

Phenotypic Variability, Cell Size and Population Fitness

A thesis presented for the degree of
Doctor of Philosophy of Imperial College London
and the
Diploma of the Imperial College (DIC)
by

Pablo Crotti

Department of Mathematics
South Kensington Campus
Imperial College
London, SW7 2AZ

DECEMBER 2015

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Pablo Crotti

Copyright

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

To my family and friends.

Abstract

In all biological systems, phenotypes are quantitative measures of an organism's traits. To understand the relation between a genotype and phenotypes, the influence of genetic perturbations, such as gene deletions, on phenotypes is studied. Although it is common practice to look at the mean value of phenotypes to detect changes, it is less common to model such modifications by considering phenotypic variability. However, phenotypic variability could also contain important information about the phenotype-genotype map. Also, phenotypic variability is a determinant of an organism's robustness and fitness.

This thesis is a collection of five studies devised to identify and analyse phenotypic variability in *S. pombe*, *S. cerevisiae* and *E. coli*. The first chapter determines the probabilistic nature of the dynamic transition from quiescence to proliferation in *S. pombe*. Based on the viability and the random nature of the quiescence of knockout genes, we demonstrate that two competing stochastic models explain the data equally well.

The second and third chapters provide novel approaches to identifying mutant genes that increase/decrease phenotypic variability throughout the cell cycle of *S. cerevisiae*. Using machine learning algorithms, we observe that mutants divide in four categories acting positively and/or negatively on phenotypic variability. Additionally, we show that phenotypic robustness is inversely related to cell fitness.

The fourth chapter builds morphological interaction networks in *S. cerevisiae*. By employing a Bayesian network framework, we show that cell size dictates nuclear size throughout most of the cell cycle.

The fifth chapter investigates the growth rate of bacterial populations subject to fluctuations in the variance of their cell size distributions. By employing a population balance equation combined with an agent-based model, our results indicate that the cell size distribution affects the growth rate. Moreover, we show that cells possess a cell size regulator within their cycle.

Acknowledgements

I want to express my deepest gratitude towards my supervisor Vahid Shahrezaei who, throughout the years of my research, guided and mentored me with the right words and his fruitful advice. His investment in my work allowed me to broaden my knowledge in mathematics, physics and biology and I am really thankful for this.

This research would not have been possible without the great advice and help from my collaborators: Michael Cook, Mike Tyers, Bruno Lages, Jurg Bähler, Marco Scutari, Sebastian Krumscheid and Samuel Marguerat.

I also want to thank the Department of Mathematics for awarding me a Roth studentship, and the Engineering and Physical Sciences Research Council which covered my tuition fees and expenses.

Finally, I would like to thank my parents, Sylvie and Claudio, for their encouragement, advice and support along the way, my brother David and his lovely family: Luca, Salome and Virginia, for their love, and all my colleagues and friends with whom I shared great moments and interesting conversations.

Pablo Crotti

Contents

Abstract	5
1 Outline	13
2 Biological and Mathematical Background	21
2.1 Mathematical Prerequisites	22
2.1.1 Linear Models and Regression Analysis	22
2.1.2 Machine Learning Algorithms	25
2.2 Biological Prerequisites	34
2.2.1 From DNA to Phenotypic Variability	34
2.2.2 Cell Cycle	37
3 Modelling the dynamic transition from cellular quiescence to proliferation in <i>S. pombe</i>	39
3.1 Introduction	39
3.2 Biological Experiments and Acquisition of the Growth Curves	41
3.3 Models	43
3.3.1 Stochastic Lag-time	44
3.3.2 Fractional Growth of the Population	47
3.4 Results	47
3.4.1 Stochastic Lag	49
3.4.2 Fractional Growth of the Population	55
3.5 Discussion	58
4 Global trends in Phenotypic Variability and Robustness	59
4.1 Introduction	59
4.2 Processing of the <i>Saccharomyces Cerevisiae</i> Morphological Database and Properties of the Morphological Phenotypes	62
4.2.1 Removal of the Phenotypes and Mutants with Non-significant Number of Single-cell Measurements	63
4.2.2 Clustering of the Phenotypes into Classes of Representatives	64
4.2.3 Genetic Properties of the Morphological Phenotypes	66

4.3	Scoring of Phenotypic Variability	68
4.3.1	Residuals for the Coefficient of Variation	69
4.3.2	Global Phenotypic Variability of the Mutants	71
4.3.3	Comparison with Other Measures of Variability	73
4.4	Neutral, Maintainer, Inducer and Capacitor Mutants	78
4.4.1	Segregation of the Mutants into Categories according to their Phenotypic Variability	79
4.4.2	Gene Ontology Annotations	85
4.4.3	Analysis of Robustness and Phenotypic Variability	87
4.5	Control of the Methodology	89
4.6	Discussion	91
5	Extraction and Analysis of Haploid Genes Modifying Phenotypic Variability and Cell Cycle Regulation in <i>S. cerevisiae</i>	93
5.1	Introduction	93
5.2	Gene Selection Based on Mean Distance and Residuals of the Wild types	99
5.2.1	Mean Distance Method	101
5.2.2	Residuals Based on the Wild Types	102
5.2.3	Permutation Tests	103
5.2.4	Results	105
5.3	Algorithms for Identification of Related Groups of Multi-gene Outliers in 2-dimensional Data Sets	109
5.3.1	Distance between Ellipsoids	110
5.3.2	K-medoids clustering	112
5.3.3	Acceptance/Rejection Test	113
5.4	Results	116
5.5	Discussion	120
6	A Primer on Cell Size and Nuclear Size Regulation in <i>S. cerevisiae</i> using Bayesian Network Analysis	123
6.1	Introduction	123
6.2	Single-cell Measurements & Motivation	126
6.2.1	Morphological Phenotypes and Single-cell Distributions	126
6.2.2	Conditional Dependence between the Mother Cell Size and the Nuclear Size in Mother	129
6.3	Bayesian Networks Analysis of the Wild Types	133
6.3.1	Cell Cycle Stage Networks and Equivalence Classes	134
6.3.2	Bootstrapping the Networks for all Cell Cycle Stages	139
6.4	Bayesian Networks Analysis of the Mutants	142
6.5	Discussion	145

7	Bacterial Population Growth and Variability in Cell Size and Division Process	147
7.1	Introduction	147
7.2	Time-dependent Growth Rate of the Population	150
7.2.1	Elongation of a Single Cell	150
7.2.2	The Population Balance Equation	153
7.2.3	The Agent-Based Model	159
7.2.4	Computation of the Growth Rate $K(t)$	162
7.3	Growth Rate of the Population at Steady State	164
7.3.1	Computation of the Cell Cycle Time Distribution	168
7.3.2	Relationship between the Growth Rate at Steady State and the Average Cell Cycle Time	172
7.4	Cell Size Regulation	175
7.4.1	Linear Dependence between Birth and Division Lengths	176
7.4.2	Data Analysis of Biological Experiments on <i>E. coli</i>	181
7.5	Discussion	188
8	Conclusion and Outlook	191
8.1	Stochastic Lag-time and Fractional Growth of the Population	191
8.2	Inducer, Maintainer and Capacitor Mutants	192
8.3	Phenotypic Variability and Cell Size Regulation in <i>S. cerevisiae</i>	193
8.4	Cell Size and Nuclear Size Regulation with Bayesian Networks in <i>S. cerevisiae</i>	194
8.5	Bacterial Growth and Size Regulation in <i>E. coli</i>	195
	Bibliography	197
	Appendix	213
A	Modelling the dynamic transition from cellular quiescence to proliferation in <i>S. pombe</i>	213
B	Global trends in Phenotypic Variability and Robustness	219
C	Extraction and Analysis of Haploid Genes Modifying Phenotypic Variability and Cell Cycle Regulation in <i>S. cerevisiae</i>	227
D	A Primer on Cell Size and Nuclear Size Regulation in <i>S. cerevisiae</i> using Bayesian Network Analysis	233
E	Bacterial Population Growth and Variability in Cell Size and Division Process	245

List of Figures

2.1	Fundamental connections in Bayesian networks	32
2.2	Cell cycle in eukaryotes and prokaryotes	38
3.1	Schematic of the dynamic transition from quiescence to proliferation	41
3.2	Typical growth curve of a population of cells leaving quiescence	42
3.3	Simulation of the transition from quiescence to proliferation	49
3.4	Representation of the wild types at 25 degrees Celsius	50
3.5	Representation of the wild types and mutants at 25 degrees Celsius	52
3.6	Representation of the wild types at 32 degrees Celsius	53
3.7	Representation of the wild types and mutants at 25 degrees Celsius	54
3.8	Growth of the mutants with no quiescence	55
3.9	Dispersion of the fraction of wild-type cells growing	56
3.10	Growth of the fraction of mutants and wild types	57
4.1	Partitioning around medoids of the morphological phenotypes	66
4.2	Comparison of genetic and stochastic variability between the mutants and wild types	68
4.3	Construction of the residuals based on a $L1$ -regression	72
4.4	Distribution of the Global Phenotypic Variability score	73
4.5	Validation of the GPV with other genetic scores	75
4.6	Distribution of the GPV for the wild types	81
4.7	Separation of the mutants into categories according to their GPV	82
4.8	Identification of complex increasing/decreasing burst frequency	84
4.9	Gene Ontology annotations of the four categories of mutants increasing/decreasing phenotypic variability	86
4.10	Robustness of phenotypes to the fitness of the cells	88
5.1	Variation in the distribution of the morphological phenotypes for the mutants and wild types	95
5.2	Comparison of mean and cv of the mother cell size throughout the cell cycle	98
5.3	Gene Ontology annotations for the genes found using the residuals and mean distances	106

5.4	Illustration of the mean distances and residuals for the nuclear size of post-mitotic buds	108
5.5	Score phase allowing to accept or reject a GO annotation	115
5.6	Separation algorithms devised to identify gene perturbing phenotypic variability in both the mean and cv	117
5.7	Variability of the mutants for the DNA damage category	119
6.1	Phenotypic distributions of the wild types in stage A	128
6.2	DREVI analysis of the relationship between the cell and nuclear sizes . . .	132
6.3	Tuning of the Max-Min Hill Climbing algorithm	134
6.4	Standard Bayesian network of the wild types in stage A	135
6.5	Standard Bayesian network of the wild types in stage A1B	136
6.6	Standard Bayesian network of the wild types in stage C	138
6.7	Bootstrapped Bayesian networks for the wild types in stages A, A1B and C	141
6.8	Heatmap of the strength and direction of the arcs for the mutants in stages A, A1B and C	144
7.1	Description of the morphological parameters for the population balance equation	152
7.2	Illustration of the probability distribution and division rate employed in the population balance equation	153
7.3	Description of the accumulation of rates employed in the population balance equation	154
7.4	Schematic of the agent-based model employed to simulate the population balance equation	161
7.5	Sample trajectories produced by the agent-based model	165
7.6	Analytical distribution of the cell cycle time	174
7.7	Variation of the switching parameter s between the Timer and Adder models	176
7.8	Comparison between the analytical and numerical distribution of cell cycle times	178
7.9	Effect of a cell cycle regulator on the cell cycle times	180
7.10	Typical elongation of a cell in the mother machine	184
7.11	Extraction of the growth parameters from the mother machine data	187

List of Tables

2.1	Algorithm: K-medoids	26
3.1	Akaike Information Criterion for the stochastic lag and fractional growth of the population	48
4.1	Medoids found using the Partitioning Around Medoids algorithm on the morphological phenotypes	66
4.2	Correlation between the GPV and the local and global genetic scores from Rinott et al. [114]	75
4.3	Correlations between the GPV and a smaller set of medoids	77
4.4	Proportion of shares ORFs between the categories found using either the CV or $L1$ -regression	90
5.1	Phenotypes employed in the analysis of cell size variability	100
5.2	Algorithm: Hypergeometric test on the Gene Ontology annotations	104
5.3	Algorithm: Ellipsoid separation	111
5.4	Algorithm: K-medoids separation	113
5.5	Algorithm: Acceptance/Rejection test for the Gene Ontology annotations	114
6.1	Raw morphological phenotypes available in the SCMD	127
6.2	Number of single-cell measurements for the phenotypes for the stages A, A1B and C	128
6.3	Algorithm: DREVI visualisation	130
7.1	Description of the parameters and variables employed in the agent-based model	162
7.2	Morphological parameters available via the mother machine device	181
7.3	Algorithm: Simulation of the exponential elongation	183

Chapter 1

Outline

All biological organisms and their machinery are characterised by their DNA and the transcription and translation of it into proteins. More specifically, DNA is fractioned into sub-sequences called genes. Each gene can encode for one or multiple proteins. Once created, the proteins allow the organisms to perform specific tasks, such as the creation of other proteins or to regulate the molecular environment within the cell.

To understand how genes regulate the cell machinery, it is convenient to look at a mapping of the genome, i.e., the set of genes, to the phenotypes. Phenotypes are observable features of the organisms and can be separated into two broad categories: molecular phenotypes and morphological phenotypes. For instance, molecular phenotypes can represent the level of gene expression, i.e., how DNA encodes for a particular gene. The second category of phenotypes refers to the morphology of the cell. For example, a morphological phenotype can be the volume of the nucleus or the cytoplasm.

At both the molecular and morphological levels, phenotypes show variability. At the molecular level, gene expression has two types of variability (see Swain et. al [131] and Shahrezaei & Swain [123]). Firstly, the stochastic nature of the transcription and translation processes creates an "intrinsic" variability (or also called noise). Secondly, external factors within the cell, such as variations in the amount of cellular components lead to an "extrinsic" noise. For the morphological phenotypes, variability can be seen as variations of a specific phenotype among a collection of identical cells. These variations can come either from genetic perturbations, such as gene deletions, or environmental stress, such as

resistance to polluted environments (see Holland et al. [49]). For example, the volume of the nucleus might slightly differ from one cell to another, thus creating a distribution of nucleus volume among the population of cells.

In this thesis, we present five investigations to identify the sources and subsequently the consequences of phenotypic variability on cell size and population fitness. We examine phenotypic variability as a novel approach to characterise and understand the cell cycle. Past analyses have focused on examining morphological distributions of the cells without taking into account variability. Hence, this thesis emphasises on variability. Here, we focus on three unicellular organisms: *Schizosaccharomyces pombe* (*S. pombe*), *Saccharomyces cerevisiae* (*S. cerevisiae*) and *Escherichia coli* (*E. coli*). We employ the aforementioned organisms to examine phenotypic variability for different genetic backgrounds. The five projects are outlined in the next paragraphs as independent chapters.

In chapter 3, we introduce a probabilistic approach to understand the dynamic transition from quiescence to proliferation in *S. pombe*. During quiescence, the cells do not grow but divide until they reach a minimum cell size. Quiescence is induced by nitrogen deprivation. Thereafter, cells quit quiescence when nitrogen is re-injected in the medium. Hence, it is possible to keep the cells in a quiescent state for a certain amount of time. The experiments performed by our collaborator B. Lages at the Bähler Lab (University College London) show that after quiescence, the revived cells will start growing again but only after a certain amount of time called "lag-time". Depending on the number of days spent in quiescence, the lag-time varies greatly from one mutant to another.

Here, we quantify the lag-time by devising two models: a probabilistic model based on the random nature of the lag-time in *S. pombe* and a model based on the viability of the cells when they leave quiescence. The viability hypothesis comes from the idea that a large fraction of the revived cells will start growing immediately after leaving quiescence and the remaining fraction will not grow. We employ the viability in accord with similar results found by Kotte et al. [62] with populations of *E. coli* cells. The second hypothesis, on the random nature of the quiescence phase, has already been investigated theoretically by Baranyi [9, 8]. In this setting, we model the lag-time as a random variable following a probability distribution. Moreover, to control the growth, the elongation rate of the cells is

considered as an additional parameter.

We test the two hypotheses, i.e., the viability and the random nature of the lag on the experimental data gathered by our collaborator. The model on viability contains two parameters: the fraction of growing cells and the elongation rate of the cells. The model on the random nature of the lag contains three parameters: the average lag-time, its variance and the elongation rate of the cells.

Using the Akaike Information Criterion, we show that both models are statistically equivalent. Either the viability or the random nature of the lag-time can explain the dynamic transition from quiescence to proliferation. Moreover, both models allow to extract mutants not following the wild-type trend. In this context, we find that the Set1C complex perturbs the transition.

In the next three chapters, we are interested in understanding phenotypic variability in *S. Cerevisiae*. Here, we employ the *Saccharomyces Cerevisiae* Morphological Database [99], which contains a collection of 4618 gene deletion strains and 126 wild types measured for 501 morphological phenotypes. Phenotypes, such as the cell or nucleus areas, are measured from single-cell measurements and allow to create empirical distributions for all the gene deletion strains and wild types.

In chapter 4, we extend the concept of phenotypic capacitance introduced by Levy & Siegal in their pioneering work on *S. cerevisiae* [76] to identify mutants possibly having an opposite effect (compared to their analysis) on phenotypic variability. By considering the residuals calculated on the pair of: mean phenotypic values and coefficients of variation of certain distributions of morphological phenotypes, they showed that there exist mutants capable of producing more phenotypic variability. Here, we investigate the possibility of having mutants capable of increasing phenotypic variability but also decreasing it. In addition, we look at the robustness of morphological phenotypes relative to the fitness of the cells.

We quantify the effect of a mutant on the morphological phenotype by looking at modified $L1$ -residuals, i.e., residuals computed using an $L1$ -regression, over a range of selected

phenotypes. Here, the phenotypes are chosen after clustering them with the partitioning around medoids algorithm. To see a fold change compared to the wild types, we normalise the residuals by the standard deviation of the wild-type residuals. The essence of this study relies on the creation of a score: the "Global Phenotypic Variability" (GPV), which is a global measure of variability amongst the phenotypes. For a given mutant, the GPV is computed as the median of all its residuals. Using the same approach with the wild types, we then define thresholds for the GPV and categorise the mutants according to their global effect on phenotypic variability.

We validate the GPV score by comparing it with pre-existing scores in the literature. Firstly, we look at the correlation between the scores from Levy & Siegal. Secondly, we look at the scores of Rinott et al. [114] who created a measure of the local and the global effect of genetic perturbations in *S. cerevisiae*. Finally, we compute the GPV on a smaller set of morphological phenotypes to identify the phenotypes highly correlated with the GPV. Both scores from Levy & Siegal and Rinott et al. show good correlations with our GPV score. Therefore, we pursue the analysis with the segregation of the mutants into categories, which indicate the effect of the gene deletion on phenotypic variability. Here, we validate our segregation method by comparing mutants with low and high phenotypic variability against the chromatin regulation complexes found by Weinberger et al. [148]. In their study, they found complexes increasing the burst frequency. A burst is a short amount of time during which proteins are produced for different (smaller or larger) amount. After proceeding with a visual inspection of the complexes we see similar results as Weinberger et al.

Our analysis shows that the mutants can be categorised in four groups. The capacitors, which are mutants increasing variability, the inducers which decrease variability, the maintainers that act as both capacitors and inducers and finally the neutrals who do not show significant phenotypic variability. In addition, we show that robustness is inversely related to the fitness of the cells. More specifically, mutants with low phenotypic variability have faster growth rates than mutants with high phenotypic variability.

Chapter 5 focuses on the variability of the cell size during the cell cycle of *S. cerevisiae*.

Here we present algorithms based on a geometrical separation of the mutants and the wild types, to extract Gene Ontology (GO) annotations. Unlike the approach of chapter 4, here we carefully select phenotypes related to the cell, the nucleus and the bud size. Each phenotype has a mean phenotypic value and a coefficient of variation, calculated on single-cell measurements. In this study, we investigate Gene Ontology annotations compared to the location of the wild types in the mean-cv space, for a given phenotype.

