_{2,2} \quad (3.1)$$

$$\sigma_{1,2} = s_{1,3} + s_{2,3} \quad (3.2)$$

$$\sigma_{2,2} = s_{3,3} \quad (3.3)$$

This fixes $s_{3,3}$ equal to $\sigma_{2,2}$, but the other entries of $\Sigma_\theta$ are underdetermined. In addition, however, covariance matrices are positive semi-definite, i.e. all eigenvalues are (real and) nonnegative. By a known characterization of this class of matrices, this is equivalent to all principal minors (the determinants of all square matrices obtained by extracting the same rows and columns from the given matrix) being nonnegative. For our case study, among other inequalities, this criterion yields $s_{1,1} \geq 0$, $s_{2,2} \geq 0$ (variances are nonnegative) and $s_{1,1} * s_{2,2} - s_{1,2}^2 \geq 0$, which is equivalent to the two inequalities $s_{1,2} \leq \sqrt{s_{1,1}} * \sqrt{s_{2,2}}$ and $s_{1,2} \geq -\sqrt{s_{1,1}} * \sqrt{s_{2,2}}$. Used in conjunction with Eq. (3.1) this yields $\sigma_{1,1} \leq s_{1,1} + 2\sqrt{s_{1,1}} * \sqrt{s_{2,2}} + s_{2,2}$ and $\sigma_{1,1} \geq s_{1,1} - 2\sqrt{s_{1,1}} * \sqrt{s_{2,2}} + s_{2,2}$, or equivalently

$$\left(\sqrt{s_{1,1}} - \sqrt{s_{2,2}}\right)^2 \leq \sigma_{1,1} \leq \left(\sqrt{s_{1,1}} + \sqrt{s_{2,2}}\right)^2 \quad (4)$$

Thus, knowledge of $\sigma_{1,1}$ (output statistics) implies constraints on the variances of parameters $\theta_1$ and $\theta_2$. Similar constraints involving other entries of $\Sigma_\theta$ can be derived algebraically using (3.1)-(3.3) in conjunction with other implications of the positive-semidefiniteness of $\Sigma_\theta$.

The resulting constraints can be graphically represented for particular values of $\Sigma_y$. Assuming for example that $\sigma_{1,1} = 1$, $\sigma_{1,2} = \sigma_{2,1} = -0.2$ and $\sigma_{2,2} = 0.5$, we obtain the plots in Fig. 1. For all pairs of entries of $\Sigma_\theta$, scatter plots illustrate what values of these unknown entries are compatible with the known output statistics $\Sigma_y$ for at least some values of the remaining entries of $\Sigma_\theta$ (i.e. satisfy $\Sigma_y = L\,\Sigma_\theta L^T$ with a positive-semidefinite $\Sigma_\theta$). For example, Inequalities (4) determine the parabolic shape visible in the first-row, fourth-column plot. In Fig 2, a similar plot shows the relationships that must hold among $s_{1,1}$, $s_{2,2}$ and $s_{1,2}$, the second-order moments of the two unidentifiable parameters $\theta_1$ and $\theta_2$. This analysis shows that knowledge (or accurate estimate) of $\Sigma_y$, together with structural properties of covariance matrices, result in significant knowledge about the (yet underdetermined) values of the underlying statistics of interest, i.e. $\Sigma_\theta$.

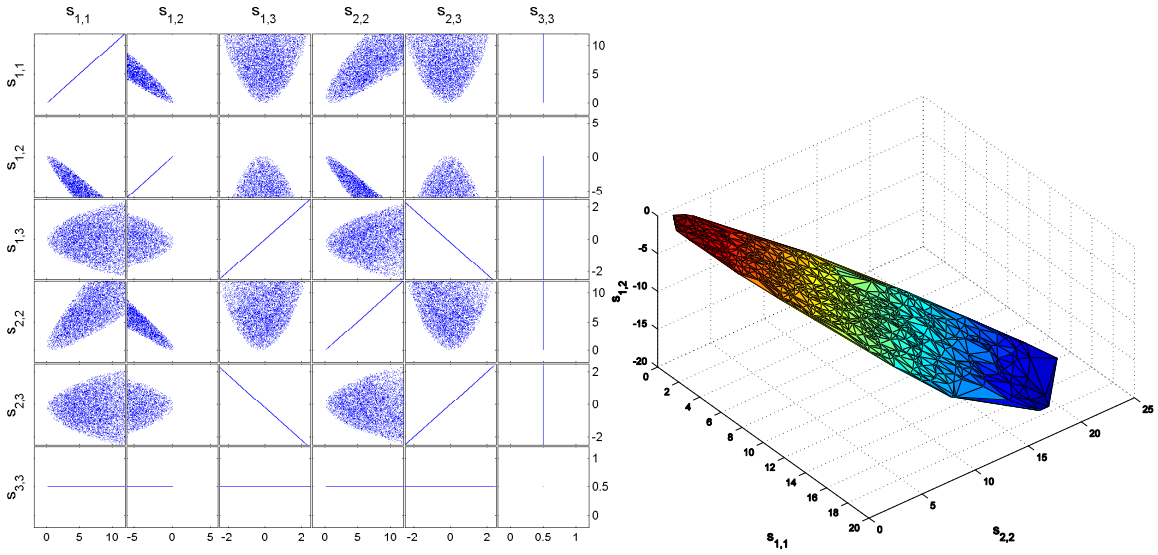



**Figure 1. Second-order output statistics constraint second-order parameter statistics.** (**Left**) Scatter plots of feasible value pairs for the unknown second-order statistics of parameter vector θ. For the given $\boldsymbol{\Sigma_y}$, all possible $\boldsymbol{\Sigma_\theta}$ are computed by first determining the affine space of symmetric solutions of the linear equation $\boldsymbol{\Sigma_y = L\Sigma_\theta L^T}$. Then, $\boldsymbol{10^6}$ candidate $\boldsymbol{\Sigma_\theta}$ are generated at random from within this space, and only the positive semidefinite solutions (*i.e.* the solutions with nonnegative eigenvalues) are retained and reported in the plot. (**Right**) Surface of feasible value triples for the unknown (joint) second-order statistics of the unknown parameters $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$. Sample solution triplets are obtained by the method described above, and the plotted solution surface is obtained from the samples by triangulation.