Firstly, we analyse the mean phenotypic value and the coefficient of variation of cell-size-related phenotypes independently. We employ a similar approach as in chapter 4, by considering the residuals of the coefficients of variation. This approach considers only one dimension of the problem, either the mean phenotypic value or the coefficient of variation. Secondly, we devise two algorithms considering both the mean phenotypic value and its coefficient of variation. For both algorithms, the mutants belonging to a given GO annotation are compared to the wild types for a given phenotype. The first algorithm processes the GO annotations by considering the groups of mutants and wild types as ellipses in the mean-cv space. Then, the ratio of the distance between the ellipses is used as a score to measure their separation. The second algorithm processes the mutants in a similar fashion, except that the separation method is based on a labelling/clustering approach, to see whether a mutant should belong to the wild-type group and vice versa. Here, each group (mutants or wild types) is given a label. Then, after clustering the mutants and wild types together in two clusters, we regard the rate of misplaced labels to create a score of the goodness of geometrical separation of the two groups.

The significance of the scores provided by the two algorithms is measured using permutation tests on the GO annotations. Thereafter, we only consider annotations with significant p-values.

For both methodologies, in one and two dimensions, we find similar results. Mutants increasing variability are related to DNA damage, spindle and chromosome segregation, chromatin remodelling and vacuolar pH. Mutants decreasing variability are more related to amino acids, vesicle function, tRNA and rRNA processing.

Chapter 6 is a Bayesian network (BN) framework created to analyse the relationship

between the nucleus and the cell size of *S. cerevisiae*. Throughout the cell cycle of budding yeast, the cell and the nucleus grow until the cell divides to produce a new daughter. However, it is unclear whether the cell (size) dictates the nucleus (size) or if it is the opposite relationship. We investigate this relationship by considering Bayesian networks constructed on the distributions of the morphological phenotypes. By their nature, Bayesian networks can represent the causal relations of random variables, here representing morphological phenotypes.

The investigation starts by looking at the networks of the wild types. Here, we construct the networks according to the cell cycle stages in which the wild types belong. Firstly, we look at the standard network, i.e., the network learned by simply processing the morphological distributions. Secondly, in order to validate the direction of the arcs between the nodes, i.e., the phenotypes, we look at the equivalent class of the standard networks. Finally, we assess the causation by looking at the bootstrapped analysis of the standard networks. The resulting bootstrapped networks provide information on two criteria: the strength and the direction of the arcs. Therefore, it is possible to confirm whether an arc should be directed toward a specific node. Thereafter, we process in a similar fashion to analysis the mutants. The mutants do not necessarily have enough single-cell measurements, which produces non significant morphological distributions. Here, we remove the mutants with such insignificant distributions.

The database of morphological distributions is divided in three cell cycle stages. The first stage "A" corresponds to the G1 phase. The second stage "A1B" lies in between the S and the G2 phases. Finally, the last stage "C" represents cells at the M phase. Here, we construct the wild-type and mutant networks for the three stages.

Our results indicate that the cell (size) dictates the nucleus (size) in stage A and A1B. However, in stage C the causation between the cell and the nucleus disappears and instead of a direct relationship, the causation needs more phenotypes to go from the cell to the nucleus. The networks constructed for the mutants show that a large proportion of the mutants possess the same arc between the cell and the nucleus. However, there is a non negligible number of mutants that do not have this relationship.

In chapter 7, we provide a theoretical analysis of the growth rate of a population of unicellular cells subject to variability in their cell size distribution. More specifically, we are interested in the cell size distribution of bacterial cells. Here, we focus on two phenotypes: the birth and the division lengths of the cells. We investigate the relationship between variations in the cell size and the growth rate of the population. In addition, we measure the effect of a cell cycle regulator on the fitness of the population. Here, we see the cell cycle regulator as a dependence between the birth and division lengths of the cells, to investigate whether or not both lengths should be related to each other.

Here, the mathematical model devised is a modification of a population balance equation (PBE) adapted to the analysis of *E. coli* cells. In this setting, we provide the PBE with the probability distributions of the birth and division lengths. The PBE is characterised by an integro-differential equation. To our knowledge, no analytical solution has been provided to it. However, the PBE can be simplified to express the growth rate of the population. In this setting, the growth rate of the population is a time-dependent function of the cell size distribution (in the population) and the division rate of the cells. Here, we study the effect of variations in the birth and division lengths on the growth rate of the population. As an intermediary result, we show that the growth rate oscillates for a certain amount of time and then reaches a stationary value. To understand why the growth rate converges to a constant value and also to see whether a cell cycle regulator might affect this convergence, we devise a theoretical formulation of the growth rate at steady state, subject to the cell cycle regulator.

We simulate the PBE with an agent-based model that considers the growth of the cells of the bacterial colony. The modularity of the agent-based model allows varying parameters such as the variance of the birth and division length distributions. In addition to our results found *in silico*, we extract experimental data from *E. coli* cells, collected using a microfluidic device [144], to test our model.

Our results indicate that the growth rate of the population depends on the variance of the birth and division lengths distributions. In addition, we show that the growth rate converges to a steady state as the cell distribution converges to its steady state. At steady state, the variance of the birth and division length distributions still affects the growth rate of the population. Finally, our theoretical model of the cell cycle regulator provides a new

insight into the dependence of the birth and division lengths of the cells when compared to experimental data.

After this outline, a background chapter provides the foundation of the mathematics and biology employed in our projects. Then, the thesis chapters are structured according to the projects we present. The first investigation on *S. pombe* is explained in Chapter 3. The second and third questions for the cell size variability in *S. cerevisiae* are presented in Chapter 4 and Chapter 5, respectively. The Bayesian networks framework for the relationship between the cell and the nucleus sizes is introduced in Chapter 6. The analytical model of the effect of variability on the growth rate of *E. coli* cells is developed in Chapter 7. Finally, a summary of our results is presented in the last chapter.

Chapter 2

Biological and Mathematical Background

The broad field of mathematical biology encompasses multiple areas of the aforementioned fields: mathematics and biology. To understand biological phenomena, it is necessary for mathematicians to understand the key concepts involved in the biological question. Similarly, biologists need to understand the mathematics involved to answer the question. This chapter therefore intends to provide the tools employed to model the diverse biological questions we have studied in this thesis. In addition, we recall main concepts of cellular biology and we illustrate their importance for the models we devise in this thesis.

The first section introduces the mathematics employed in this thesis. The tools are mainly borrowed from statistics, regression analysis and machine learning. We emphasise the last part by introducing Bayesian networks.

The second section presents the concepts of DNA, transcription/translation and phenotypic variability. Finally, we give a short introduction on the cell cycle of eukaryotic and prokaryotic cells.

2.1 Mathematical Prerequisites

When confronted with the modelling of biological phenomena, mathematicians employ tools coming from different areas of mathematics. Although the tools and models are theoretically well defined, it is sometimes impossible to get an accurate calculation by hand. Thus, computers are now the main help to simulate the models, find patterns in data or simply perform standard statistical analyses.

In this section we introduce basic knowledge of probability and statistics and present the tools we employ to categorise groups, as for instance mutants or phenotypes, in the different projects of this thesis. The tools we employed mainly belong to the fields of regression analysis and machine learning. For more information about probability and statistics, the reader is invited to have a look at Davison [32] and Ross [116]. To understand the concepts employed in regression analysis, the following books: Everitt & Hothorn [36], Cohen & Cohen [22] and Stumpf et al. [129] offer an introduction to the topic as well as practical applications using the software R. In order to comprehend the key concepts of machine learning, we recommend the following references: Baldi & Søren [7] and Lantz [68]. For a deeper understanding of machine learning algorithms with an emphasis on probability theory, we recommend Murphy [91].

2.1.1 Linear Models and Regression Analysis

Statistical estimation and regression analysis are two fields largely employed in mathematical biology as it requires estimating or inferring quantities coming from experiments. Here we start our explanation by introducing the Least Squares Estimation which provides a method to fit data and assess their dependence. Following this, we introduce two additional regression methods: the $L1$ -regression and the Lowess regression. The former is similar to the least squares estimation except that it considers a different norm to estimate the fit. The Lowess regression is a weighted regression that gives more importance to points with small variance in the data.

In biology, it is often required to test the dependence between two variables, for exam-

ple, the response of an organisms to a gene deletion. Here, two variables are measured by the biologists. The variable X (called *predictor*), which is for example the gene deletion and the variable Y (called *response*), which is the response to the gene deletion. Let us suppose that at the end of the experiment, both variables X and Y were measured for n organisms, i.e., we removed a gene for n different organisms. Using regression methods, it is possible to decide the relationship between X and Y . Here, we present three regression methods.

Least Squares Estimation

Let X and Y be the two variables as explained previously. Their relationship can be measured as

$$Y = f(X) + \text{noise} \quad (2.1)$$

where f is a function (called regression function) supposed to model the biological problem.

Suppose now that the regression function f is known up to a finite number p of parameters $\beta = (\beta_1, \dots, \beta_p)$. We then denote the function as $f(\mathbf{X}, \beta)$.

For the special case where

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} \quad , \quad i = 1, \dots, n, \quad (2.2)$$

the relationship is called *linear*.

The Least Squares Estimation (LSE) is a method that minimises

$$S_{L2} = \sum_{i=1}^n (y_i - f(x_i, \beta))^2 = \sum_{i=1}^n r_i^2, \quad (2.3)$$

i.e., it needs to find an estimator of β such that

$$\frac{\partial S_{L2}}{\partial \beta_i} = 0 \quad (2.4)$$

for all possible β_i 's. Here, the r_i 's are called *residuals*. These residuals are the "vertical" distance between the original y -data and the fitted line by the LSE.

Here, the "goodness of fit" is assessed by computing the coefficient of determination R^2 . The coefficient is calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - f(x_i, \beta))^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2.5)$$

where $\bar{y}$ is the sample mean of Y , i.e.,

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i. \quad (2.6)$$

By its definition, the values of the coefficient of determination range from 0 to 1. The numerator of Eq.(2.5) is called the *total sum of squares* and the denominator is called the *residual sum of squares*.

A coefficient of 1 indicates a perfect fit of the model to the data whereas a coefficient of 0 indicates that the model does not fit the data at all.

■

L1-Regression

The $L1$ -regression is similar to the LSE except that it tries to minimise

$$S_{L1} = \sum_{i=1}^n |y_i - f(x_i, \beta)|. \quad (2.7)$$

The $L1$ -regression gets its name from the $L1$ -norm used in the optimisation process.

The coefficient of determination can also be utilised for the $L1$ -regression.

■

Lowess Regression

We showed that the LSE minimises the sum of the square of the residuals. Similarly, the Lowess regression (also labelled "Loess"), which stands for "Locally Weighted Scatterplot Smoothing", tries to minimise a weighted sum of the square of the residuals, i.e., Lowess minimises the sum

$$S_{Lowess} = \sum_{i=1}^n w_i [y_i - f(x_i, \beta)]^2 = \sum_{i=1}^n w_i r_i^2 \quad (2.8)$$

where w_i is the inverse of the variance of the measurement i . This forces the fitted curve to be closer to the points with low variance.

Here, the coefficient of determination is not often employed to assess the goodness of fit as the Lowess regression tends to smooth the data.

■

2.1.2 Machine Learning Algorithms

For decades now, progress in biology has not only led to new discoveries but also increased the amount of data produced or gathered within biological experiments. In order to draw conclusions, data is extracted and analysed by specialists that either use the raw data or plug it into mathematical models. However, sometimes data is either too big or not classified enough to be treated "as is". This is where machine learning is necessary. As its name suggests, machine learning is a field in between computer science and mathematics, allowing algorithms to learn from the data. There are different approaches to "learn from the data". Algorithms in machine learning are separated into two categories depending on the degree of learning involved in the process. The first category concerns supervised learning. Supervised learning algorithms are models that infer predictions, based on training labelled examples (or training data) for a similar new set of data. The second category, the unsupervised learning models, tries to understand and re-create the hidden structure of the unlabelled data.

In this thesis, we employ unsupervised learning algorithms, namely the k-medoids al-

Algorithm : k-medoids

1. Randomly choose k initial medoids amongst the n data points.
2. Associate non-medoids points to their closest medoids, i.e., the medoid with the lowest pairwise L1 distance.
3. For each medoid c compute:
 1. For each non-medoid point p
 1. Swap c and p and recompute the total cost of the system
4. Pick the system with the lowest total cost.
5. Repeat points 2. to 4. until the total cost does not change anymore.

Algorithm 2.1: By switching from one data point to another, the k-medoids algorithm finds the largest distance between clusters of points.

gorithm and Bayesian Networks. The later actually belongs to a particular category of models that are based on the joint probability of the data, whereas supervised or unsupervised learning are more of heuristic. In the following subsections, we introduce the k-medoids algorithm and the idea behind Bayesian networks.

K-medoids Algorithm

The k-medoids algorithm is an unsupervised clustering algorithm allowing to partition a dataset into K groups. Each group possesses a representative (called medoid) that is a member of the dataset. To determine which points can be the medoids, the algorithm computes the distances between all points and the medoids and decides whether a medoid should be interchanged with a non-medoid, to decrease a cost function, usually described as a distance function between medoids and non-medoids. The cost function is often the Euclidean distance or the L1-norm.

This interchange of medoids is called partitioning around medoids (PAM) and was first proposed by Kaufman & Rousseeuw [57]. A description of the PAM algorithm is given in Algorithm 5.4.

Once PAM has been performed, the data points are categorised in K groups. A natural

question arising from this algorithm is the choice of K . For certain problems where it is known that a population has to be separated into K groups, the use of the algorithm is straight forward. However, when K is not known, we need to find it such that it maximises the overall distance between the medoids. This maximisation process is called *silhouetting*. The idea behind silhouetting (or the silhouette of the clustering algorithm) is to compute the average distances between clusters and points between clusters.

To find the optimum K , one computes the average distance $a(i)$ between a point i and all the other points belonging to the same cluster as the point i . For the same point i , then, one computes the average distance with the other clusters and one repeats this calculation for all available clusters. We set $b(i)$ to be the minimum average distance measured between all clusters. The cluster which has the smallest average distance, i.e., the smallest $b(i)$, is said to be the closest neighbour of point i (without including the cluster in which point i belongs to). We define the *silhouette* of point i as

$$s(i) := \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad , \quad -1 \leq s(i) \leq 1, \quad (2.9)$$

such that when $s(i)$ is close to 1, point i is well placed in its cluster and when $s(i)$ is close to -1 , point i should belong to its neighbouring cluster. Finally, when $s(i)$ is close to zero, point i lies at the border of both clusters.

By computing the silhouette of all points in the cluster and then averaging the silhouettes, it is possible to assign a *silhouette width* for the number of clusters, i.e., K . Therefore, one procedure to find the optimal number K of clusters is to maximise the silhouette width. A concrete example is shown in Figure 4.1.A.

Bayesian Networks

An important question in biology and more generally in Science is to determine causal relations between two or more variables. More than determining whether two or more variables are related to each other, it is important to see whether one variable acts on other

variables or vice versa. A possible answer to this question lies in the field of Bayesian networks (BNs). Bayesian networks are graphical models that capture the probabilistic nature of a given set of random variables of interest. Bayesian networks are acyclic graphs and therefore, can model causality between nodes (the variables). Here, we present the general concepts and definitions of Bayesian networks. For a full description and examples of usage, we refer the reader to Koller & Friedman [60], Nagarjan et al. [94] and Korb [61].

A Bayesian network is a directed acyclic graph $G = (\mathbf{V}, E)$ where the nodes v_i 's represent random variables $\mathbf{X} = (X_1, \dots, X_n)$ and where the edges (also called arcs) are the probabilistic dependencies between the nodes. To understand how Bayesian networks are created, in the next paragraphs, we define the concepts of graphical separation, path and blocked path. The definitions allow us to then define the notion of d-separation which is the key point in the theory of Bayesian networks. We first start by defining the graphical separation between two nodes or sets of nodes.

Definition 2.1.1. (*Graphical separation*) Two nodes or sets of nodes $\mathbf{A}$ and $\mathbf{B}$ in an acyclic graph $G = (\mathbf{V}, E)$ are *graphically separated* (denoted $\mathbf{A} \perp\!\!\!\perp_G \mathbf{B}$) if there is no arcs joining $\mathbf{A}$ and $\mathbf{B}$.

To understand the notion of causation in a graph, we then follow with the definition of parents and descendants of a node.

Definition 2.1.2. (*Parents of a node*) In an acyclic (and thus ordered) graph, the *parents* of a particular node $\mathbf{T}$, denoted $Par[\mathbf{T}]$, are the set of all nodes preceding node $\mathbf{T}$. The *descendants* of node $\mathbf{T}$ are the set of all nodes following node $\mathbf{T}$.

The relationship between the parents and descendant of nodes in a Bayesian network can be seen using the idea of path between nodes. Here, we define what is a path and a blocked path in a Bayesian network.

Definition 2.1.3. (*Undirected path and blocked path*) For two sets **A** and **B** of nodes in an acyclic graph $G = (\mathbf{V}, E)$, a *path* is any sequence of nodes between a node of **A** and a node of **B** such that all pairs of adjacent nodes to **A** and **B** are connected by an undirected arc and the nodes do not appear twice in the sequence.

A path is said to be *blocked* given a set of nodes **C**, if there is a node **T** in the path and at least one of the three following conditions holds:

- The node **T** is in **C** and **T** has one arc (in the path) entering it and one arc (in the path) exiting it.
- The node **T** is in **C** and **T** has both arcs (in the path) exiting it.
- Neither **T** nor its descendants are in **C**, and both arcs (in the path) enter **T**.

We follow with the identification of random variables with an acyclic graph. The probabilistic independence of two random variables (denoted here $\perp\!\!\!\perp_P$) can be related to the graphical separation between the same two variables (previously defined as $\perp\!\!\!\perp_G$) using the correspondence principle [103].

Definition 2.1.4. (*Correspondence Principle*) A graph $G = (\mathbf{V}, E)$ is an *independence map* (I-map) for the probability P (of the random variables of **X**) if there is a one-to-one correspondence between the random variables of **X** and the nodes of G such that

$$\mathbf{A} \perp\!\!\!\perp_P \mathbf{B} \mid \mathbf{C} \iff \mathbf{A} \perp\!\!\!\perp_G \mathbf{B} \mid \mathbf{C}$$

where **A**, **B** and **C** are disjoint sets of **V**. This must be read as: "there is an independence map if node **A** is not connected to node **B** (given the connectivity between node **B** and node **C**) and this relation implies a similar structure for the random variables, i.e., the variable **A** is independent of the variable **B** given the variable **C**".

Conversely, G is a *dependency map* (D-map) if

$$\mathbf{A} \perp\!\!\!\perp_P \mathbf{B} \mid \mathbf{C} \implies \mathbf{A} \perp\!\!\!\perp_G \mathbf{B} \mid \mathbf{C}.$$

Finally, G is a *perfect map* of P if

$$\mathbf{A} \perp\!\!\!\perp_P \mathbf{B} \mid \mathbf{C} \iff \mathbf{A} \perp\!\!\!\perp_G \mathbf{B} \mid \mathbf{C}.$$

For the later relation, P is also said to be isomorphic to G .

In order to find the correspondence between the probabilistic structure of G and the conditional independence, we employ the *directed separation* criterion (d-separation).

Definition 2.1.5. (*d-separation*) For a set $\mathbf{C}$ of nodes in an acyclic graph $G = (\mathbf{V}, E)$, $\mathbf{C}$ *d-separates* two other sets of nodes $\mathbf{A}$ and $\mathbf{B}$, if every path from a node in $\mathbf{A}$ to a node in $\mathbf{B}$ is blocked given $\mathbf{C}$.

Following Definition 2.1.5, it is possible to investigate the *Markov Property of Bayesian networks*. To understand this property it is important to recall the Markov property as stated in probability theory. In probability theory, the Markov property states that the probability of seeing the future of an event of a (specific) random variable only requires the knowledge of the present and not the sequence of events that occurred in the past. For Bayesian networks, the past, present and future are represented as the connection between nodes. If node $\mathbf{A}$ goes to node $\mathbf{B}$ which goes to node $\mathbf{C}$, then $\mathbf{B}$ can be seen as the present, $\mathbf{C}$ the future and $\mathbf{A}$ the past. With the d-separation, it is then possible to determine whether the nodes are d-separated, and using Definition 2.1.4 if there are probabilistic independences. Given an associated graph to the variable $\mathbf{X} = (X_1, \dots, X_n)$, the Markov property of Bayesian networks can provide a representation of the joint probability distribution of $\mathbf{X}$ as the product of the conditional distributions of the variables X_i 's. Thus, for discrete variables X_i , the distribution of $\mathbf{X}$ under the probability P is

$$P_{\mathbf{X}} = \prod_{i=1}^n P_{X_i}(X_i \mid \text{Par}[X_i]), \quad (2.10)$$

where $\text{Par}[X_i]$ is the set of all parents of X_i and P_{X_i} is the marginal distribution of the variable X_i . In the case of continuous random variables, the same applies and the distribution functions P_{X_i} are replaced by the density functions f_{X_i} .

After presenting the d-separation in Definition 2.1.5, we introduce the *fundamental connections* concept (Jensen [53]), illustrated in Figure 2.1. The fundamental connections are the three possible configurations connecting three nodes with two arcs. As we can see, the converging (v-structure) connection has incoming arcs from **A** and **B**, therefore violating the conditions of Definition 2.1.5. Thus, **A** and **B** are not independent given **C**. In this setting, we have

$$P(A, B, C) = P(C|A, B)P(A)P(B) \quad (2.11)$$

from the Markov property and because the parents of **A** and **B** are empty and $\text{Par}[\mathbf{C}] = \{A, B\}$.

In the serial and diverging configurations, the conditions of Definition 2.1.5 are satisfied. We see for the serial configuration that $\text{Par}[A] = \{\emptyset\}$, $\text{Par}[B] = \{C\}$ and $\text{Par}[C] = \{A\}$. Thus,

$$P(A, B, C) = P(B|C)P(C|A)P(A). \quad (2.12)$$

Similarly, for the diverging structure we have $\text{Par}[A] = \{C\}$, $\text{Par}[B] = \{C\}$ and $\text{Par}[C] = \{\emptyset\}$ and therefore

$$P(A, B, C) = P(A|C)P(B|C)P(C). \quad (2.13)$$

Using the fundamental connections, it is then possible to restate the idea of d-separation. Two sets of nodes A and B in a graph G are d-separated by a third set D (excluding the sets A and B) if and only if every path in between A and B is blocked by a node C (in D and not in A or B) such that:

- the connection through C is diverging or serial
- or the connection through C is converging and neither C nor its descendants have evidence (i.e., available quantitative information, also said to be "known").

The node A being d-separated from the node B means that C actually blocks the transmission of information between A and B in the case where C is the only (known) informative node in the graph between A and B . An example of d-separation is illustrated in Figure 2.1.B In Figure 2.1.A, the serial and diverging connections result in equivalent factorisations. Each factorisation can be obtained from the other by repeating Bayes' theorem.

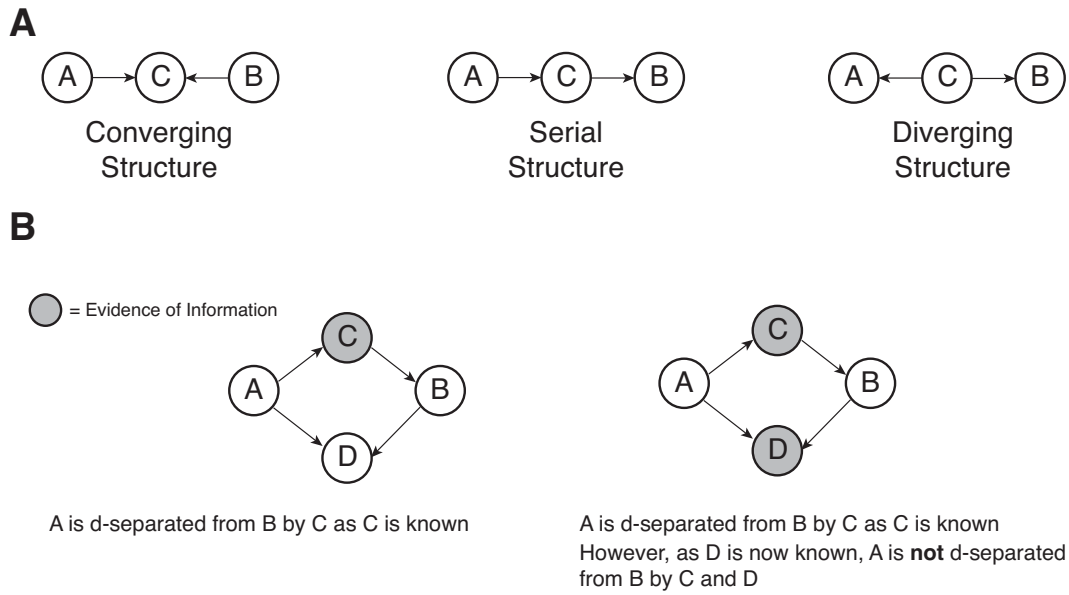


Figure 2.1: Fundamental connections in Bayesian networks and d-separation. A. There are three types of structures connecting three nodes with only two arcs. The converging structure, where two nodes A and B converge to the node C. The serial structure where the arcs flow from A to C and then from C to B and finally, the divergent structure where the arcs go from C to A and B. In the second panel B., the d-separation is explained with a visual example. The known information about a node having evidence allows us to determine if two nodes are d-separated.

Such equivalent structures in Bayesian networks are called *Markov equivalent* structures. By the nature of equivalence relations, the set of equivalent structures in a graph forms an equivalence class. Equivalence class are often represented by a completed partially directed acyclic graph (CPDAG) where the arcs belonging to v-structures and the arcs introducing new v-structures (or cycles) are directed. A simple illustration of the concept of equivalence class is given by Chickering [20] who cites the theorem stated by Verma & Pearl [140]. The theorem says that two directed acyclic graphs are equivalent if and only if they have the same skeletons and the same v-structure. Here, the skeleton refers to the undirected graph that results from ignoring the directionality of all edges in the graph. Additionally, a v-structure in a directed acyclic graph is an ordered triplet (A, B, C) of nodes having the connections: $A \rightarrow B$, $B \leftarrow C$ and the nodes A and C are not adjacent in the graph. Equivalence classes in Bayesian networks are usually found by performing a greedy search on all possible networks.

As a machine learning framework, Bayesian networks learn from the data given to them and then construct the structure of the networks. This *structure learning* step learns the network by finding the minimal independence map of it, i.e., by finding a directed acyclic graph G which is an independence map of a probability P and for any subgraph G' of G , G' is not an independence map of P . A number of algorithms are available in the literature and are described in [94]. The algorithms are separated into three categories. First, constraint-based structure learning algorithms learn the structure of the network by testing the conditional independence of the variables. Second, score-based learning algorithms assign a score to the network and try, via optimisation technique, to maximise the score. Finally, the last class of algorithms represents the hybrid structure learning algorithms. This last class combines the score-based and constraint-based algorithms to produce reliable networks. Among this class of algorithms, the Max-Min Hill-Climbing Algorithm (MMHC, Tsamardinos et al. [136]) is one of the most accurate.

Here, we will not describe the entire process beneath the MMHC algorithm. However, it must be noted that the algorithm performs following two steps. The first step "restricts" the candidate set of variables representing the parents of the node X_i to a subset of nodes whose behaviour has been related to X_i . The second step "maximises" the score of the network according to the constraints applied to the parents in the first step. To compute the score, based on the structure of the network, the MMHC algorithm performs conditional tests on the whole set $\mathbf{V}$ of nodes. To this end, the algorithm assesses the strength and the conditional independence of two variables by computing a G^2 statistic. The G^2 statistics is a likelihood-ratio used to replace χ^2 statistics. It can be shown that the approximation of the χ^2 statistics to the G^2 statistic is obtained by taking a Taylor expansion of the natural logarithm around 1 in the χ^2 statistic.

The general formula for G^2 is

$$G^2 = 2 \sum_i Obs_i \log \left(\frac{Obs_i}{Exp_i} \right), \quad (2.14)$$

where Obs_i is the observed frequency of the event i and Exp_i is the expected frequency of

event i . The reader is invited to read McDonald [89] for more information and use of the G^2 statistic.

Following the notation of Spirtes et al. [126], for a set of random variables $\mathbf{X} = (X_1, \dots, X_n)$, let S_{ijk}^{abc} be the number of times one sees: $X_i = a$, $X_j = b$ and $X_k = c$, in the data. They also define in a similar manner, the number S_{ik}^{ac} , S_{jk}^{bc} and S_k^c . The G^2 statistic is then calculated ([126, 95]) as

$$G^2 = 2 \sum_{a,b,c} S_{ijk}^{abc} \log \left(\frac{S_{ijk}^{abc} S_k^c}{S_{ik}^{ac} S_{jk}^{bc}} \right). \quad (2.15)$$

It can be shown that the G^2 statistic is asymptotically identical to a χ^2 statistic under the null hypothesis that the observed frequencies of events are the result of random sampling from distributions with the expected frequencies of events. In this setting, it is then possible to define a p-value α and tune the MMHC algorithm to retrieve statistically significant network structures.

2.2 Biological Prerequisites

In this section, a brief introduction to molecular biology is presented. Firstly, we introduce the connection between DNA, genes and phenotypic variability. Secondly, as this thesis focuses on variability during the growth and division of unicellular organisms, we present the cell cycle of prokaryotes and eukaryotes. We recommend the books of Lodish et al. [78] and Alberts et al. [3] for a complete description of the processes involved in molecular biology and the book of Phillips et al. [104] for an introduction to physical biology. Finally, for the readers willing to learn and apply key concepts of biology, the book of Waite & Waite [142] and Cohen [21].

2.2.1 From DNA to Phenotypic Variability

All living organisms are genetically described by a chain of nucleic acids called *deoxyribonucleic acid* (DNA) or *ribonucleic acid* (RNA). The entire machinery of living organisms is based on the correct utilisation of these chains. More specifically, in molecular biology the *central dogma* stated by Francis Crick [28] says that *genes*, which are sequences of

DNA, are processed in two steps to create *proteins*. The first step (called *transcription*) copies the gene into a *messenger RNA* chain. The second step (called *translation*) brings the mRNA to a ribosome (a specialised protein/RNA complex), which operates protein synthesis, and then transforms the mRNA into a protein. Proteins are essential in all living organisms as they provide most of the cell machinery components and are involved in cellular functions. Proteins can also join and form protein complexes. For instance, *actin* is a protein that produces filaments controlling different structural aspects of the cell, in particular related to cell division and cell shape.

The cell shape and more generally the cell structure can be differentiated according to the presence of certain *organelles* within the cell. Organelles are subunits of the cells which perform specific functions for the cell. An example of such an organelle is the *nucleus*. The nucleus is a *membrane-enclosed* organelle, i.e., it has a membrane, that contains the genetic material (DNA) of the cell. Organisms having a nucleus are called *eukaryotes*. On the contrary, organisms such as bacteria and lacking a cell-membraned nucleus are called *prokaryotes*. The discernment between organisms having a membrane-enclosed nucleus or not is fundamental in biology to classify organisms.

An example of a eukaryote is yeast. Yeast is *unicellular*, which means that it is an organism only consisting in one cell. Yeast normally reproduces asexually, i.e., the organism can produce an offspring by itself. However, it has mating types and can take part in mating under specific conditions. In this thesis, we study two species of yeast: *Saccharomyces cerevisiae* (also denoted *S. cerevisiae* or budding yeast) and *Schizosaccharomyces pombe* (denoted *S. pombe*). These organisms operate *cytokinesis*, the process of dividing the *cytoplasm*, differently. *S. cerevisiae* has an *asymmetric division*, meaning that the *mother cell* is larger than the *daughter cell* when the cell divides. In contrast, *S. pombe* operates a *binary fission*, which produces two identical (in size) cells at division. Therefore, their *phenotypes*, i.e., their observable features, are different.

A well studied example of prokaryotes are bacteria. Here, we study the bacterium *Escherichia coli* (denoted *E. coli*). Bacterial cells such as *E. coli* are unicellular cells that operate a binary fission.

Phenotypes of organisms vary greatly depending on the type of (eukaryotic/prokaryotic) cells. It is important to distinguish between molecular and morphological phenotypes. Molecular phenotypes are, for example, the level of gene expression and morphological phenotypes are related to the cell morphology, such as the volume of the cell. Therefore, it is important to differentiate phenotypes by their organisms to which they belong to and their level of description.

Once an organism has a defined set of phenotypes, it is possible to observe phenotypic changes by altering the genome of the organism. To this end, biologists often employ *gene deletion* which consists in removing one or multiple genes. When a gene is deleted, it is common practice to rename the organism by writing its name followed by Δ . To signify that one or multiple genes have been deleted, biologists employ the term: "gene knock-out". It must be noted that usually, gene deletion is performed on *non-essential* genes, to avoid triggering cell death. Genetically modified organisms are usually called *mutants*, to distinguish their modified nature from the organisms that have not been modified, which are called *wild types*.

Wild types and mutants allow to perform empirical analyses when a gene deletion is performed. By comparing the phenotypes of the wild types against the phenotypes of the mutants, biologists can draw conclusions on the nature of the deleted gene. As a common procedure and for statistical reasons, the same gene is deleted on multiple organisms of the same strain to provide the analysis with an acceptable amount of statistical samples. Therefore, each gene deletion can be assimilated to a *phenotypic* distribution of changes by looking at the statistical distributions of samples recorded from the gene deletion.

In general, phenotypic distributions, like any probability distributions, possess a mean and a variance. In other words, phenotypic distributions have variability that can be related to the gene deletions. Variability of the distribution depends on what type of phenotype is regarded. At both the molecular and morphological level, variability arises.

At the molecular level, there are two types of variability: intrinsic and extrinsic. In-

trinsic variability represents the stochastic nature of gene expression. Transcription and translation occur in a fluctuating environment. Transcription and translation are noisy in the sense that they are not performed perfectly and therefore they show variability. Extrinsic variability is related to the variation occasioned by the other components within the cells.

Variability in morphological phenotypes relates to the variations seen amongst similar cells for a specific trait of the cells morphology. For example, the nucleus size slightly varies from one cell to another as genetic perturbations or environmental changes can occur.

Some phenotypes, when observed, can show more or less variability in their distributions. We denote this variability by "*non-genetic phenotypic variability*". In this thesis, we are interested in understanding phenotypic variability by analysing the genes responsible in high and low variability. Moreover, we investigate the effect of phenotypic variability on the cell cycle, and we look at how phenotypic variability affects the cell cycle of some microorganisms.

2.2.2 Cell Cycle

The growth of the cells and consequently their division needs to be investigated in two different manners, according to the type (eukaryotic/prokaryotic) of cells. The process in which the cell will grow and divide is called the *cell cycle*. The cell cycle is controlled by a periodic chain of molecular events in the cell. Cell cycle regulation in eukaryotic cells differs from prokaryotic cells as the latter lack a nucleus.

In prokaryotes, the cell cycle is composed of two states: the *interphase* and the *cell division*. Each cell cycle state can have one or multiple phases, depending on the nature of the molecular events. The interphase is separated into three phases. Firstly, the *Gap1* (G1) phase which ensures that the DNA content is ready to be replicated. Secondly, the *synthesis* (S) phase in which the DNA is replicated. Finally, as the cell continues growing, the *Gap2*

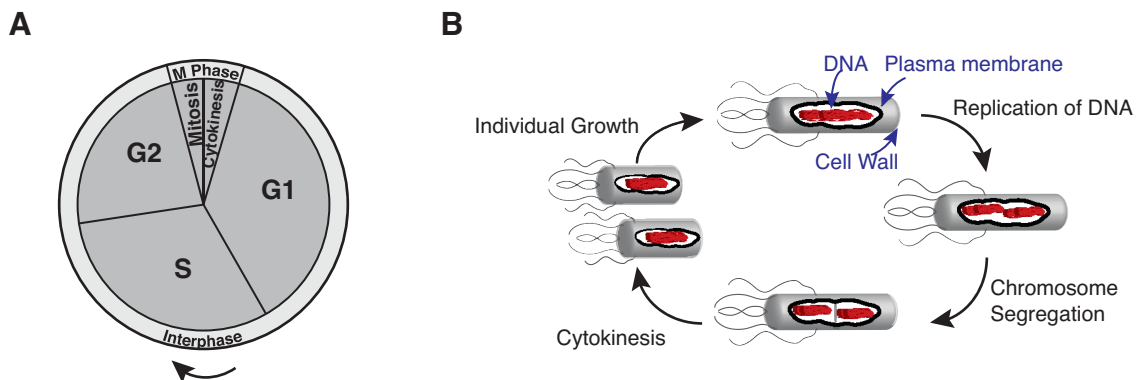


Figure 2.2: The growth and division of unicellular organisms operates differently depending on the type, either eukaryotic or prokaryotic cells. A. Prokaryotes go through a cell cycle composed of three phases where the DNA is replicated and then transmitted to the daughter cell after division (mitosis). B. Eukaryotes process a binary fission where DNA is replicated to allow a segregation of the chromosomes which are then transferred to the daughter cell, as illustrated in panel B.

(G2) phase verifies that the DNA has been replicated and that the cell is ready to divide. Then the cell division also called *mitosis* or *M* phase occurs and the cell stops growing to divide and produce two *daughters*. The complete prokaryotic cell cycle is illustrated in Figure 2.2.A.

In eukaryotes, the cell cycle follows a binary fission where the DNA and more especially the *chromosomes* are replicated and then attached to both sides of the *cell membrane*, i.e., the envelope of the cell. Once the cell starts dividing, the original set of chromosome and its replicate are separated and then attributed to the daughter cells separately. The complete binary fission process is illustrated in Figure 2.2.B.

Chapter 3

Modelling the dynamic transition from cellular quiescence to proliferation in *S. pombe*

3.1 Introduction

In cell biology, quiescence is a resting state where the cells have stopped growing. In particular experimental settings, it is possible to induce and stop quiescence. In their work, Sjiki et al. [118] studied quiescence in *Schyzosaccharomyces pombe* by depleting nitrogen from the medium and forcing the cells to divide without growing. More specifically, under nitrogen depletion, the cell cycle of fission yeast is perturbed and forces the cell to divide without growing. This process causes the cell to divide twice until it reaches its minimum size. At this stage, when no additional divisions are possible, the cell enters a dormant phase where no division and no growth occur. After having spent a certain amount of time in quiescence, nitrogen is re-introduced in the medium to allow the cell to start growing and proliferating again. The whole process is illustrated in the schematic of Figure 3.1.

Under nitrogen starvation, it is unclear how the cell readapts its cell cycle (Laporte et al. [69]) and how much time it needs to wait until it starts growing and dividing again. Indeed, at the end of the dormant phase, when cells leave quiescence, they spend a certain

amount of time, called "lag-time", before fully proliferating. Hence, the aim of our analysis is to provide a framework of mathematical models capable of reproducing the effect of a lag-time based on experimental data. The data was collected at the Bähler lab at the University College London. Around 2000 mutants were screened using a bar-seq experiment. The bar-seq, or "Barcode analysis by Sequencing", is a novel type of experiment allowing geneticists to identify genes by attributing barcodes to them [45]. Hence, after a first bar-seq screen, the set of knockout genes was reduced to 19 candidates potentially affecting the lag-time, as they diverged from the wild-type trend.