### The statistics of non-identifiable parameters are constrained by correlations between identifiable and non-identifiable parameters

We now pose the question how correlation between an identifiable parameter and a non-identifiable one may help the estimation of the latter. For simplicity let $\mu_\theta = 0$ (arguments below can be generalized to $\mu_\theta \neq 0$). Consider again the case where the identifiable parameter $\theta_3$ is perfectly determined via $y_2$ by the observation of the single-cell output $y$. Regardless of the additional information provided by $y_1$, what can we say about, *e.g.*, $\theta_1$? From the theory of linear estimation, the optimal linear estimator of $\theta_1$, which is also optimal over all possible estimators in the Gaussian case, is $\theta_1^* = s_{1,3}s_{3,3}^{-1}\theta_3$, and the variance of the estimation error is

$$var(\theta_1^* - \theta_1) = s_{1,1} - s_{1,3}s_{3,3}^{-1}s_{3,1} = s_{1,1} - s_{1,3}^2/s_{3,3}$$

Thus, relative to the a priori variance $s_{1,1}$, observation of $y_2 = \theta_3$ decreases the uncertainty about $\theta_1$ by the amount $s_{1,3}^2/s_{3,3}$, which is positive if the correlation $s_{1,3}$ between $\theta_1$ and $\theta_3$ is nonzero. The residual uncertainty about $\theta_1$ is captured by the so-called Fraction of Unexplained Variance, defined as

$$FUV = 1 - {s_{1,3}^2}\Big/{s_{1,1}s_{3,3}}$$

The larger the correlation between $\theta_1$ and $\theta_3$, the smaller the residual uncertainty about $\theta_1$ given the knowledge of $\theta_3$. Because $s_{1,1}$, $s_{3,3}$ and $s_{1,3}$ are only partially determined by $\Sigma_y$, the FUV cannot be computed uniquely. For the case of the previous section, however, we computed the average FUV over all sampled solutions $\Sigma_\theta$ compatible with $\Sigma_y$. We found that

$$average\ FUV \cong 0.75$$

and we found this number rather insensitive to the width of the sample space. That is, in absence of detailed information about $\Sigma_\theta$, for the given $\Sigma_y$ the sole knowledge of $\theta_3$ is expected to contribute to 25% of the knowledge of $\theta_1$. This analysis shows that indeed joint distributions may help refine the knowledge of non-identifiable parameters given the observation of identifiable parameters correlated with them.