To comprehend the mechanism behind the lag-time and determine why certain mutants have a longer lag-time than others, we model the transition from quiescence to proliferation in two distinct ways. The first model considers the lag-time at the single cell level as a random time which cells undergo before they start growing and dividing again. Here the main ingredients of the stochastic process are the average lag-time and its standard deviation as first stated by Baranyi [9, 8]. In addition to the average lag-time and its standard deviation, we affix the elongation rate of a single cell to the process to also investigate whether the lag-time has an effect on the elongation rate of the cell. It is unclear if the lag-time can be reduced by the cells and the entire population and if the cells can sense the lag-time and adapt their elongation rate. Thus, this model allows us to describe the lag-time of an entire population by considering the cells individually and describe morphological properties of them. As an amelioration of the model presented by Baranyi, which was purely theoretical, we fit the parameters of the model to real experimental data to demonstrate the accuracy of the model. Secondly, we test a broader hypothesis that instead of having all cells growing after a random amount of time, some of the cells become non-growing or dormant [62, 40, 124]. In this setting, the model consists of computing the proliferation curve (or also called growth curve) by fitting an exponential to it, subject to a parameter representing the fraction of growing cells.

This chapter is organised in four sections: the first section explains the biological experiments and the format of the data that we employ in our models. The second section introduces the two mathematical models and their application on the biological data. The

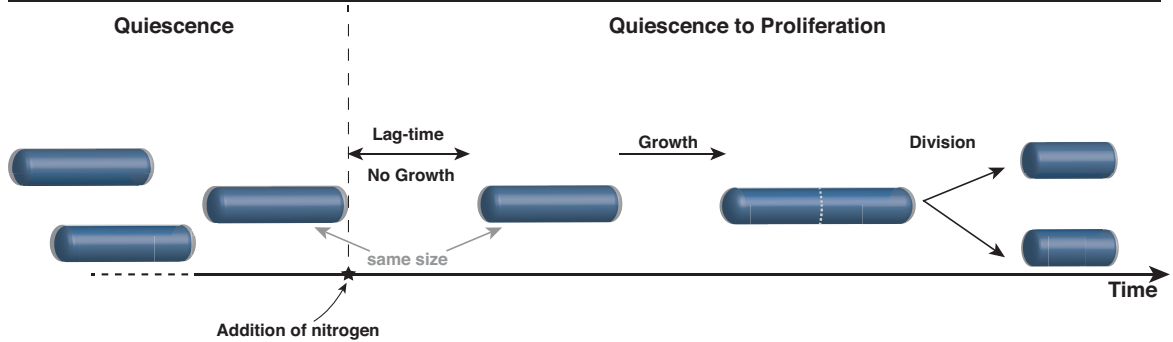


Figure 3.1: Schematic of the growth and division process with a lag-time. In a first time, the cell does not grow until the lag-time is over. Thereafter, the cell starts growing and finally divides to produce two daughters.

third section presents the results of the two models on the biological data. Finally, the last section concludes our analysis and offers a follow-up experiment.

3.2 Biological Experiments and Acquisition of the Growth Curves

Here, the growth curves for the transition from quiescence to proliferation are recorded as optical densities (ODs) of the population mass. The optical densities are measured with a BioLector® for the 19 mutants selected at the Bähler lab. In the interest of understanding the mechanisms behind the lag-time, the mutants are put in quiescence for different amount of time (in days). In addition, and to set a control group, some strains are grown in a standard fission yeast minimal media (EMM) without entering the quiescent phase. The experiments are performed for two temperatures: 25 and 32 degrees Celsius, to investigate if the temperature can also be a factor modifying the transition. As soon as the cells leave quiescence, a time series of the optical density is recorded for each strain until the whole population reaches its steady state. The growth curves for each mutant are measured three times for the same batch. Such repetitions of the measurements are called "technical replicates". Moreover, each batch is also grown three times to provide more "biological replicates".

Due to errors in the image acquisition process, the optical mass (density) of the system sometimes decreases at the beginning of the time series. This phenomenon thus affects

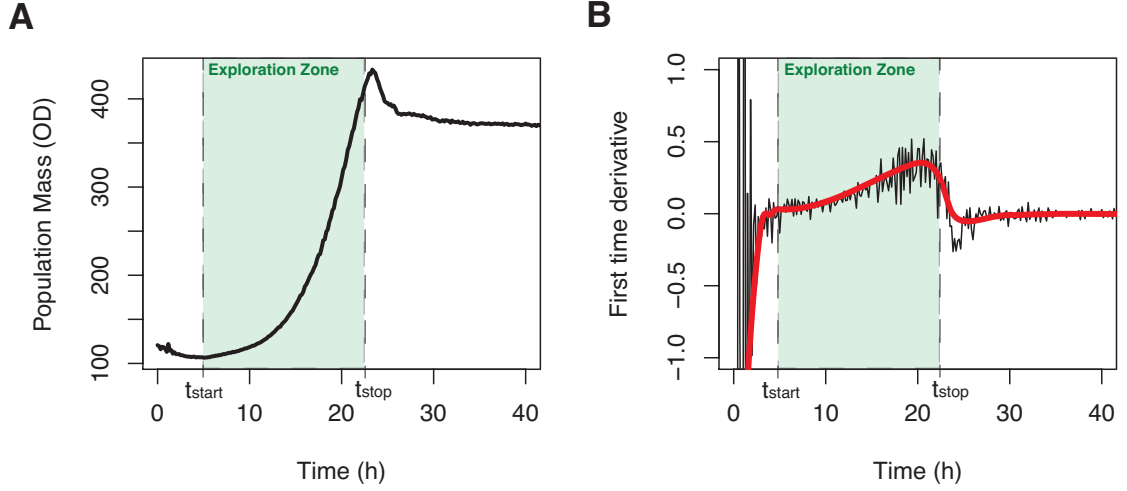


Figure 3.2: Typical growth curve for a population of cells leaving quiescence and entering the exponential growth. The exploration zone, shown in green, is delimited by the starting and finishing time points of the growth curve. (A) Due to errors in the image acquisition process, the growth curve often decreases during the course of its growth and then starts growing exponentially. (B) The first time derivative of the growth curve is noisy due to the low number of time points used to compute the derivative numerically. The thick red curve indicates a Lowess approximation used to find the extremums of the growth curve.

the modelling of the transition as the trajectory should always be constant for some time (the lag-time) and then monotonic increasing. An efficient method to fix the decrease of mass is to consider the time series only where the population effectively starts growing exponentially. It is then necessary to modify the time series and set the optical density to its minimum before the population starts growing exponentially (see Figure 3.2.A). By "shrinking" the time series we might induce a bias in the final analysis as we do not consider the original optical densities. However, as we apply the shrinking in both models, we do not think that this will affect the final comparison. If both models are analysed using the same (possibly biased) optical densities, this should not affect the comparison between the models.

We numerically find the "starting" and "finishing" time points of the trajectory by calculating

$$\begin{aligned} t_{\text{start}} &:= \arg \min_t \left\{ \frac{d\bar{s}(t)}{dt} = 0 \right\}, \\ t_{\text{stop}} &:= \arg \min_t \left\{ \frac{d\bar{s}(t)}{dt} < 0 \mid t > t_{\text{start}} \right\}, \end{aligned} \quad (3.1)$$

where $\bar{s}(t)$ is the mass of the population at time t . As shown in Figure 3.2.B, the first time derivative of the trajectory is very noisy. Therefore, finding its null points can be numerically challenging. To smooth the first time derivative of the $\bar{s}(t)$ process, we fit the first time derivative of the process with a Lowess regression, to find the zeros of $d\bar{s}(t)/dt$. Moreover, to ensure a perfect matching of the starting and finishing points, we visually inspect the starting and finishing points by hand.

We recall that the experiment consists in measuring mutants and wild types at two different temperatures: 25 and 32 degrees Celsius. For the 25 degrees Celsius, the experiment only considered 5 mutants and one wild type. For the 32 degrees Celsius, the experiment considered 18 mutants and one wild type. The growth curves of each experiment (as shown in Figure 3.2.A) were measured 9 times in total, i.e., for three biological replicates times three technical replicates.

Once the growth curves have been collected for each mutant, we then fit our models to the growth curves to identify the causes of the transition to proliferation.

3.3 Models

In this analysis, we model the lag-time either at the single cell level or the population level. Firstly, by considering the lag-time as a random amount of time the cells need to wait before entering their cell cycle again and secondly, by assuming that cells might have died or completely stopped their cell cycle. The first part introduces the stochastic lag and its properties and the second part presents the model fitting the growth curves with an exponential, subject to solely a fraction of the growing population.

3.3.1 Stochastic Lag-time

In this part, we consider proliferating cells with a lag-time at the beginning of their cell cycle. Each cell waits a certain amount of time, the lag-time, which is considered here as a random variable. We model the elongation of the cells by a stochastic process where the lag-time follows a probability distribution. Due to the stochastic nature of the process, we compute the average mass of the population to analyse the asymptotic behaviour.

We recall that in this particular experiment, the cells divided without growing, until they reach a minimum mass where no further divisions were possible. The stochastic process then associated with the model is based on the time the cell waits before starting to grow. Here, for a particular cell i , we model its length with the following time-dependent process:

$$L_i(t) \quad , \quad i = 1, \dots, n_0 \quad (3.2)$$

where n_0 is the total number of cells initially present in the population.

As a first assumption, we define the lag-time as a random vector τ following the probability distribution f_{μ, σ^2} , i.e.,

$$\tau = (\tau_1, \dots, \tau_{n_0}) \quad , \quad \tau_i \sim f_{\mu, \sigma^2} \quad , \quad i = 1, \dots, n_0, \quad (3.3)$$

where the parameters μ and σ^2 are the average lag-time and the variance of the lag-time respectively. In addition, under the hypothesis that all cells possess the same initial length l_0 (variability in the birth and division lengths is not considered here), we formally describe the length of a particular cell i at time t as

$$L_i(t) = \begin{cases} l_0 & , \quad t \in [0, \tau_i] \\ l_0 e^{\lambda(t-\tau_i)} & , \quad t > \tau_i \end{cases} \quad , \quad i = 1, \dots, n_0 \quad (3.4)$$

where the elongation of $L_i(t)$ is exponential with an elongation rate λ . Although a recent analysis suggests that the elongation of *S. pombe* can be separated in two as the population first grows exponentially and then linearly (see Nobs & Maerkl [98]), for simplicity we model the elongation as a pure exponential growth as proposed by Cooper [26].

The total mass of the system at time t is then defined as the sum of all lengths:

$$s(t, \mu, \sigma^2, \lambda) := \sum_{i=1}^{n_0} L_i(t). \quad (3.5)$$

Here, we are interested in the asymptotic behaviour of the total mass of the population. Hence, we derive the average total mass of the system $\bar{s}(t, \mu, \sigma^2, \lambda)$ by calculating

$$\begin{aligned} \bar{s}(t, \mu, \sigma^2, \lambda) &:= \mathbb{E}[s(t, \mu, \sigma^2, \lambda)] = \mathbb{E} \left[\sum_{i=1}^{n_0} L_i(t) \right] \\ &\stackrel{iid}{=} \sum_{i=1}^{n_0} \mathbb{E}[L_i(t)] \\ &= n_0 \mathbb{E}[L_1(t)] \\ &= n_0 \mathbb{E} [l_0 \mathbb{1}_{\{t \leq \tau_1\}} + l_0 e^{\lambda(t-\tau_1)} \mathbb{1}_{\{t > \tau_1\}}] \\ &= n_0 (\mathbb{E} [l_0 \mathbb{1}_{\{t \leq \tau_1\}}] + \mathbb{E} [l_0 e^{\lambda(t-\tau_1)} \mathbb{1}_{\{t > \tau_1\}}]) \\ &= n_0 (l_0 \mathbb{P}(\tau_1 \geq t) + l_0 \mathbb{E} [e^{\lambda(t-\tau_1)} \mathbb{1}_{\{t > \tau_1\}}]) \\ &= n_0 l_0 \left(\int_t^\infty f_{\mu, \sigma^2}(s) ds + \int_0^t e^{\lambda(t-s)} f_{\mu, \sigma^2}(s) ds \right). \end{aligned} \quad (3.6)$$

where $\mathbb{1}$ is the indicator function.

In his original paper, Baranyi [9], does not consider the nature of the distribution f_{μ, σ^2} . Thus, it must be noted that when the density function f_{μ, σ^2} follows a normal distribution, f_{μ, σ^2} has to be re-scaled in order to integrate to 1 when $t = 0$, as the tail of the distribution might cross the zero lag-time value.

Using $\bar{s}(t, \mu, \sigma^2, \lambda)$ to fit the growth curves, the parameters that need to be searched for are: the initial number n_0 of cells in the population, the initial length l_0 of the cell, the average lag-time μ of the cells, the variance σ^2 of the lag-time, and the elongation rate λ of the cells.

The total number of cells n_0 and the initial length l_0 can be deduced from the experimental data. Hence, the initial length of the cells is the ratio of the real total mass of the system, denoted by $s_{\text{exp}}(t)$ and found experimentally, at time $t = 0$ to the fixed initial number n_0 of

cells. In other words,

$$l_0 = \frac{s_{\text{exp}}(0)}{n_0}. \quad (3.7)$$

Moreover, and for simplicity, we normalise the time series to obtain

$$n_0 l_0 = 1. \quad (3.8)$$

This ansatz reduces the problem to three unknown parameters: μ , σ^2 and λ .

The averaged trajectory of a single strain, computed from the biological and technical replicates, is then normalised with the minimum mass, to start the process with a unit mass and thus agree with Eq.(3.8).

For each triplet of parameters in

$$\{(\mu, \sigma^2, \lambda) \mid \mu > 0, \sigma^2 > 0, \lambda > 0\}, \quad (3.9)$$

we run an optimisation algorithm on the error function

$$e(\mu, \sigma^2, \lambda) := \sum_{t \in T} \left(\frac{\bar{s}_{\text{exp}}(t) - \bar{s}(t, \mu, \sigma^2, \lambda)}{\sigma_{\text{exp}}(t)} \right)^2, \quad (3.10)$$

where T is the discrete time interval of the real data, $\bar{s}_{\text{exp}}(t)$ is the average trajectory over multiple biological and technical replicates, $\sigma_{\text{exp}}(t)$ is the standard deviation of the replicates and $\bar{s}(t, \mu, \sigma^2, \lambda)$ is the theoretical averaged trajectory as defined in Eq. (3.6).

The average lag-time, its variance, and the elongation rate of each mutant are then found by minimising the quadratic standardised error function in Eq.(3.10). The optimisation algorithm and the parameters obtained here are explained in the results section.

3.3.2 Fractional Growth of the Population

In this part, we study an alternative approach to the problem by considering a fraction of cells that may have stopped their cell cycle. Therefore, the population consists of a fraction of cells not growing, while the remaining part can grow and divide.

The growth being recorded as an optical density, we measure the number of cells in terms of fraction. Hence, we define the size of the population, denoted by $N(t)$, as:

$$N(t) = N(t, f, \lambda) := f \exp\{\lambda t\} + (1 - f) \quad (3.11)$$

where f is the fraction of cells growing, $1 - f$ is the remaining fraction of cells not growing, λ is the growth rate of the population, and t is the time.

Here, the parameters to fit and to adjust are: the fraction of cells f and the growth rate λ , i.e.,

$$\{(f, \lambda) \mid 0 \leq f \leq 1, \lambda > 0\}.$$

To find the optimal f and λ , we define the quadratic error function

$$e(f, \lambda) := \sum_{t \in T} \left(\frac{\bar{s}_{\text{exp}}(t) - N(t, f, \lambda)}{\sigma_{\text{exp}}(t)} \right)^2, \quad (3.12)$$

where T is the discrete time interval of the real data and $\bar{s}_{\text{exp}}(t)$ is the average trajectory over multiple biological and technical replicates, and $\sigma_{\text{exp}}(t)$ is the standard deviation of the replicates.

3.4 Results

We compared both models by using the Akaike Information Criterion (Aho et al. [2]) to distinguish and separate the models by their number of parameters. The AIC score is formally defined as:

$$\text{AIC} = 2k - 2 \log(L) \quad (3.13)$$

where k is the number of parameters used in the model and L is a maximised value of the likelihood function of the model. Assuming that the residuals coming from the fit of the

models are normally distributed, it can be shown that the log-likelihood is a function of the residual sum of squares. Hence, the AIC becomes (see Appendix A):

$$\text{AIC} = 2k + C + \sum_{i=1}^n \frac{(y_i - f(x_i))^2}{\hat{\sigma}_i^2} \quad (3.14)$$

where n is the number of data points, the y_i 's are the data, $f(x_i)$ is the fit of the model and C is a constant, usually set to zero.

The residuals obtained after fitting our models to the growth curves showed skewed normal distribution (see Appendix A). In our analysis, we still decided to use the residuals to compute the Akaike Information Criterion, as in both models we see the same skewness in the QQ-Plots in the appendix.

We recall that the Stochastic Induced Lag Model takes 3 parameters (μ , σ^2 and λ) as input and the Fractional Growth of the Population takes 2 parameters (f and λ). For each model, we look at the median and standard errors of the AIC scores amongst the mutants and wild types, to obtain the comparisons between the two models, as shown in Table 3.1. The results in Table 3.1 indicate that both models are close to each other as their AIC score differ by a factor of two.

Model	Number of Parameters	Median AIC	Standard Error AIC
Stochastic Lag	3	6.0026	0.25
Fractional Growth	2	4.012077	0

Table 3.1: Akaike Information Criterion for the stochastic lag and the fractional growth of the population. The addition of a third parameter for the stochastic lag model does not provide more information compared to the fractional growth of the population.

For the stochastic lag model, we analyse both the 25 and 32 degrees Celsius fits by first comparing the scatter plots of the average lag-time, the coefficient of variation (labelled "cv" and calculated as the standard deviation divided by the mean lag-time), and the elongation rate of the cells and second by looking at the residuals of a linear regression computed on the aforementioned plots. Due to the low number of measurements for the 25 degrees experiment, we only analyse the 32 degrees experiment for the fractional growth model. This section is separated in two: the first part gives the extraction of the mutants

being outliers according to the stochastic lag model and the second part extracts the outliers for the fractional growth model.

3.4.1 Stochastic Lag

To fit the average lag-time μ and its variance σ^2 to the biological and technical replicates with the stochastic process, we assume a gaussian distribution for τ , i.e.,

$$f_{\mu, \sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right). \quad (3.15)$$

To minimise the quadratic error function $e(\mu, \sigma^2, \lambda)$ we employ two minimisation techniques: a simulated annealing and a gradient descent L-BFGS-B (Byrd et al. [15]). Both techniques showed the same results; however, the L-BFGS-B algorithm performed faster. Therefore, here, we only use the L-BFGS-B algorithm to fit both models. A typical trajectory found by the algorithm is shown in Figure 3.3.