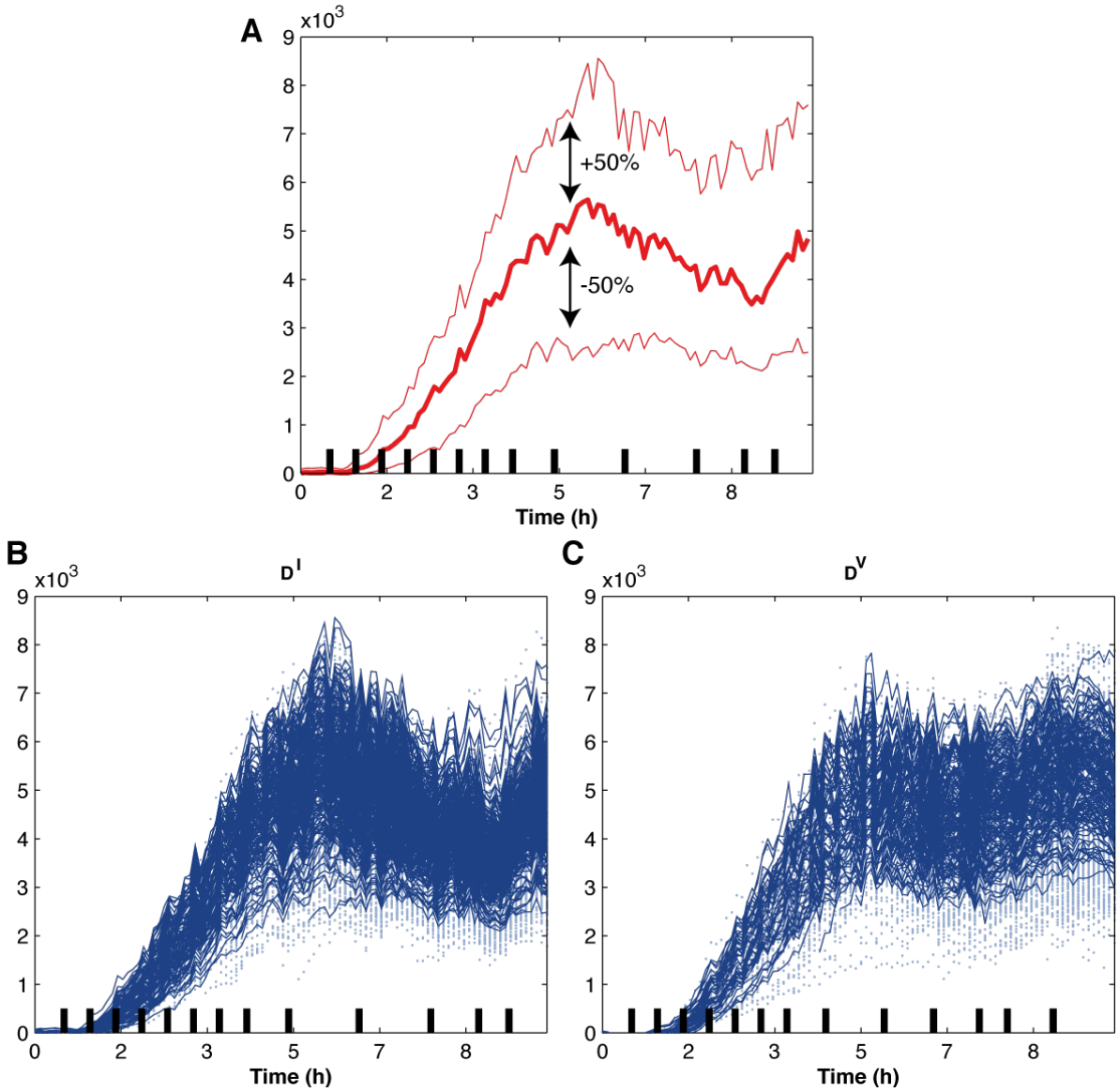



**Figure S1. pSTL1 expression in response to repeated osmotic stresses shows a high level of variability between cells.** A. Minimum, maximum and average cellular fluorescence levels in the identification dataset $\mathcal{D}^I$. Back bars represent input shocks. B. Set of single cell trajectories present in the identification dataset $\mathcal{D}^I$ (solid lines). Samples that did not pass all quality tests described in Text S1 appear as light blue dots. C. Set of single cell trajectories present in the validation dataset $\mathcal{D}^V$.

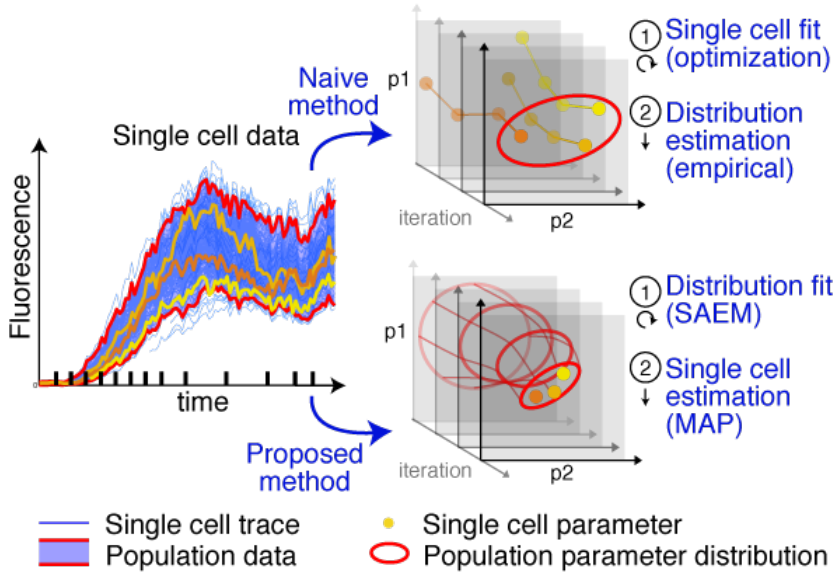
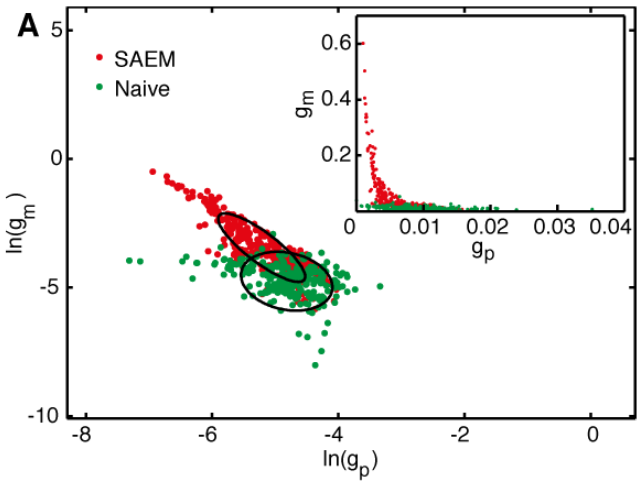



**Figure S2: Statistical inference methods for single-cell and population parameter estimation.** In the naive approach, optimization is used to seek -for each cell- parameter values fitting the individual behavior of the cell via residual minimization (top, step 1). The distribution describing all of the estimated parameter values is then deduced (top, step 2). In the proposed method, the SAEM tool is used to infer a distribution that explains the set of individual behaviors at the distribution level (bottom, step 1). Parameter values for single cells are then estimated based on the particular behavior of the cell and the inferred distribution for the population, using maximum *a posteriori* estimation (bottom, step 2).




B

| | Parameter distribution structure | Parameter distribution spread |
|---|---|---|
| SAEM | 0.62 | 6.7 |
| Naive | 0.19 | 37 |

**Figure S3: The distribution that better describes the entire population is more compact and more structured.** A. 2D plot describing the distribution of the (logarithm of) single-cell parameters for two parameters (insert: same data shown in natural scale). The ellipses represent the region in which 50% of the parameters are distributed. B. Two metrics were computed to quantify the difference in the structure of the parameter distributions at a more global level. The first metric was the average of the coefficients of the variation matrix (i.e. of the off-diagonal terms $cov_{ij}/(\mu_i\,\mu_j)$) between the parameters of the model; this represents the amount of structure in the parameter distribution and shows that SAEM yielded a more structured parameter distribution. The second metric was the volume in the parameter space of the 95%-confidence ellipsoid associated with the covariance matrix. This yields a measure of the typical volume of parameter space occupied by the parameter distribution, and therefore, quantifies the spread of the parameter distributions. This showed that the SAEM approach described the population with a smaller distribution.