We separated the analysis in two, as some of the wild types and mutants were measured

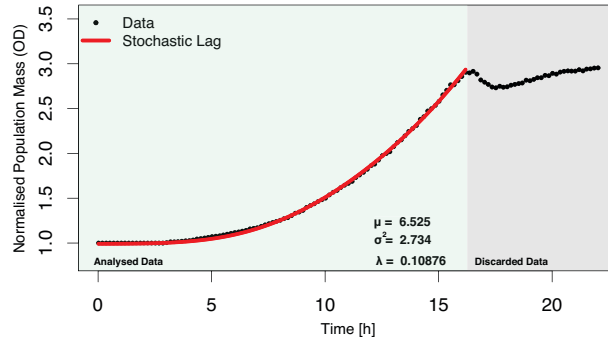


Figure 3.3: Trajectory produced by the stochastic algorithm for a wild type which went under quiescence for one day at 32 degrees Celsius. The green area represents the data points (from the real experiment) used to estimate the lag time. The grey area represents the discarded data points. The thick red line is the approximated trajectory by the stochastic algorithm. The average lag-time is μ , the standard deviation of the lag-time is σ^2 , and the elongation rate is λ .

at different temperatures; at 32 degrees Celsius and at 25 degrees Celsius. The 25 degrees Celsius wild types and mutants were underrepresented in terms of number of measurements (only 6 for the WT and 20 for the mutants) so we only show their results in Figure

3.4 and Figure 3.5. The analysis of the wild types and mutants grown at 32 degrees Celsius is shown in Figure 3.6 and Figure 3.7.

Wild types and Mutants at 25 Degrees Celsius

The wild types indicate that the longer the number of days in quiescence, the longer the average lag-time, as seen in Figure 3.4.A. Similarly, in panel B, one can see that the longer the number of days in quiescence, the smaller the coefficient of variation. Thus, by strictly comparing the mutants and the wild-type trends (Figure 3.5), one can first see that the wild types show strong coefficients of correlation in each setting and second, by strictly looking at the linear trends, it is possible to determine the mutants diverging from the wild-type trend. When comparing the average lag-time with the elongation rate (Figure 3.5.B)

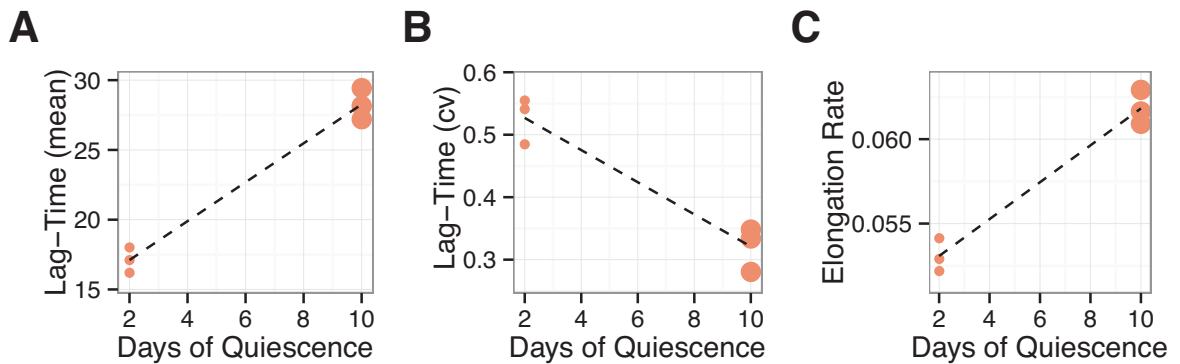


Figure 3.4: Parameters found from the fits for the wild types at 25 degrees Celsius. The black dashed line is the regression line amongst all measurements. The plots represent (A) the average lag-time against the number of days in quiescence (B) the coefficient of variation of the lag-time against the number of days in quiescence and (C) the elongation rate against the number of days in quiescence. Here and for the rest of this chapter, the units for the mean lag-time and the elongation rate are in days and optical density per day respectively. The coefficient of variation (cv) has no unit.

coming from the fits, we see that four mutants (*spf1* Δ , *stc1* Δ , *pyp2* Δ , and *pst2* Δ), which are measured after 10 days of quiescence, are possible outliers. The same observation applies when comparing the coefficient of variation of the lag-time with the elongation rate (Figure 3.5.C). The aforementioned mutants do not follow the wild-type trend. A visual inspection of the number of days in quiescence for the average lag-time (Figure 3.5.D)

also shows that the four mutants are outliers. We validate these observations by looking at the residuals of the mutants calculated using the wild-type trends (see Appendix A.1). However, a deeper inspection of the distribution of the lag-time, cv and elongation would be necessary to confirm that the four mutants are outliers. Unfortunately, the number of measurements for the 25 degrees Celsius wild types (6 in total) and mutants (19 in total) is not enough to perform such analysis.

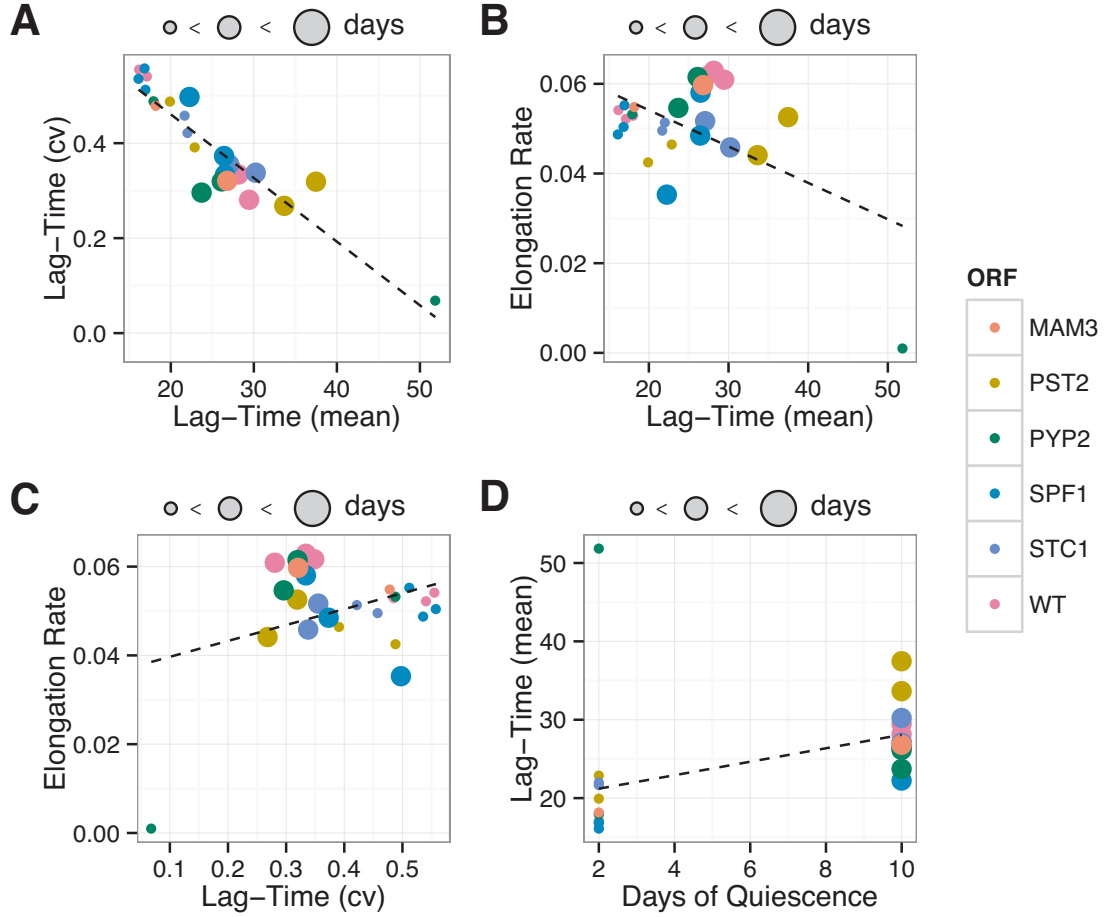


Figure 3.5: Representation of the wild types and mutants at 25 degrees Celsius. The larger the symbol, the longer the number of days in quiescence. The black dashed line is the regression line amongst all measurements. (A) Coefficient of variation against the average lag-time. The Pearson coefficients of correlation for the wild types and mutants are respectively $r = -0.992$, $P < 0.001$, and $r = -0.895$, $P < 0.001$. (B) Elongation rate against the average lag-time. The Pearson coefficients of correlation for the wild types and mutants are respectively $r = 0.945$, $P < 0.001$, and $r = -0.009$, $P < 0.001$. (C) The elongation rate against the coefficient of variation. The Pearson coefficients of correlation for the wild types and mutants are respectively $r = -0.921$, $P < 0.001$, and $r = -0.204$, $P < 0.001$. (D) The average lag-time against the number of days in quiescence. The Pearson coefficients of correlation for the wild types and mutants are respectively $r = 0.986$, $P < 0.001$, and $r = 0.801$, $P < 0.001$.

Wild types and Mutants at 32 Degrees Celsius

For the wild types and mutants measured at 32 degrees Celsius, we perform the same analysis as for the 25 degrees Celsius. As a first comment, one can see that the longer the days in quiescence, the longer the average lag-time (Figure 3.6.A). Similarly to the 25 degrees Celsius wild-type measurements, the 32 degrees Celsius wild-type measurements indicate that the correlations found previously hold here. Therefore, one can again compare the wild types and the mutants to see what the trends are and which mutants behave as outliers.

When compared to the elongation rate (Figure 3.7.B), the average lag-time of the wild types clearly shows a linear trend. Here, we isolate four mutants: *set1* Δ , *swd1* Δ , *spf1* Δ , and *rhp6* Δ as outliers. The same mutants are as well outliers in the other figures (Figure 3.7.A-C&D) where the mutants seem to group together, indicating a similar general behaviour. We notice that the mutant *spf1* Δ , found previously, is again an outlier here. Again, here we validate our observations by looking at the residuals of the mutants calculated using the wild-type trends (see Figure A.2 in appendix).

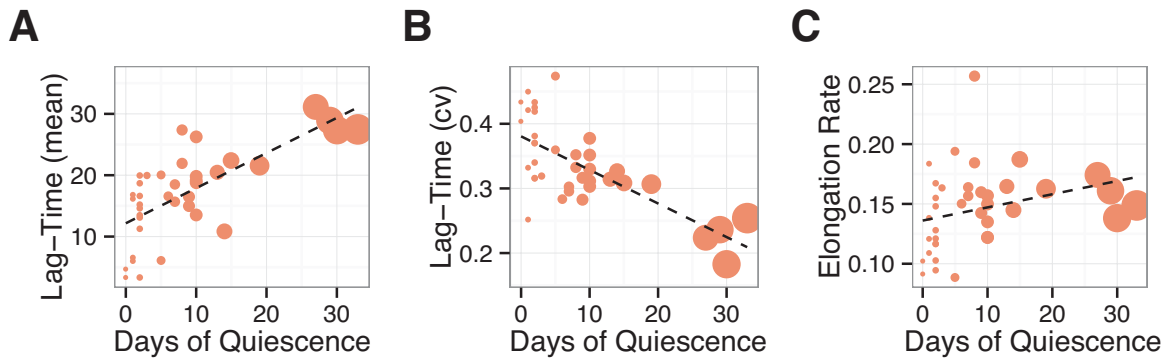


Figure 3.6: Representation of the wild types at 32 degrees Celsius. The larger the symbol, the longer the number of days in quiescence. The black dashed line is the regression line amongst all measurements. The plots represent (A) the average lag-time against the number of days in quiescence, (B) the coefficient of variation of the lag-time against the number of days in quiescence and (C) the elongation rate against the number of days in quiescence.

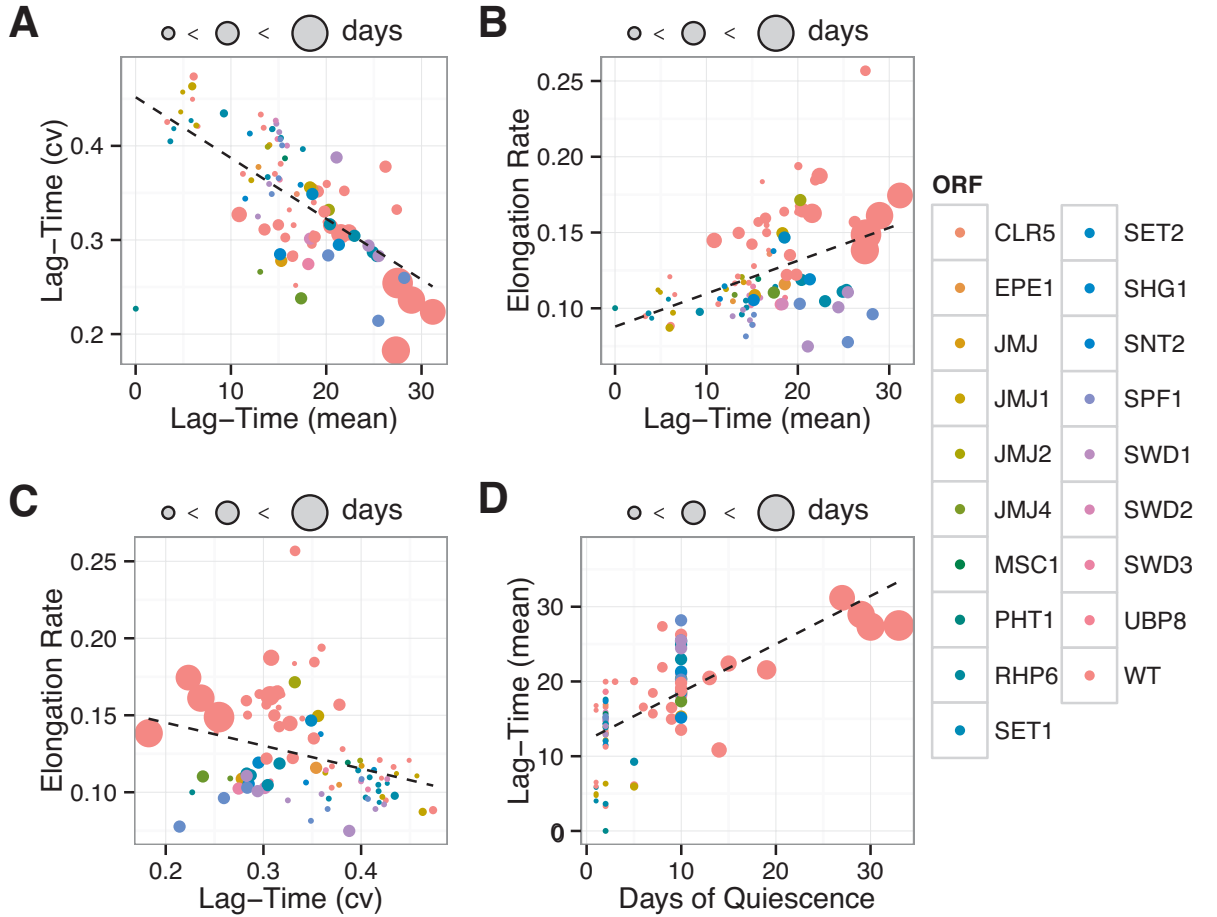


Figure 3.7: Representation of the wild types and mutants at 32 degrees Celsius. The larger the symbol, the longer the number of days in quiescence. The black dashed line is the regression line amongst all measurements. (A) Coefficient of variation against the average lag-time. The Pearson coefficients of correlation for the wild types and mutants are respectively $r = -0.452$, $P < 0.001$, and $r = -0.605$, $P < 0.001$. (B) Elongation rate against the average lag-time. The Pearson coefficients of correlation for the wild types and mutants are respectively $r = 0.637$, $P < 0.001$, and $r = -0.081$, $P < 0.001$. (C) Elongation rate against the coefficient of variation. The Pearson coefficients of correlation for the wild types and mutants are respectively $r = -0.227$, $P < 0.001$, and $r = -0.065$, $P < 0.001$. (D) Average lag-time against the number of days in quiescence. The Pearson coefficients of correlation for the wild types and mutants are respectively $r = 0.74$, $P < 0.001$, and $r = 0.676$, $P < 0.001$.

As one can see in Figure 3.8, the mutants grown with no quiescence, indicated by a gray ellipsoid on the plots, do not show the same behaviour as the mutants grown with quiescence. The scatter plot in Figure 3.8.D shows the relationship between the elongation

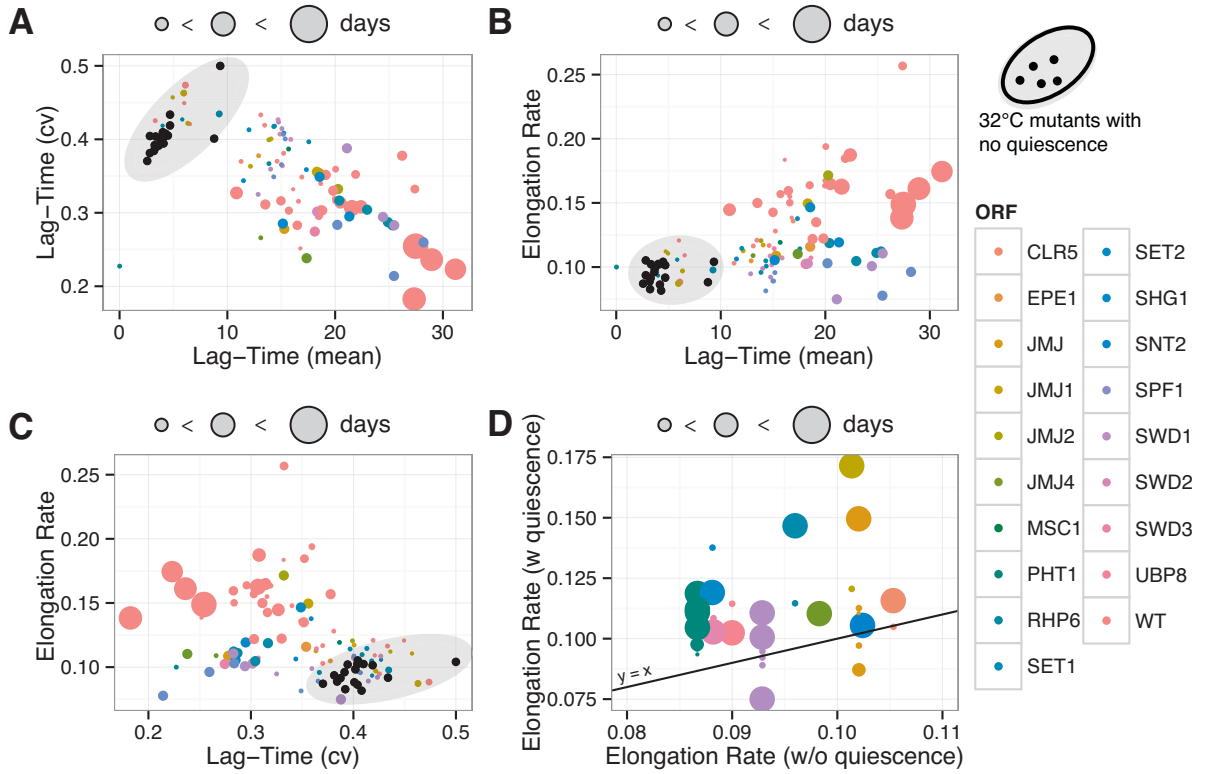


Figure 3.8: Plots of the mutants growing without quiescence. The four plots show (A) the coefficient of variation of the lag-time against the average lag-time, (B) the elongation rate against the average lag-time, (C) the elongation rate against the coefficient of variation of the lag-time and (D) the elongation rate with quiescence against the elongation rate without quiescence.

rates of the mutants grown, with and without quiescence. Although the Pearson correlation ($r = 0.34$, $P < 0.001$) is significant, one can see that the pace of growth of the mutants is different as they grew under quiescence or not.

3.4.2 Fractional Growth of the Population

In this particular model, the cell itself is no longer considered as a crucial element of the analysis, as we consider the growth of the population to be biased by a fraction of cells not growing. We also recall that this part of the analysis only concerns the mutants grown at 32 degrees Celsius, as the mutants grown at 25 degrees Celsius do not have much measurements and therefore do not provide the best statistical estimations.