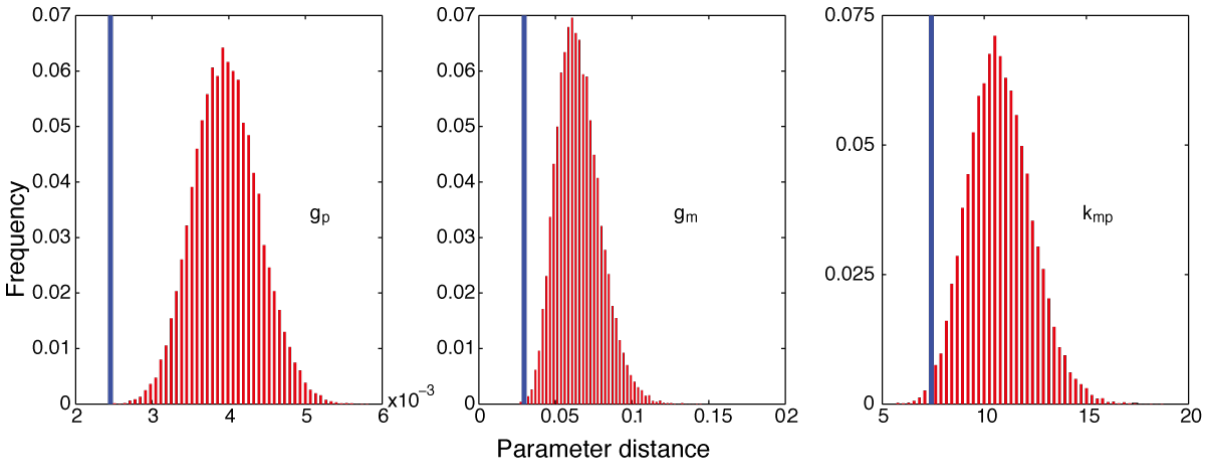



**Figure S4. Average parameter distance of Mother-Daughter pairs against random pairs from the same experiment.** The blue bar represent the average distance in parameters between 55 mother-daughter pairs from experiment $\boldsymbol{D^I}$. The red distribution is obtained by bootstrapping 20000 sets of 55 random pairs of cells (from the same experiment). We see that the distance is very significantly smaller for mother-daughter pairs.

**Table S1. Parameter estimates for the mixed-effects model using the naive inference approach (A), using SAEM on the identification dataset $D^I$ (B) and using SAEM on the validation dataset $D^V$ (C).** (A) Initial values for the search have been obtained by global optimization (CMAES) on the mean behavior starting from literature-based parameters. The value of the delay $\tau$ has been fixed for all cells to its mean-cell. Therefore, statistics on its variability have been shaded. The dataset used is the identification set $D^I$. (B and C) The parameter search is initialized with parameter means extracted from the literature and a diagonal covariance matrix. The parameter search has been adapted to account for the structural non-identifiability relation of $k_m$ and $k_p$ (only their product is relevant in single-cell models): the mean of $k_p$ is kept at a constant value during the search. No constraints are placed on its variance though. The value of the delay $\tau$ is estimated but is set identical for all cells. The dataset used for identification is $D^I$ (B) and $D^V$ (C). The relative standard errors of the estimated moments are typically less than 2%, with the exception of the estimate of $\mathrm{SD}[k_m]$ where it was 8%.

**A**

| Parameter | Inferred Values | | Initials Values | Units | Parameter | Inferred Values | | Initials Values |
|---|---|---|---|---|---|---|---|---|
| | Nat. Values | Log. Values | Log. Values | | | Nat. Values | Log. Values | Log. Values |
| $\mathrm{E}[k_m]$ | 5.47 | 1.61 | 1.49 | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_m g_m)$ | 0.128 | 0.167 | NA |
| $\mathrm{E}[g_m]$ | $1.38{\cdot}10^{-2}$ | -4.75 | -4.74 | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_m k_p)$ | -0.172 | -0.196 | NA |
| $\mathrm{E}[k_p]$ | 1.23 | $7.81{\cdot}10^{-2}$ | $5.66{\cdot}10^{-2}$ | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_m g_p)$ | 0.297 | 0.328 | NA |
| $\mathrm{E}[g_p]$ | $9.76{\cdot}10^{-3}$ | -4.82 | -4.81 | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_m \tau)$ | 0 | 0 | NA |
| $\mathrm{E}[\tau]$ | 23.4 | 3.15 | 3.15 | min | $\mathrm{Corr}(g_m k_p)$ | 0.342 | 0.424 | NA |
| $\mathrm{SD}[k_m]$ | 2.41 | 0.421 | NA | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(g_m g_p)$ | -0.136 | -0.205 | NA |
| $\mathrm{SD}[g_m]$ | $1.73{\cdot}10^{-2}$ | 0.971 | NA | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(g_m \tau)$ | 0 | 0 | NA |
| $\mathrm{SD}[k_p]$ | 0.667 | 0.508 | NA | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_p g_p)$ | 0.216 | 0.244 | NA |
| $\mathrm{SD}[g_p]$ | $6.60{\cdot}10^{-3}$ | 0.614 | NA | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_p \tau)$ | 0 | 0 | NA |
| $\mathrm{SD}[\tau]$ | 0 | 0 | NA | min | $\mathrm{Corr}(g_p \tau)$ | 0 | 0 | NA |
| $\varepsilon_a$ | 45.3 | 3.81 | 3.62 | AU | | | | |
| $\varepsilon_b$ | $9.13{\cdot}10^{-2}$ | -2.39 | -1.51 | - | | | | |