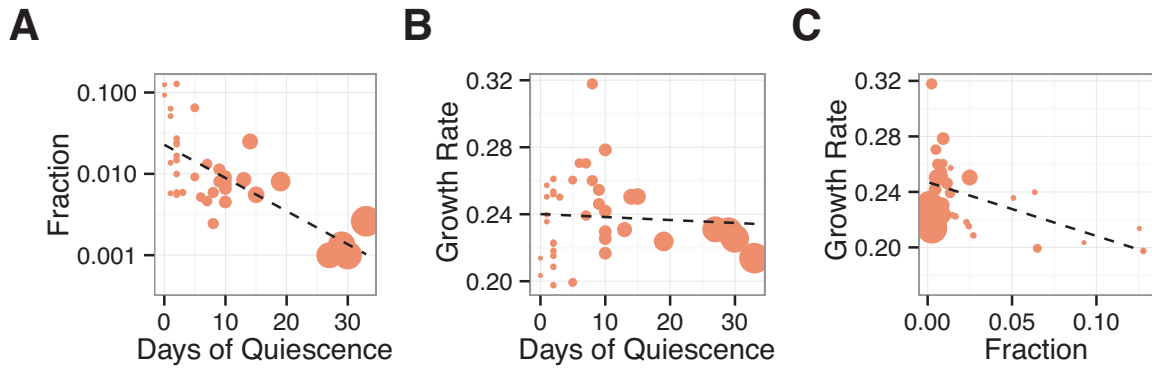


Figure 3.9: Representation of the fraction of wild types growing. (A) The plot represents the log-dependance between the fraction of cells growing and the number of days in quiescence. The longer the number of days in quiescence and the smaller the fraction of cells growing. (B) Relation between the number of days in quiescence and the growth rate. The longer the number of days in quiescence, the faster the growth rate. (C) Dependence between the fraction of cells growing and the growth rate.

For the wild types, one can see (Figure 3.9) that there is a high negative correlation ($r = -0.7$, $P < 0.001$) between the number of days in quiescence and the fraction of cells growing (Figure 3.9.A). This negative correlation indicates that the longer the number of days in quiescence, the smaller the fraction of cells growing, possibly telling that some of the cells might either die or not wake up after a long time in quiescence. There is a small correlation ($r = 0.23$, $P = 0.013$) between the number of days in quiescence and the growth rate (Figure 3.9.B). A possible explanation of this positive correlation might be that the cells that have been quiescent for a long time, might need to grow faster to catch up with their usual growth rate. Finally, there is a clear negative correlation ($r = -0.58$, $P < 0.001$) between the fraction of cells growing and the growth rate (Figure 3.9.C). The larger the fraction of cells growing, the smaller the growth rate, as the population has more cells to populate it.

In general, we observe the same trends for the mutants dataset (see Figure 3.10). However, one can see that there are some outliers. In Figure 3.10.C, the mutants: *set1* Δ , *spf1* Δ , *swd1* Δ , and *rhp6* Δ are outliers. According to their fractions, they should have a larger growth rate. Interestingly, these outliers are also found previously, in the stochastic lag

model. Here, we validate our observations by also looking at the residuals of the mutants, calculated using the wild-type trends (see Figure A.3 in appendix).

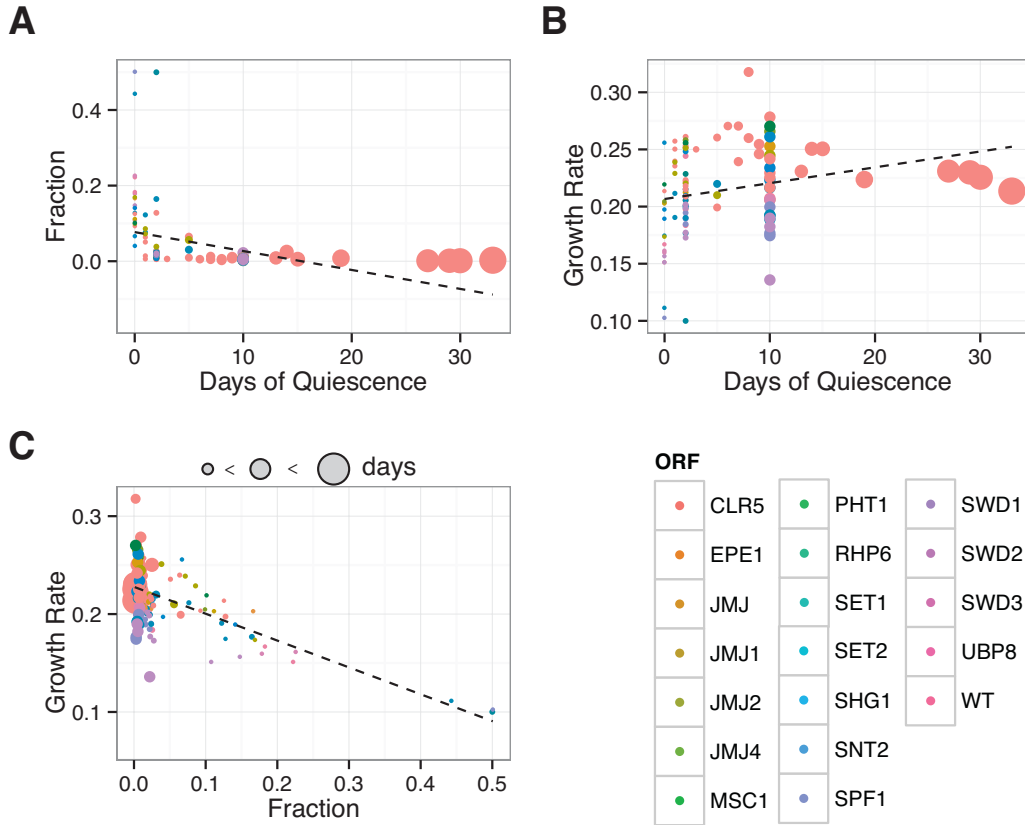


Figure 3.10: Plots of the mutants and wild types growing without lag but only a fraction of the organisms grow. (A) The plot represents the dependance between the fraction of cells growing and the number of days in quiescence. The longer the number of days in quiescence and the smaller the fraction of cells growing. (B) Relation between the number of days in quiescence and the growth rate of the ORF. The longer the number of days in quiescence, the faster the growth rate. (C) Dependance between the fraction of cells growing and the growth rate.

3.5 Discussion

In this study, we modelled the dynamic transition from quiescence to proliferation in *S. pombe*. The aim was to develop two mathematical models to comprehend and explain the lag phenomenon occurring in between the exit of quiescence and the beginning of the population growth.

The two models differed in their approach of modelling the growth of the population. The first model is a single-cell level based process, predicting the behaviour of a cell leaving quiescence and waiting a random time (the lag-time) before growing again. The second model, based on a macroscopic level, studies the growth of the population when a fraction of the cells does not grow and the remaining fraction grows. The first model, designed by Baranyi [9, 8], did not provide any experimental data to be validated. In this research we validated the model by adding a normalised lag-time distribution and we fitted experimental data to find the morphological parameters involved in the model.

Both models showed comparable results. Indeed, we were able to extract a list of mutants diverging from the wild-type trends. Here, we found that *set1* Δ , *spf1* Δ , and *swd1* Δ were the mutants diverging the most from the wild-type trends. We found, using the Pombase search engine [151], that the three genes are members of the Set1C complex. This complex catalyses the methylation of histone H3. In addition, we found that the *rhp6* Δ gene, also perturbed the transition from quiescence to proliferation.

The novelty of our analysis was to apply Baranyi's model to experimental data and compare it to our new model which suggests that a fraction of cells might have stopped growing after leaving quiescence. Our models were designed to perform a preliminary analysis that will lead to further experiments aimed at identifying which model fits the data correctly. The follow-up experiments will control the viability of the cells throughout the entire growth of the population, to detect whether only a fraction of cells is growing or the cells actually undergo a stochastic lag-time.

Chapter 4

Global trends in Phenotypic Variability and Robustness

4.1 Introduction

All biological systems and more especially single-celled organisms show variability in their phenotypes. This type of variability, called "phenotypic variability" is often interpreted as a consequence of the genetic heterogeneity of the population, for example as single nucleotide polymorphism (SNP). For decades now, biologists have agreed that even in populations formed of genetically identical organisms, the organisms exhibit significant phenotypic variability (Spudich & Koshland [128]). Recent progress in molecular biology and mathematical modelling have produced new insights to comprehend the source of non-genetic variability (Raj & van Oudenaarden [110], Shahrezaei & Swain [123] and Sanchez & Golding [119]). Non-genetic variability can be attributed to two sources of noises which are "intrinsic" and "extrinsic". Intrinsic variability comes from the probabilistic nature of the molecular collisions inside the organisms [135]. Extrinsic variability arises due to external factors surrounding the organism. The intrinsic and extrinsic types of variability have been investigated and characterised for the case of noise in gene expression (Swain et al. [132]). In addition, it was shown that variability in gene expression is spread across sub-processes and cellular phenotypes, as for example for the cell cycle (Nachman et al. [92]), cell signalling (Colman-Lerner et al. [24]), cell fate (Suel et al. [130]) and the response

towards drugs of cancer cells (Spencer et al. [125]).

As demonstrated by Kussel & Leibler [65], Acar et al. [1] and Pancaldi [101], phenotypic variability can be beneficial for a population, to overtake environmental stress and fluctuations and therefore survive. For example, to protect the population from environmental stress, it has been showed by Levy et al. [77] that there exists a resistance gene acting as a bet-hedging mechanism, allowing some cells to survive while others die.

The capacity for a population to suppress variability in certain phenotypes is termed "robustness" [58, 10, 88, 35]. Robustness is useful for the organisms as it can ensure reliability and reproducibility. Following this idea of robustness, phenotypes exhibiting large phenotypic variability are not robust. On the contrary, robust phenotypes will exhibit small variability. Here, we refer to robustness with respect to genetic modification and molecular/environmental fluctuations as genetic robustness and stochastic robustness respectively.

In the literature, most of the experiments and models devised to study phenotypic variability have been focused on a single or few phenotypes. However, phenotypes can be seen at the molecular, organellar and cellular level. Therefore the number of phenotypes available is large. Each phenotype is characterised by its own level of variability and also robustness. In this research, we investigate global trends in phenotypic variability and robustness. To this end, we employ a high-dimensional database of microscopy-derived data in *S. cerevisiae* [99]. The database contains a large collection of gene knockouts (mutants) and wild types. An additional dataset containing 37 wild strains [152] is also added to the original database. The morphological phenotypes analysed in the database allow us to cross-compare phenotypic variability amongst all phenotypes. This yeast morphological phenotypic database has been, in the past, used in several studies to investigate the genetic nature and the causes of phenotypic variability. Amongst all studies, the pioneering work by Levy & Siegal [76] identified several hundred mutants capable of potentially increasing phenotypic variability. These mutants are termed phenotypic capacitors, according to their capacity of buffering (storing) phenotypic variability. Subsequently, it has been shown that the deletion of phenotypic capacitors has a cost on fitness [76, 72, 143].

In this study we explore global trends in phenotypic variability. Firstly we observe correlations between genetic and stochastic variability. Secondly, we introduce a novel measure of global phenotypic variability (GPV) to classify genetic modifications and quantify their effect on phenotypic variability. We define four classes of mutants: the phenotypic inducers whose deletions decrease phenotypic variability, the phenotypic capacitors whose deletions increase variability, the phenotypic maintainers which redistribute variability across the phenotypes and keep the GPV relatively constant and finally the phenotypic neutrals which do not significantly affect phenotypic variability in any phenotype. Our analysis demonstrates a clear relationship between robustness, phenotypic variability and fitness by employing growth rates of the mutants, measured for similar experimental conditions as in Ohya et al. [99].

Our methodology is divided in five parts. In the first section, the yeast morphological phenotypic data is cleaned and processed to investigate genetic and stochastic variability on a subset of "representative" phenotypes found by clustering the dataset. The second section introduces the mathematical tools employed to construct the global phenotypic variability score. Moreover, to validate our score, we compare it to existing genetic scores available in the literature. In the third section, we segregate the mutants into four categories, according to their effect on phenotypic variability. In addition, we investigate robustness and fitness of the four categories. To validate our methodology, we analyse a second computation of the global phenotypic variability score using another mathematical formulation. Finally, the last section concludes our analysis.

4.2 Processing of the Saccharomyces Cerevisiae Morphological Database and Properties of the Morphological Phenotypes

In this research, we analyse phenotypic variability and robustness of *S. cerevisiae* by employing a dataset of morphological measurements, the Saccharomyces Cerevisiae Morphological Database (SCMD) from Ohya et al. [99]. This dataset contains morphological measurements for a collection of 4718 mutants and 126 biological replicates of a wild-type strain (here, *his3* Δ). We also integrate a second dataset containing 37 wild strains (see Yvert et al. [152]) analysed in a similar manner as the SCMD. Both sets were created with the CalMorph software, specifically developed by Ohya et al. to process and analyse images of yeast cells.

In the past, the SCMD has proven to be of great use in the analysis of phenotypic variability. For instance, Lehner [73] employed the SCMD and a similar approach as Levy & Siegal [76] to show that mutational robustness does not require to be affected by stochastic perturbations, as his analysis showed that the nonessential genes conferring robustness for stochastic changes also buffer genetic changes. In the same idea, Cooper et al. [27] tested the hypothesis that random mutations might affect, on average, the fitness of the cells. Finally, the pioneering work of Levy & Siegal [76] who employed the SCMD to detect phenotypic potentials, mutants capable of buffering phenotypic variations. The later analysis greatly motivated our present study.

The SCMD was created using the CalMorph software, developed by Ohya et al. CalMorph analyses morphological phenotypes via an image acquisition process on growing cells. For each strain, the number of cells available for the analysis ranged from a few to hundreds. In this section we remove the mutants having fewer than 25 single-cell measurements to avoid employing mutants and wild types with non statistically significant number of single-cell measurements. Thereafter, we cluster the phenotypes (as in [76]) to only consider medoids, i.e., phenotypes being representatives of their cluster. Finally, we present properties of the morphological phenotypes and show the difference between environmental and genetic variability.

In this project, all the computations are realised with The R project for Statistical Computing software (<http://www.r-project.org>). The packages used for specific computations are mentioned within the sections using them.

4.2.1 Removal of the Phenotypes and Mutants with Non-significant Number of Single-cell Measurements

The *Saccharomyces Cerevisiae* Morphological Database (SCMD) is delivered in the form of two sets of data: one containing 4718 analysed mutants (ORFs) and one for the 126 biological replicates of the wild types (*his3Δ* strain). Each dataset is structured as a matrix with rows referring to the mutants (or the wild types) and columns to the morphological phenotypes. In total, there are 501 morphological phenotypes measured for different features, such as nucleus size, actin patches sizes...etc. A complete description of the SCMD is provided in the next section. It must be noted that the morphological phenotypes are separated into two categories: the mean phenotypic value and its coefficient of variation. The mean phenotypic value is the average value computed on single-cell measurements of a specific mutant or wild type. Similarly, the coefficient of variation is calculated on the same set of single-cell measurements.

In this project, we consider phenotypic variability by employing the coefficient of variation of the mutants, the wild strains and the wild types. Therefore, to ensure statistically significant calculations of the coefficient of variation we remove the phenotypes (and some of the mutants) that were created using only a small number of cells.

For each phenotype, i.e., each column of the mutants dataset, we compute the median number of single-cell measurements. If this number is less than 25, we remove the phenotype. We repeat the same procedure for the ORFs, i.e., the rows of the mutants dataset, to remove the mutants with small numbers (less than 25) of single-cell measurements.

After removal, the mutants dataset finally contains 4681 ORFs and 428 morphological phenotypes.

In the next part we introduce a method (similarly as in [76]) to work on classes of independent phenotypes instead of the full available phenome.

4.2.2 Clustering of the Phenotypes into Classes of Representatives

The *Saccharomyces Cerevisiae* Morphological Database provides a large collection of morphological phenotypes. Initially there are 501 phenotypes and after cleaning the dataset, 428. The phenotypes are measurements of different subunits/cellular compartments of the cell. One can find phenotypes related to the cell, the nucleus and actin patches. A phenotype can be represented, in most cases, by two values, the mean phenotypic value and the coefficient of variation. The terminology used in the SCMD is clear as phenotypes related to the cell have labels starting with a "C", phenotypes related to the nucleus have labels starting with a "D" and finally phenotypes related to the actin patches have labels starting with an "A". The phenotypes have been measured at different stages of the cell cycle of *S. cerevisiae* and have been divided into three stages, namely: stage A (G1 phase), stage A1B (S/G2 phases) and stage C (M phase). As mentioned earlier, a phenotype contains two values, the mean phenotypic value and its coefficient of variation.

In this study, we want to analyse phenotypic variability of mutants and see how variability affects robustness. To this end, we want to employ the morphological phenotypes available in the SCMD and, in particular the coefficient of variation of the phenotypes. Instead of using the complete available phenome, we cluster it to create classes of phenotypes and we look at the representatives of the classes. This approach was already employed by Levy & Siegal [76] to consider groups of phenotypes and simplify the analysis. Their approach is slightly different as they used the standard deviation instead of the coefficient of variation.

To restrain the number of phenotypes and only work on representatives, we use a parti-

tioning around medoids algorithm (PAM, see Reynolds et al. [112]) on the mutants dataset (here, denoted M). Due to the diversity of morphological phenotypes, i.e., length, surface, angles, brightness, ...etc. we standardise the dataset M to produce a new dataset M_{standard} which contains phenotypes with no units. The standardisation is:

$$M_{\text{standard}}^i = \frac{M^i - \langle M^i \rangle}{sd(M^i)}, \quad i = 1, \dots, 428 \quad (4.1)$$

where i stands for the i^{th} column of the dataset.

As part of the K-medoids algorithm family, the partitioning around medoids algorithm must be run with a pre-defined number K of clusters. To find the optimum K , i.e., the best separation between the clusters, we compute the average silhouette width (Rousseeuw [117]) and pick the K for which the average silhouette width is maximal. Using the R package *cluster*, the optimum K is 67 as illustrated in Figure 4.1.A. The 67 medoids (phenotypes being representatives) are listed in Table 4.1. The complete list with the phenotypes and their clusters is available in Appendix B.1. Here, the medoids comprise all three types of phenotypes: the cell, the nucleus and actin patches. A complete definition of the phenotypes is available in the SCMD.

We employ the medoids found previously to remove the wild-type experiments showing "abnormal" coefficients of variation, i.e., coefficients not following the general wild-type trend. We want to remove these wild-type experiments as they might bias the outcome of the whole analysis by providing wrong "information" on the nature of the wild types. Hence, we look at the clustered coefficient of variations of the medoids with a dendrogram and we remove the wild-type experiments having abnormal variations. In Figure 4.1.B, four wild-type experiments show such abnormalities in their coefficients of variation. The removal reduces the number of wild-type measurements from 126 to 122.