**B**

| Parameter | Inferred Values | | Initials Values | Units | Parameter | Inferred Values | | Initials Values |
|---|---|---|---|---|---|---|---|---|
| | Nat. Values | Log. Values | Log. Values | | | Nat. Values | Log. Values | Log. Values |
| $\mathrm{E}[k_m]$ | 14.7 | 2.63 | 2.30 | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_m g_m)$ | 0.320 | 0.432 | 0 |
| $\mathrm{E}[g_m]$ | $6.00{\cdot}10^{-2}$ | -3.45 | -1.22 | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_m k_p)$ | -0.0647 | -0.0753 | 0 |
| $\mathrm{E}[k_p]$ | 1.19 | $-5.45{\cdot}10^{-2}$ | $-5.45{\cdot}10^{-2}$ | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_m g_p)$ | -0.324 | -0.376 | 0 |
| $\mathrm{E}[g_p]$ | $6.45{\cdot}10^{-3}$ | -5.22 | -5.52 | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_m \tau)$ | 0 | 0 | 0 |
| $\mathrm{E}[\tau]$ | 37.0 | 3.61 | 3.40 | min | $\mathrm{Corr}(g_m k_p)$ | 0.627 | 0.746 | 0 |
| $\mathrm{SD}[k_m]$ | 4.80 | 0.319 | 0.330 | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(g_m g_p)$ | -0.416 | -0.843 | 0 |
| $\mathrm{SD}[g_m]$ | $9.66{\cdot}10^{-2}$ | 1.13 | 0.330 | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(g_m \tau)$ | 0 | 0 | 0 |
| $\mathrm{SD}[k_p]$ | 0.902 | 0.674 | 0.330 | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_p g_p)$ | -0.289 | -0.382 | 0 |
| $\mathrm{SD}[g_p]$ | $4.12{\cdot}10^{-3}$ | 0.585 | 0.330 | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_p \tau)$ | 0 | 0 | 0 |
| $\mathrm{SD}[\tau]$ | 0 | 0 | 0 | min | $\mathrm{Corr}(g_p \tau)$ | 0 | 0 | 0 |
| $\varepsilon_a$ | 63.9 | 4.16 | 5.99 | AU | | | | |
| $\varepsilon_b$ | $8.67{\cdot}10^{-2}$ | -2.45 | -1.20 | - | | | | |

**C**

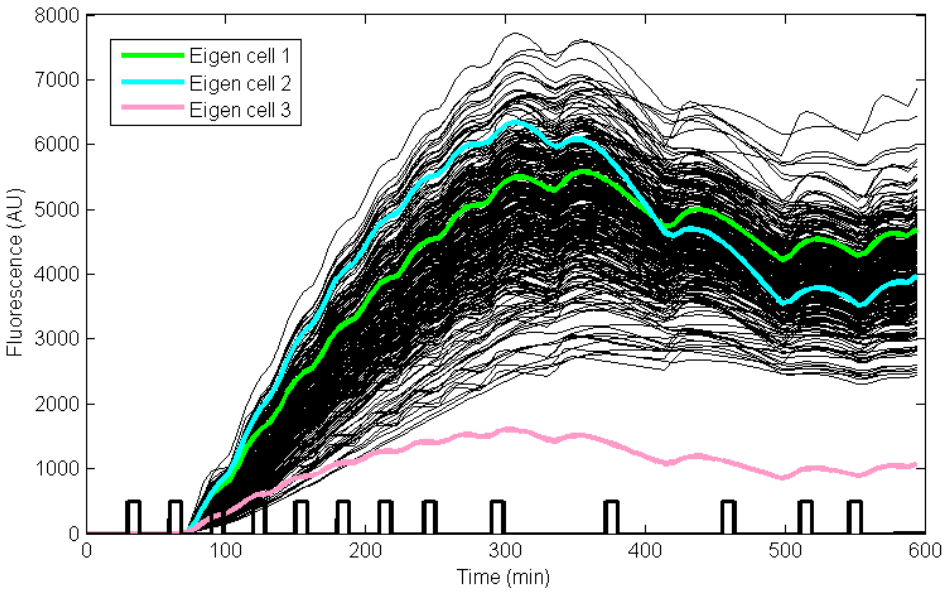
| Parameter | Inferred Values | | Initials Values | Units | Parameter | Inferred Values | | Initials Values |
|---|---|---|---|---|---|---|---|---|
| | Nat. Values | Log. Values | Log. Values | | | Nat. Values | Log. Values | Log. Values |
| $\mathrm{E}[k_m]$ | 9.33 | 2.10 | 2.30 | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_m g_m)$ | 0.604 | 0.733 | 0 |
| $\mathrm{E}[g_m]$ | $4.17{\cdot}10^{-2}$ | -3.80 | -1.22 | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_m k_p)$ | 0.329 | 0.359 | 0 |
| $\mathrm{E}[k_p]$ | 1.08 | $-5.45{\cdot}10^{-2}$ | $-5.45{\cdot}10^{-2}$ | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_m g_p)$ | $2.52{\cdot}10^{-2}$ | $2.99{\cdot}10^{-2}$ | 0 |
| $\mathrm{E}[g_p]$ | $4.36{\cdot}10^{-3}$ | -5.65 | -5.52 | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_m \tau)$ | 0 | 0 | 0 |
| $\mathrm{E}[\tau]$ | 50.4 | 3.92 | 3.40 | min | $\mathrm{Corr}(g_m k_p)$ | 0.578 | 0.706 | 0 |
| $\mathrm{SD}[k_m]$ | 5.08 | 0.509 | 0.330 | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(g_m g_p)$ | -0.247 | -0.458 | 0 |
| $\mathrm{SD}[g_m]$ | $6.57{\cdot}10^{-2}$ | 1.12 | 0.330 | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(g_m \tau)$ | 0 | 0 | 0 |
| $\mathrm{SD}[k_p]$ | 0.605 | 0.521 | 0.330 | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_p g_p)$ | $-4.05{\cdot}10^{-2}$ | $-4.88{\cdot}10^{-2}$ | 0 |
| $\mathrm{SD}[g_p]$ | $3.17{\cdot}10^{-2}$ | 0.652 | 0.330 | $\mathrm{min}^{-1}$ | $\mathrm{Corr}(k_p \tau)$ | 0 | 0 | 0 |
| $\mathrm{SD}[\tau]$ | 0 | 0 | 0 | min | $\mathrm{Corr}(g_p \tau)$ | 0 | 0 | 0 |
| $\varepsilon_a$ | 68.6 | 4.23 | 5.99 | AU | | | | |
| $\varepsilon_b$ | $6.42{\cdot}10^{-2}$ | -2.75 | -1.20 | - | | | | |