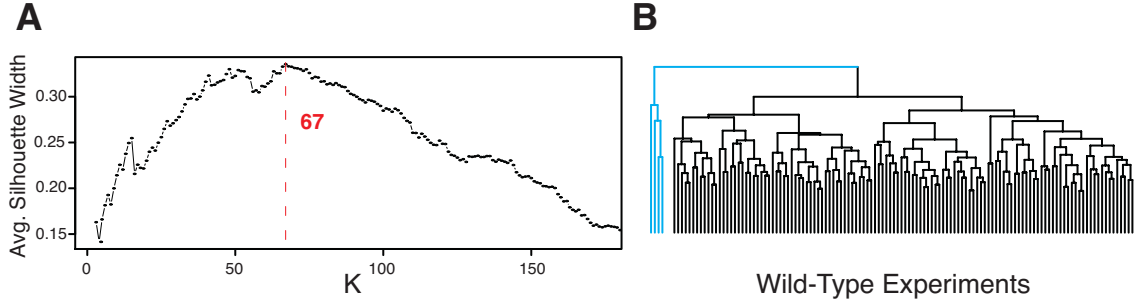


Figure 4.1: The partitioning around medoids algorithms confers the possibility to cluster the morphological phenotypes into classes, represented by their medoids. This classification allows to work on a smaller set of phenotypes instead of the complete available phenome. The average silhouette width is employed to find the optimum number K of clusters as presented in panel A. Then, similarly to the removal of the mutants having few single-cell measurements, we look at the dendrogram of the wild-type experiments according to the values of their coefficients of variation. Here we see that the dendrogram shows two parents. We remove the first branch (highlighted in blue) which contains the following experiments : HIST-023, HIST-032, HIST-036 and HIST-100.

4.2.3 Genetic Properties of the Morphological Phenotypes

Here, we compare the average variability of the wild types against the coefficient of variation of the mean phenotypic values of the mutants (also denoted YKO) and the wild strains (denoted WS). One must regard this comparison as a way of understanding genetic and phenotypic variability. Each mutant possesses a mean phenotypic value, computed as the average value amongst all single-cell measurements, and a coefficient of variation, also computed amongst all single-cell measurements, for all phenotypes. The same applies for the wild types and the wild strains. Therefore, it is possible to look at a particular phe-

A7-1_A	A8-1_A	C115_A1B	D147_A	D117_C	D155_A	D182_A
A101_A1B	A9_A1B	C116_A1B	D106_C	D119_C	D155_A1B	D17-2_C
A7-1_C	A9_C	C116_C	D147_A1B	D121_C	D155_C	D184_A1B
A102_A1B	C12-1_A1B	C118_A1B	D108_C	D123_C	D157_C	D182_C
A102_C	C102_C	C117_C	D109_C	D176_A	D193_C	D188_A
A104_C	C12-1_A	C126_A1B	D110_A1B	D14-1_C	D193_A1B	D196_C
A120_C	C104_C	C127_A	D152_C	D14-2_C	D163_C	D198_C
A122_A	C106_A1B	C13_A1B	D153_C	D178_A1B	D167_C	
C11-2_A1B	D158_C	D141_C	D114_A1B	D15-3_C	D170_A1B	
A7-2_C	C111_C	D104_A1B	D135_A	D151_C	D170_C	

Table 4.1: Medoids found using the PAM algorithm. The algorithm returns 67 medoids for the largest silhouette width.

nototype and calculate the coefficient of variations of all mean phenotypic values across all mutants, wild types and wild strains. We call this coefficient the "genetic variability" as it is a measure of the variability produced by the mutants. Similarly, by averaging the coefficients of variation of the wild types, we define a measure of "stochastic variability". By its nature, the coefficient of variation of the wild types is a measure of variability in a neutral setting, i.e., with no genetic mutations. Therefore, averaging the coefficients of variation of the wild types for a given phenotype provides an idea of how variable is the organism in nature.

Here, we measure genetic and stochastic variability for the 67 medoids. We are interested in the correlations between the genetic variability of the mutants and wild strains and the genetic variability of the mutants compared to the stochastic variability of the wild types. We see in Figure 4.2.A that the genetic variability of the mutants and the wild strains is well correlated ($r = 0.54$, $P < 0.001$). The three categories of phenotypes, i.e., actin, cell and nucleus, are also labelled on the scatter plot. In addition, the genetic variability of the mutants is also well correlated with the stochastic variability of the wild types ($r = 0.59$, $P < 0.001$). This suggests that phenotypes are consistent with the robust nature of variability or in other words, that more variable phenotypes are less robust.

For both comparisons in Figure 4.2, the phenotypes are consistent with an increase of variability. For a small amount of variability in, for instance, the genetic variability, there is a small amount of stochastic variability. The same observation applies between both genetic variability. Only phenotypes related to the actin patches show variability not following the general trend. A possible answer to this is the appreciation of the actin patches in the image acquisition process of the SCMD. Unlike size related phenotypes, actin is measured with fluorescence, which could have technical noise, i.e., noise coming from the image acquisition process and therefore bias the image acquisition process.

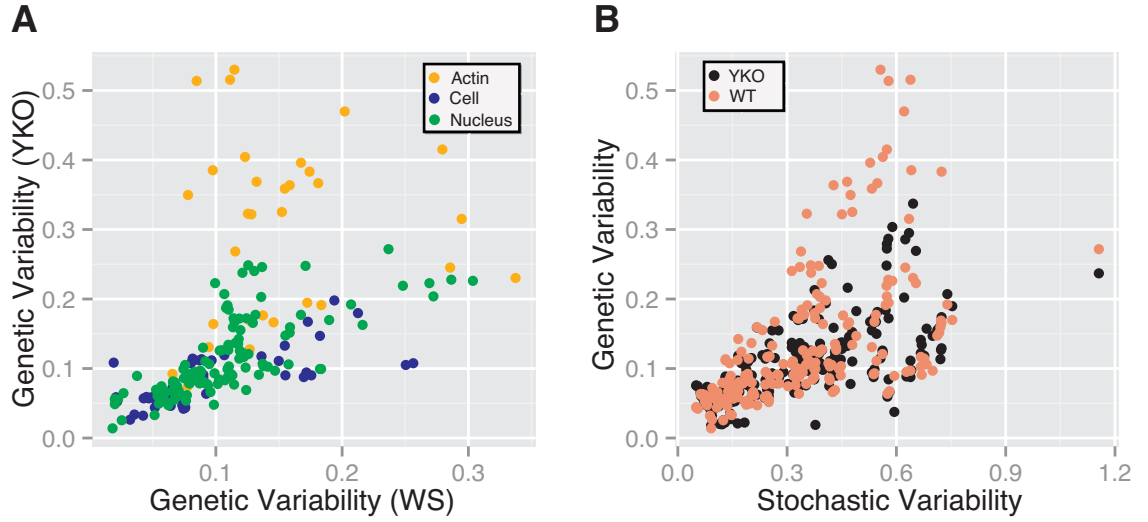


Figure 4.2: Comparison between stochastic and genetic variability for the WS, the YKO and the WT. Genetic variability is seen here as a coefficient of variation, calculated using the mean values of a given phenotype across all mutants or wild strains. Stochastic variability, for the wild types, is calculated by taking the average coefficient of variation. A. Genetic variability of the 37 wild strains and the YKO shows a high correlation ($r = 0.54$, $P < 0.001$), indicating that a phenotype's robustness with respect to single gene knockouts is a good predictor of the phenotype's robustness with respect to evolution in different environment. When the correlation is computed without considering the actin phenotypes, the correlation between the WS and YKO increases to $r = 0.72$ ($P < 0.001$). B. In comparison with stochastic variability of the wild types, the genetic variability for the YKO (plotted here) and the WS is also highly correlated (respectively $r = 0.59$, $P < 0.001$ and $r = 0.67$, $P < 0.001$), suggesting that phenotype's robustness with respect to non-genetic stochastic effects is a good indicator of phenotype's robustness with respect to genetic modifications.

4.3 Scoring of Phenotypic Variability

The analysis of phenotypic variability in *S. cerevisiae* has already been studied in the past, as for example in the work of Levy & Siegal [76]. In their analysis, they provided a list of scores for the mutants and wild types according to their capacity of buffering phenotypic variability. Here, we also want to create scores that can allow us to differentiate between the mutants and the wild types. Therefore, we employ both the mean phenotypic value and its coefficient of variation to acquire information on variability. We utilise the wild types as a control group and refer to them to compute a special form of residuals on the

mutants, calculated using the coefficient of variation. The residuals indicate the dispersion of a mutant compared to the general trend. A positive residual will signify an increase of variability and a negative residual a decrease of variability. A first score, functioning as a ratio, is then given to each mutant according to its median residual and the median of the absolute value on all residuals. Finally, we compare our genetic score with other scores provided in the literature.

4.3.1 Residuals for the Coefficient of Variation

In the *Saccharomyces Cerevisiae* Morphological Database, the phenotypes are paired with the mean phenotypic value and its coefficient of variation. The wild types and the mutants can be plotted together to have a first insight in the dispersion of the mutants for the coefficients of variation. Here, we employ a similar approach as in the work of Levy & Siegal [76] who looked at phenotypic variability by considering the residuals of the mutants calculated with a Lowess regression on the whole mutants dataset. In the work of Levy & Siegal, the residuals are computed between the mean phenotypic value and the standard deviation. However, our methodology differs from the methodology of Levy & Siegal as we look at the residuals calculated using a standardised $L1$ -regression between the mean phenotypic value and the coefficient of variation. The $L1$ -regression has the particularity of being more sensitive to outliers and therefore allows us to determine mutant with high variability better. Here and throughout the rest of this chapter, we employ the term "residual" to define a modified residual scaled by the standard deviation of the wild-type residuals. The standard deviation of the wild-type residuals is employed to obtain a fold change measure of phenotypic variability. The regression between the mean phenotypic value and the coefficient of variation is employed, here, to identify mutants that have high or low coefficient of variation.

Here, we utilise an $L1$ -regression to give the outliers more importance to our analysis. Unlike the Lowess regression, the $L1$ -regression is more sensitive to outliers and therefore gives them larger residuals. This must be understood as a wish to capture mutants showing large coefficients of variations and thus high phenotypic variability. Hence, for each

medoid, we compute the $L1$ -regression between the mean phenotypic value and its respective CV, across all mutants. In addition, we aggregate a synthetic weight to the wild-type data by including their dataset 38 times ($38 \times 122 = 4636$ data points) to almost equalise the number of mutants. Each mutant is attributed a residual, calculated as the difference in height (here the coefficient of variation) between the original CV value and the fitted CV value (from the $L1$ -regression). The residual is then normalised by the standard deviation of the wild type residuals. We repeat the process for all mutants and medoids and construct a matrix R of residuals. The mathematical formulation of the matrix is

$$R = (R)_{ij} = \log_2 \left(\frac{o_{ij}^m}{f_{ij}^m} \right) \times \left[\text{SD} \left(\log_2 \left(\frac{o_{kj}^w}{f_{kj}^w} \right) \right) \right]^{-1}, \quad i = 1, \dots, 4681, \quad k = 1, \dots, 122, \\ j = 1, \dots, 67, \quad (4.2)$$

where o_{ij}^m and f_{ij}^m are the data point and respectively fitted point for the mutant i and the phenotype j and o_{kj}^w and f_{kj}^w are the data point and fitted point for the wild-type experiment k for the phenotype j . As mentioned earlier, the residual defined in Eq.(4.2) is not a proper residual as usually considered in the literature, i.e., the distance between the data point and its fit. Here the use of the logarithm provides us with a fold change between the data point and the fitted regression. In addition, we divide the logarithm by the standard deviation across all the "residuals" computed for the wild types for the same phenotypes. The scaling is performed on the fold change of the mutants to obtain a better dispersion of the mutants compared to the wild types.

We also perform the same calculations of the residuals on the dataset containing the 37 wild strains (WS). The calculation of the residuals for the mutants and the wild strains is illustrated in Figure 4.3.

As depicted in Figure 4.3 in panels A and B, the residuals (signalised by the blue arrow) can have either a positive or a negative value, depending on their position compared to the $L1$ -regression. The value of the residual must be interpreted as a fold change compared to the wild types. We plot the number of residuals being smaller than -1 or larger than 1 for

all medoids in panels C and D of Figure 4.3. For both the mutants and the wild strains, the number of positive residuals is higher than the number of negative residuals across the medoids. We investigate this difference in the number of negative and positive residuals at the mutant level in the next part of this section, by introducing a score measuring the global phenotypic variability of a mutant.

4.3.2 Global Phenotypic Variability of the Mutants

The essence of our research resides in the study of variability, seen as the coefficient of variation of the strains. Previously, we showed that the sign of the residuals varies largely depending on the phenotype (medoid). Here, we want to attribute a score to the mutants, to see whether their residuals are negative or positive. However, considering each phenotype one by one would complicate the analysis at the mutant level. Therefore, to regroup the phenotypes and create a score for the mutants, we look at the global effect of the mutants on phenotypic variability. To this end, we use the matrix R of residuals computed in Eq.(4.2) and we create two scores for the mutants: the Global Phenotypic Variability (denoted GPV) and the Global Absolute Phenotypic Variability (denoted GAPV).

The Global Phenotypic Variability (GPV) is the median value of the residuals computed across all phenotypes for a given mutant. We utilise the median as it is unbiased compared to the mean. By taking the median of all residuals of a mutant, we propose a qualitative information for the phenotypic variability of the mutant. Hence, the GPV of mutant i is defined by

$$\text{GPV}_i = \text{median}(R_{ij}) , \ i = 1, \dots, 4681 , \ j = 1, \dots, 67. \quad (4.3)$$

Similarly to standard statistics requiring the mean and the variance of a probability distribution, we define a second score for the mutants, the Global Absolute Phenotypic Variability (GAPV), based on the absolute value of the residuals to study the spread of the global phenotypic variability and observe it across the set of mutants. Thus, the second score for

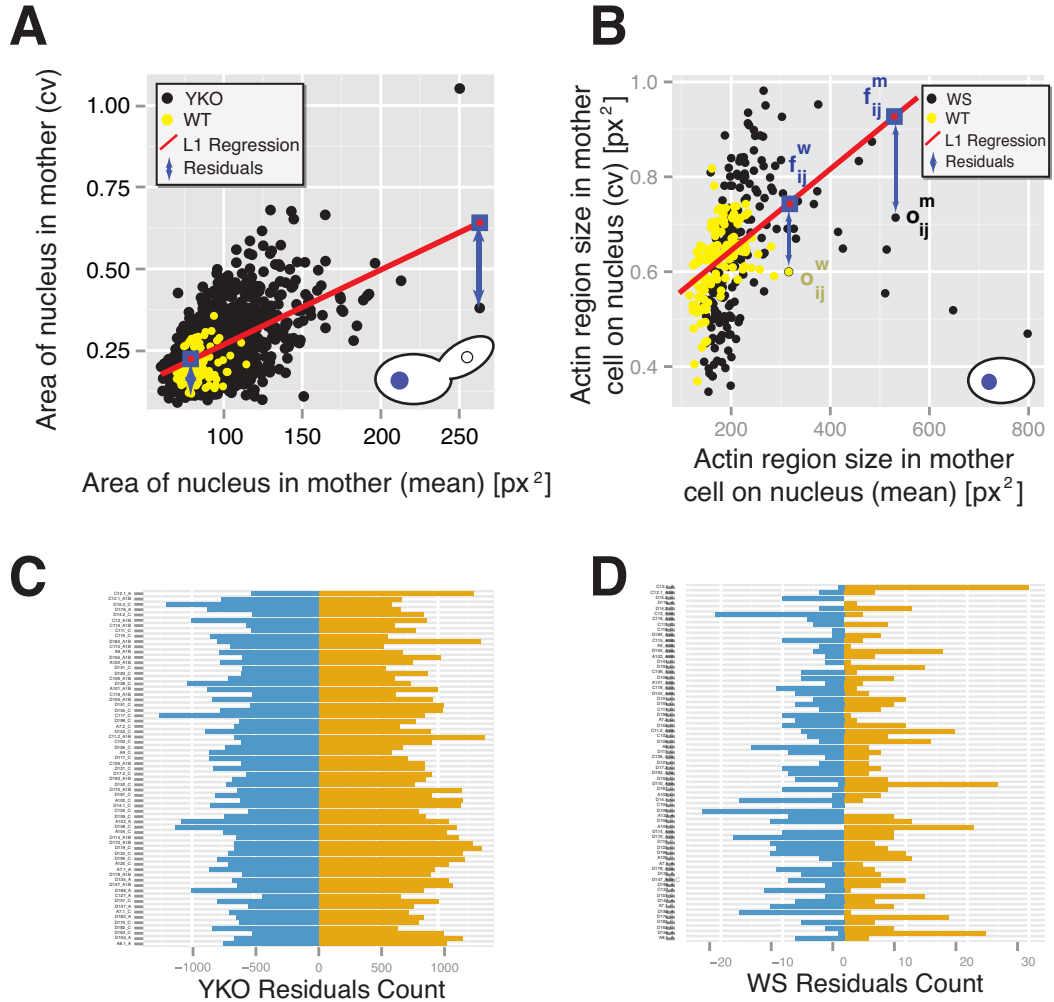


Figure 4.3: The $L1$ -regression (the red lines in panels A and B) allows to construct a matrix of residuals, based on the wild types (WT, yellow dots) and the mutants (YKO, black dots in A) & the wild strains (WS, black dots in B). The wild types are employed here to compare the mutants or wild strains and see the fold change for the coefficients of variation. The scatter plots in panels A and B represent the size of the nucleus in stage C. In panel B, the wild strains show a large dispersion of their mean phenotypic values and coefficients of variation compared to the wild types. The residuals greater than 1 or less than -1 in the fold change are represented in panels C and D for the 67 medoids.

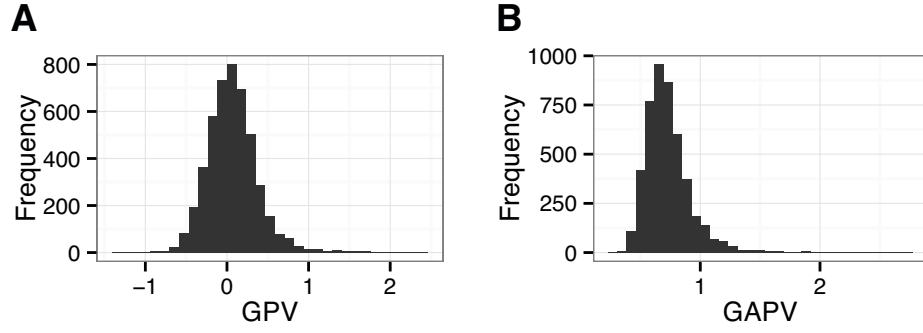


Figure 4.4: The global phenotypic variability (GPV) score and the global absolute phenotypic variability (GAPV) are measures of phenotypic variability at the mutant level. By their definitions, the scores discern between mutants having positive residuals, i.e., increasing phenotypic variability globally and negative residuals, i.e., decreasing variability. The distribution of the GPV scores for all 4681 mutants is represented in panel A. Similarly, the distribution of the GAPV scores is showed in panel B. Both distributions inform on the presence of mutants having mainly three effects: the aforementioned increase or decrease of variability and mutants with zero GPV. The latter mutants both increase and decrease phenotypic variability.

mutant i is

$$\text{GAPV}_i = \text{median}(|R_{ij}|), \quad i = 1, \dots, 4681, \quad j = 1, \dots, 67. \quad (4.4)$$

The distributions of the GPV and GAPV scores are shown in Figure 4.4. As one can see in panel A, the negative value of the residuals is preserved for the GPV score. We also see that there are mutants having a GPV score of zero. This suggests that there exist mutants having both positive and negative residuals. This observation indicates that for some mutants, their phenotypic variability can be either positive or negative. In order to employ the GPV and GAPV scores further, we pursue our investigation by looking at the correlation of both scores with existing and already validated measures of genetic variability.