## 7. Simulation of Eigen cell behavior

Performing a PCA yields a new parametrization where the new parameters (called principal components) are linearly independent from each other. The basis in which these new parameters are decomposed can be expressed as a set of *eigen cell*. This theoretical cell represents which aspect of the initial variability is captured by a given principal component. Here, we represent the principal components that were obtained when performing a PCA analysis on the 325 cell of the experiment Di (see III.3.a) and which yield the following parameters for the three eigen cells.

| | kmp | gm | gp | tau |
|---|---|---|---|---|
| Eigen Cell 1 (87%) | 17,4 | 0,054 | 0,0041 | 37 |
| Eigen Cell 2 (12%) | 19,0 | 0,028 | 0,0087 | 37 |
| Eigen Cell 3 (<1%) | 9,1 | 0,065 | 0,0074 | 37 |

**Table of parameter values for the three eigen cells. The figure in parenthesis is a reminder of the proportion of the total variability which is accounted by a given principal component and therefore, by the corresponding eigen cell.**




**Simulation of the behavior corresponding to the three eigen cells. Thin black lines are the simulations for all single cells parameters estimated. Black pulses are the applied osmotic shocks. Experiment Di.**

# 8. Developing an Open Source, single-cell optogenetic system

***Objective***

In order to activate optogenetic systems at the single cell scale, we need to project an image directly on the sample through the microscope objective. The image is projected at the microscope focal plane (and is therefore on the sample when the sample is focused).

Various optogenetic systems respond to different wavelength, therefore, it should be possible to change the illumination wavelength.

***Principle***

We use a RGB LED projector which uses a Digital Mirror Device (DMD) technology. If necessary, we change the LEDs in order to project the wavelength of interest.

Coupling with the microscope is based on common optic elements and several 3d printed (or 4 axis machined) custom parts (nb : custom part can be ordered through commercial prototyping services). The coupling should allow focusing the image on the sample plane, translation and rotations for proper alignment. We enter the microscope column through a fluorescent illumination port (at the back).

We drive the system using Matlab©, allowing calibration (mapping camera pixels to DMD pixels), simple manual utilization by drawing ROI to be projected, automated utilization from segmentation/ tracking images of yeast cells.

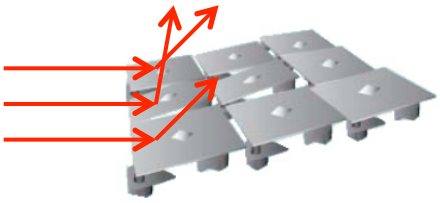


Figure 82 - DMD devices are large arrays of micrometric mirrors. Each mirror forms a pixel of the projected image and can tilt individually in order to either reflect incident light (in which case the pixel is illuminated) or divert it (in which case the pixel will be black).

***Realization***

Digital Mirror Devices are used in various commercial projectors and in many industrial applications (structured illumination, computer vision, 3D printers by stereolithography etc.). As represented in Figure 82, DMD are arrays of micrometric mirrors which can tilt independently. Projectors including a DMD have a single light source (alternating quickly Red, Green and Blue for color projectors) illuminating a DMD. Mirrors tilt determine if the pixel is ON or OFF (gray values are obtained by pulse Width Modulation, PWM)

The starting element is a DMD development kit depicted on Figure 83. The kit is composed of a control board which includes video inputs (DVI, HDMI etc.) and drives a DMD and RGB LEDs. In addition, this kit includes a light engine (Figure 83 B) where are mounted the DMD, the LEDs along with several optical elements.

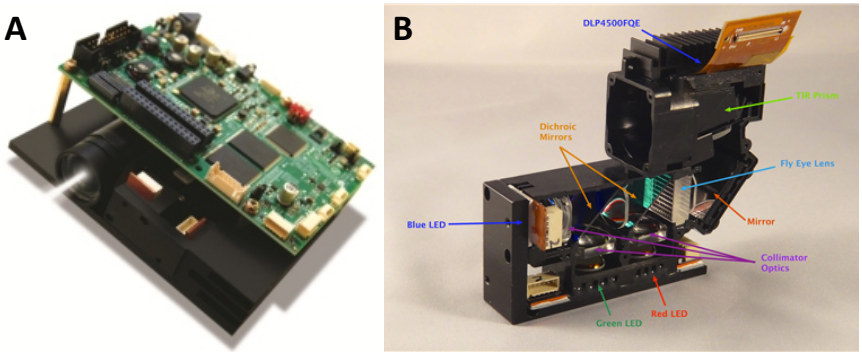



Figure 83 – A. View of the DMD development kit used in this system. B. Close up viev of the light engine (opened, objective removed) which comes with the kit. It sells fror Texas Instruments © for about 1500€. It has a resolution of 1280x800 and a refreshin rate of 120 Hz in grayscale and >4KHz in binary.

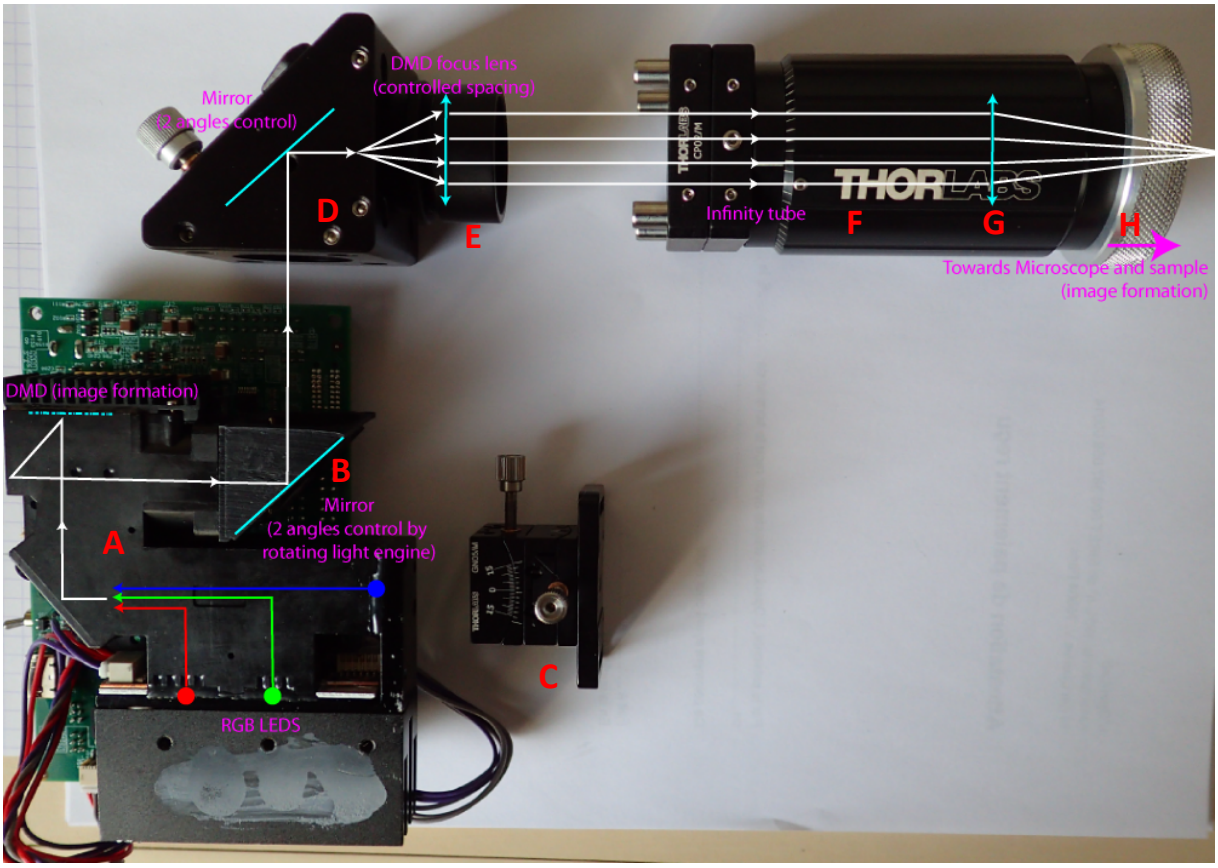



**Figure 84 – Concept view of the optical coupling of the DMD with the microscope fluorescent back port. Using the original light engine (A), we removed the objective and replaced it with a custom made 45° mirror holder (B). The whole light engine can be rotated using a 2-axis goniometer (C). The image bounces off towards a second 45° mirror with two degrees of rotation (D). A first converging lens (E) makes the image of the DMD at infinity (DMD is at the focal plane). Lights travels through an helicoidal lens holder (F) and reaches a second converging lens (G) which has the sample in its focal plane (in reality, a virtual image of the sample, located around 20 cm inside the microscope). The lens holder is connected to fluorescent illumination port at the back of the microscope using a C-mount adaptor (H).**

The projector is mounted on the microscope using a fluorescence illumination port. It could also be mounted using an unused camera port but the first option has the advantage of being able to use filter cubes in DMD light path (for now we simply use a empty filter cube with a beam-splitter in place of a dichroic). In this configuration, the projector will project exactly in the focal plane of the microscope (which means the projected image is always in focus, regardless of the objective choice or of the focus wheel position).

In order to connect the light engine to our microscope, we employed a strategy where we first produce an image at infinity of the DMD (using lens E in Figure 84) which travels at infinity in an infinity tube (portion from E to F in Figure 84) where it encounters a second lens (G in Figure 84) which focus plane is on the microscope sample plane (the effective distance depends on the microscope).

Importantly, in the montage represented in Figure 84, we have separate focusing elements for the DMD and the sample which makes the initial focusing easier. Also, the light engine as well as the mirror cube (D Figure 84) each have two axis of rotation. Combined together, this allows the fine positioning and orientation of the projected image with the sample plane.

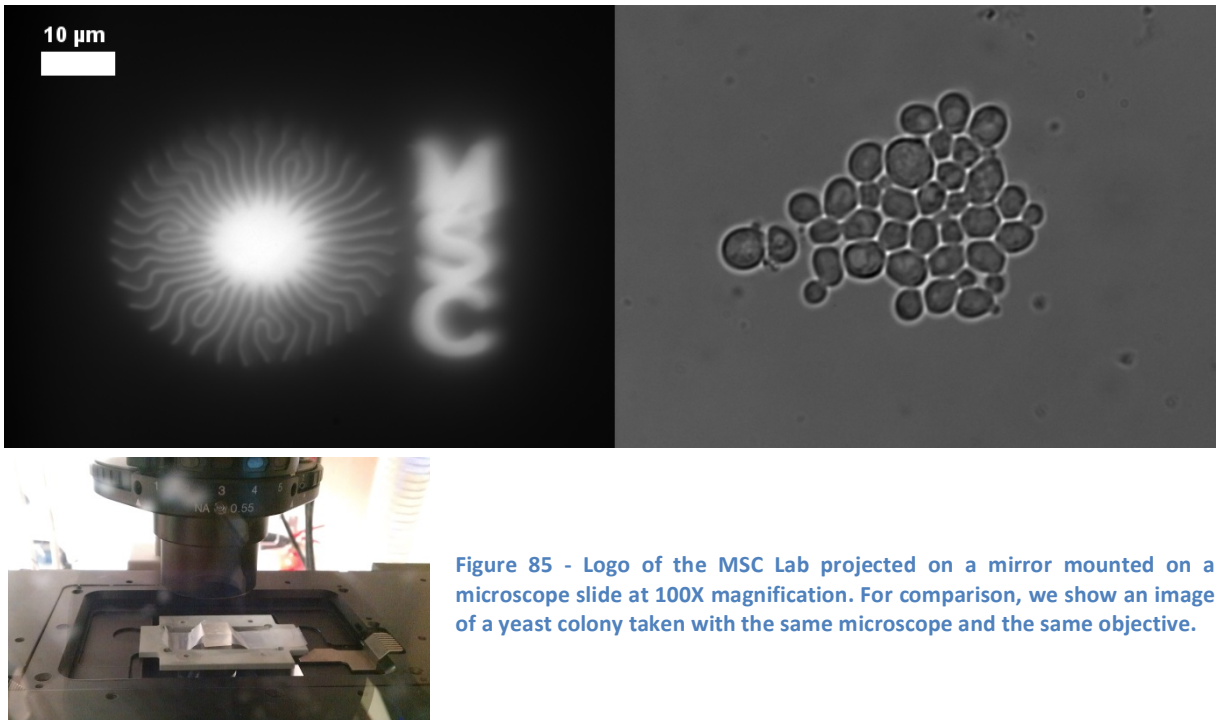



**Figure 85 - Logo of the MSC Lab projected on a mirror mounted on a microscope slide at 100X magnification. For comparison, we show an image of a yeast colony taken with the same microscope and the same objective.**

In order to visualize (under the microscope) the projected image, we place a small mirror on a glass slide and focus the microscope objective on its reflective surface. In Figure 85 we projected the Logo of the MSC lab and imaged it at 100x. We see that although slightly blurry, we can achieve precise and very small illumination. This is clearly sufficient for the purpose of single cell independent optogenetic induction, provided a safe margin around cells is maintained to avoid cross illumination at the borders.

**Software**

In order to drive the device, we use Matlab©. I designed a simple dedicated set of scripts and class which allows easily to recalibrate pixels (*i.e.* refresh the mapping between DMD pixels and Camera pixels if the device has been moved or misaligned.). To do so, the DMD projects a reference image and the user images it with the microscope. The user is prompted to indicate 3 reference points in the original and recorded images. From these reference points, an affine transformation is computed and the DMD is ready to use.

The dedicated class has methods to turn on, off, display any image and draw and image to be projected.

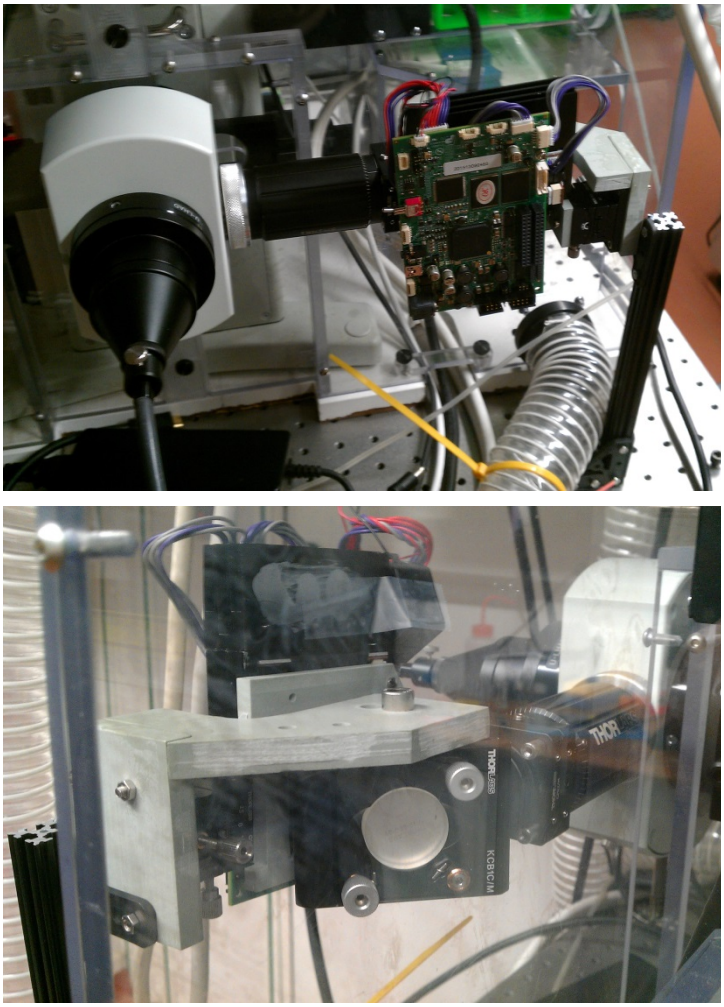



**Figure 86 - Picture of the DMD system mounted on the back of the microscope.**

In Figure 86, we can see how it is in the end integrated at the back of the microscope. Custom pieces (in grey) can also be ordered for people who do not have access to fabrication equipment. As for now, it is tested on mammalian cells, using a blue light optogenetic system by a PostDoc in our research group. Next developments are: building an enclosure and changing one LED to have a different illumination wavelength.