4.3.3 Comparison with Other Measures of Variability

The global phenotypic variability and global absolute variability, defined and calculated previously, showed that there exists mutants decreasing or increasing variability at a global

level amongst all medoids. However, these increasing and decreasing effects on phenotypic variability might be biased by the nature of the scores. Defining and employing scores not necessarily already used in the literature always requires a careful attention and a methodical validation. Therefore, we compare our GPV and GAPV scores with similar scores devised by Levy & Siegal [76] and Rinott et al. [114]. In addition, we re-calculate our scores on a smaller sets of medoids to test the efficiency of the scores to capture the nature of the mutants that increase or decrease variability in morphological phenotypes.

In their work, Levy & Siegal devised a score capable of analysing the ability of capacitance of the mutants available in the SCMD. The phenotypic potential of Levy and Siegal is defined as the top 50% residuals (ranked from the highest to the lowest) on their medoids, computed as an average across the selected residuals. In other words, for each mutant in the matrix of residuals, the medoids are ordered according to the value of the residuals and the top 50% residuals are averaged to produce a score.

A Pearson correlation analysis between our two scores: the GPV and GAPV and the phenotypic potentials of Levy & Siegal indicates high correlations (Figure 4.5). The Pearson's r for the GPV and the phenotypic potential is $r = 0.75$ ($P < 0.001$). Similarly, the correlation between the GAPV and the phenotypic potential is $r = 0.63$ ($P < 0.001$). These high correlations suggest that our scores, the GPV and GAPV can capture the same information on phenotypic potential as in Levy & Siegal but as we also showed previously, our scores provide additional information on mutants with negative impact, i.e., negative residuals, on phenotypic variability.

As a second test, we look at the correlations between our scores and the local and global variability as defined in Rinott et al. [114]. In their analysis, Rinott et al. analysed at the single-cell level, two fluorescent reporters activated by two different promoters subject to genetic deletions. They identified proteins affecting variability on the expression of the aforementioned reporters. They created two scores, depending on the effect of the perturbations. They used both a global score which represents the mutants affecting both reporters at the same time and a local score for the mutants affecting reporters independently. Each score was recorded for two fluorescent proteins: RFP and GFP. The correlations between

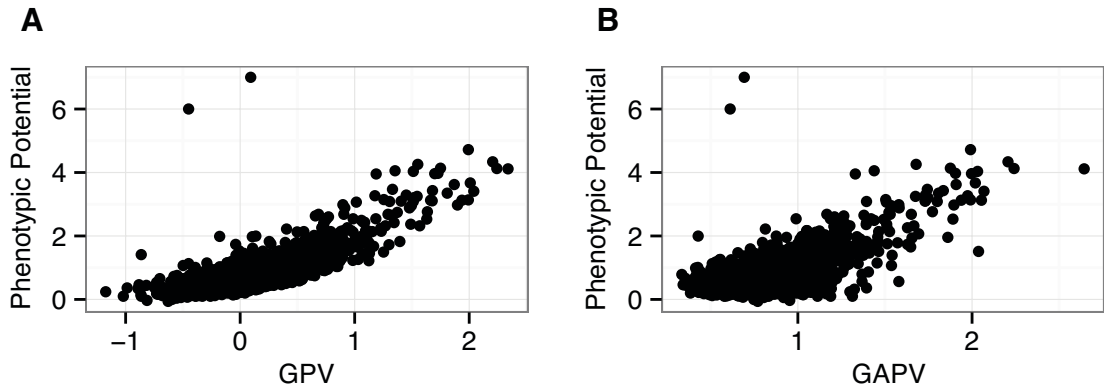


Figure 4.5: The creation of scores based on genetic modifications requires a methodic comparison with existing scores that have proved their efficiency in the literature. Here we compare the GPV and GAPV scores with the phenotypic potentials of Levy & Siegal [76]. Panel A shows a high correlation ($r = 0.75$, $P < 0.001$) between the GPV and the phenotypic potential scores. Panel B also shows a high correlation ($r = 0.63$, $P < 0.001$) between the GAPV and the phenotypic potential scores. Two mutants seem to be outliers in both plots.

the scores of Rinott et al and ours are listed in Table 4.2.

	RFP	Local	RFP	Global	GFP	Local	GFP	Global
GPV	-0.039	(0.503)	0.183	(0.002)	-0.173	(0.003)	0.113	(0.053)
GAPV	0	(0.996)	0.091	(0.119)	-0.111	(0.085)	0.077	(0.192)

Table 4.2: Correlations between the GPV and GAPV scores and the local and global variability scores from Rinott et al. [114]. The respective p-values of the Pearson's correlations are indicated in brackets.

Although some of the p-values in Table 4.2 are not significant (> 0.05), we see that the GPV score shows better correlations with the local score (calculated with a green fluorescent protein, GFP) and the global score (calculated with a red fluorescent protein, RFP). This indicates that the GPV score, to some degree, also captures global variability in other phenotypes, such as gene expression.

The correlations between the GPV and GAPV with the scores from Levy & Siegal and Rinott et al. not only indicate that our methodology is consistent but that our scores can also provide new information on the opposite effect (as in opposite to the phenotypic po-

tential) on variability. However, as a last observation, we look at the correlation between the global phenotypic variability score (as well as global absolute phenotypic variability) and a reduced set of medoids. Here we select medoids for which the number of clusters is $K = 15$, to reduce the number of phenotypes and see what type of phenotypes are well correlated with our scores. We chose $K = 15$ for two reasons. First, there is a drop in the silhouette width after 15 and second, this K reduces greatly the number of medoids to analyse, facilitating the visualisation for the correlations. This last comparison is performed to ensure that our scores are more related to cell or nucleus size than other phenotypes. As we want to record phenotypic variability, we think that phenotypes related to size are more reliable than phenotypes related to, for instance, actin patches, as the size-phenotypes are measured as length or surfaces whereas actin-phenotypes rely on the brightness of the fluorescence, which might be already experimentally noisy. The shortened list of medoids with their descriptions and the respective correlations with the GPV and GAPV scores are listed in Table 4.3.

Although the number of phenotypes related to actin patches is small, we see in Table 4.3 that our scores are more correlated for phenotypes related to the cell or nucleus size.

We pursue our analysis by introducing a method to segregate the mutants according to their global phenotypic variability and the global absolute phenotypic value. This to categorise mutants increasing or decreasing variability and analyse them separately.

Medoid	GPV ($r =$)	GAPV ($r =$)	Description
C101_C	0.42	0.49	Cell size on nucleus C
D141_C	0.39	0.45	Distance between nuclear brightest point in mother and mother hip on stage C
D14-3_C	0.39	0.55	Area of nucleus region in nucleus C
C12-2_A1B	0.37	0.44	Contour length of daughter cell on nucleus A1B
D117_A	0.30	0.25	Distance from mother cell's center to nucleus in nucleus A
D134_C	0.28	0.36	Distance between two nuclear brightest points
D17-2_C	0.27	0.26	Fitness to ellipse of the nucleus in the daughter cell
C115_A1B	0.18	0.30	Roundness of mother cell on nucleus A1B
D196_C	0.08	0.13	Maximal intensity of nuclear brightness divided by average
A120_A1B	0.04	0.22	Total length of actin patch link on nucleus A1B
A101_C	0.00	0.14	Actin region ratio on nucleus C
D193_C	-0.03	0.14	Average of nuclear brightness in whole cell
D114_A1B	-0.09	-0.12	Ratio of D110 to C128 on stage A1B
C126_A	-0.11	0.00	Unevenness of brightness on nucleus A
D158_C	-0.12	-0.27	Angle between D18.1,D1.1 and D18.1,C1.2 on stage C

Table 4.3: Correlations between the GPV and GAPV scores and a smaller set of medoids ($K = 15$). The scores are correlated with medoids related to the size of the cell or nucleus.

4.4 Neutral, Maintainer, Inducer and Capacitor Mutants

In the previous sections we showed that residuals calculated with the coefficients of variation of selected phenotypes (medoids) can be positive or negative, therefore indicating an increase and respectively a decrease in phenotypic variability. In addition, we created two scores, the global phenotypic variability and global absolute phenotypic variability of a mutant, capable of providing mutants with their global effect on phenotypic variability. We showed that the scores correlate with genetic scores already employed in the literature. Moreover, we examined the nature of the scores on a smaller set of medoids and we showed that size-phenotypes are more prevalent than actin-phenotypes in the calculation of the scores.

In this section, we are interested in using the global phenotypic variability and global absolute phenotypic value to create categories of mutants, depending on their effect on phenotypic variability. Here we segregate the mutants in four groups: the neutral, the inducer, the maintainer and the capacitor mutants. The neutral mutants are knockout genes not particularly affecting variability at the global level. The inducer mutants are mutants decreasing variability. The capacitor mutants are mutants increasing variability. Finally, the maintainer mutants are genes acting both as inducers and capacitors, therefore decreasing and increasing variability at the global phenotypic level.

4.4.1 Segregation of the Mutants into Categories according to their Phenotypic Variability

The segregation of the mutants in categories, according to their effect on phenotypic variability, is based on the idea that the wild types, by their nature, should not show large phenotypic variability. Therefore, we create the categories by considering the wild-type distributions of the global phenotypic variability and global absolute phenotypic variability scores. Hence and due to the normality of the wild-type scores (see Figure 4.6), we separate the mutants (YKO) and the wild strains (WS) by considering them as outliers of the wild-type distributions. The four groups are (mathematically) defined as follow:

$$\begin{aligned} \textbf{Capacitors} : & \text{GPV} > \text{median}(GPV_{\text{WT}}) + 2 \times \text{SD}(GPV_{\text{WT}}) \\ & \text{GAPV} > \text{median}(GAPV_{\text{WT}}) \end{aligned}$$

$$\begin{aligned} \textbf{Maintainers} : & \text{GPV} > \text{median}(GPV_{\text{WT}}) - 2 \times \text{SD}(GPV_{\text{WT}}) \\ & \text{GPV} < \text{median}(GPV_{\text{WT}}) + 2 \times \text{SD}(GPV_{\text{WT}}) \\ & \text{GAPV} > \text{median}(GAPV_{\text{WT}}) \end{aligned}$$

$$\begin{aligned} \textbf{Inducers} : & \text{GPV} < \text{median}(GPV_{\text{WT}}) - 2 \times \text{SD}(GPV_{\text{WT}}) \\ & \text{GAPV} > \text{median}(GAPV_{\text{WT}}) \end{aligned}$$

$$\textbf{Neutrals} : \text{GAPV} < \text{median}(GAPV_{\text{WT}})$$

where GPV_{WT} and $GAPV_{\text{WT}}$ stand for the 122 wild-type GPV and GAPV scores respectively. We find 586 mutants in the capacitor category, 85 mutants in the inducer category and finally 2206 mutants in the maintainer category. The remaining mutants are all in the neutral category.

To avoid falling in the trap of outlier detection by simply taking the mean and standard deviation of the distributions of the wild-type scores, we determine the probability of randomly choosing a mutant, by devising a permutation test on the residual values. We want to minimise the chance of selecting mutants for which their mean phenotypic values and coefficients of variations were randomly set. Thus, the test consists of shuffling the matrix R of residuals and then computing new GPV and GAPV scores. This to see whether mutants with random GPV and GAPV scores would produce the same results. In order to get a good approximation of the p-values of accepting "random" mutants, we repeat the permutation test 1000 times to create a distribution of random scores. Then, we compute the p-values of obtaining random ORFs according to the wild-type thresholds. In this setting, the p-values for the GPV score are: for the lower bound $p_{\min} < 0.001$ and for the upper bound $p_{\max} < 1 - 0.96 = 0.03$.

We visualise the GPV and GAPV scores in Figure 4.7. For each of the four categories, a mutant is plotted to illustrate the ranked distribution of the residuals. The capacitors show (in panel B) an increase in variability for the GPV scores whereas the inducers show a decrease of variability. To examine this phenomenon, we plot a member of the four categories in panel A. We see for the capacitors that the number of positive residuals is larger than the number of negative residuals. Oppositely, the number of negative residuals is larger than the number of positive residuals for the inducers. The two categories are opposite to each other. The capacitors are mutants capable of globally increasing phenotypic variability and the inducers decrease it. The maintainers are a special case of mutants. As one can see, they have an almost equally distributed number of positive and negative residuals, therefore either increasing or decreasing variability. Finally, the neutrals do not increase or decrease phenotypic variability due to the low values of their residuals. We recall that by looking at the GAPV and selecting the mutants, wild types and wild strains overtaking the median GAPV of the wild types, we want to see which of the aforementioned categories can at least show a high dispersion of variability.

Previously, we showed a high correlation between the GPV score and the phenotypic potentials from Levy & Siegal [76]. Here, we also have a high correlation ($r = 0.75$, $P <$

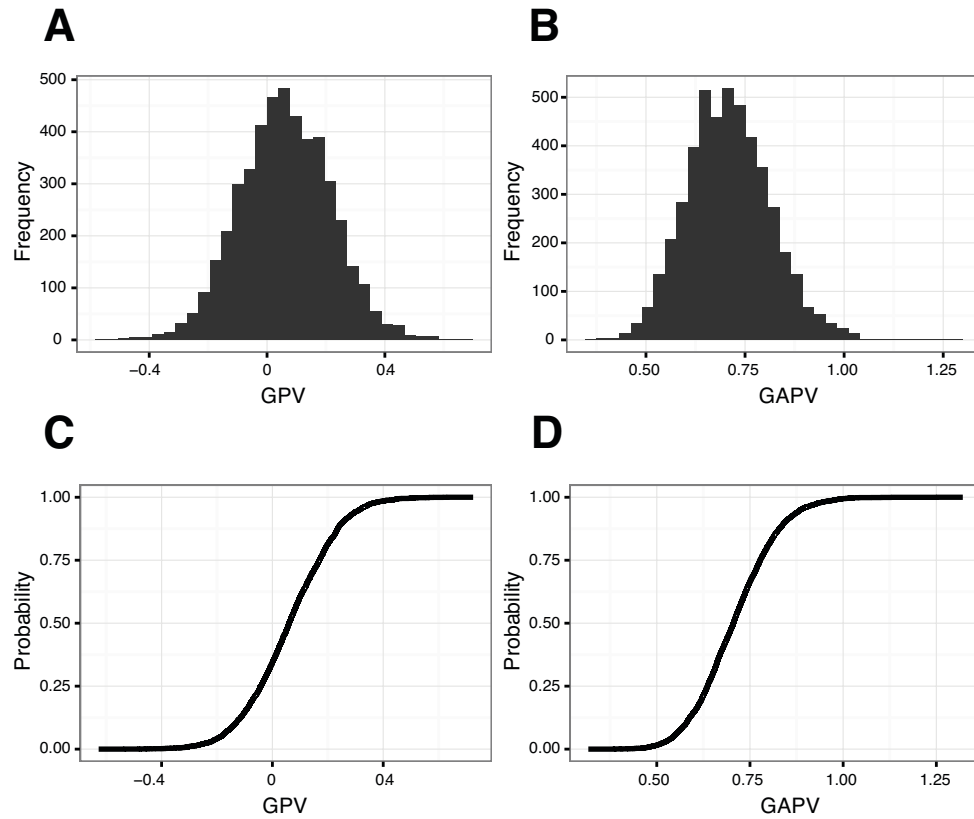


Figure 4.6: The segregation into categories of mutants affecting phenotypic variability is based on the analysis of the GPV and GAPV scores of the wild types. In panels A and B, the distributions of the GPV and the GAPV scores respectively, resemble a normal distribution centred respectively in zero and one. Therefore, we create the categories of mutants affecting phenotypic variability by considering the wild-type distributions and taking the outliers, i.e., the mutants located on the left and right tails of the wild-type GPV distribution. We compute the probability of having randomly chosen mutants falling onto the left and right tails of the GPV distribution. By permuting the wild-type GPV distribution 1000 times and looking at the average empirical cumulative distribution of GPV scores (in C and D). Here, the p-values for the lower and upper bounds are respectively $p_{\min} < 0.001$ and $p_{\max} < 1 - 0.96 = 0.03$.

0.001) between the phenotypic capacitors of Levy & Siegal and our capacitors. However, there is only an overlap of 70% between the two lists of capacitors. In addition, 30% of the phenotypic capacitors from Levy & Siegal are in fact maintainers in our study.

We see in Figure 4.7.B that the wild strains (labelled in green) are categorised as maintainers in our classification. This suggests that during the evolution of the wild strains (and

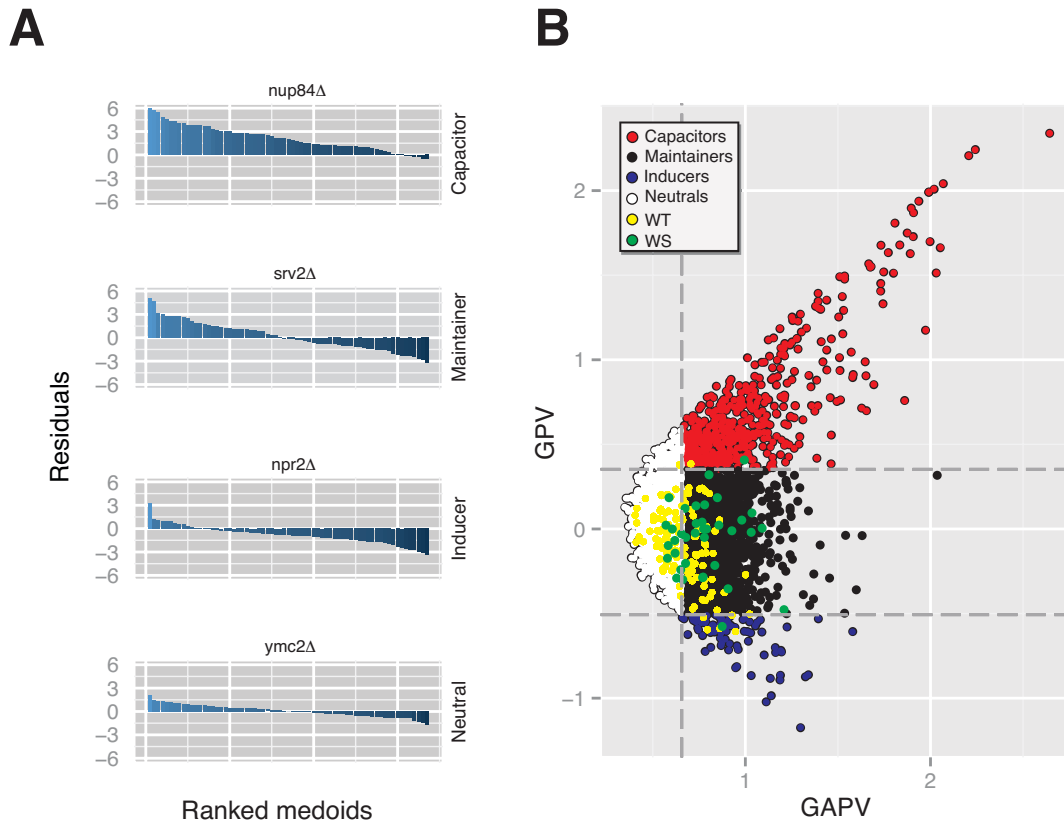


Figure 4.7: Ranked distributions of the residuals amongst the phenotypes and scatter plot of the GPV and GAPV scores. A. Bar plots of the ranked residuals, from the highest to the lowest, of the YKO's belonging in the four categories: capacitors, maintainers, inducers and neutrals. The capacitor shows a greater number of positive residuals than the inducer. The maintainer has an almost equal number of positive and negative residuals. Finally, the neutral does not show large positive or negative residuals. B Scatter plot of the GPV and GAPV. The dashed lines represent the separation between the four categories of capacitors, maintainers, inducers and neutrals. The red dots are the capacitors, the black dots are the maintainers, the blue dots are the inducers and the white dots the neutrals. In addition, the 37 wild strains are indicated by the green dots.

also for a majority of mutants), although phenotypic variability may be moved around, the phenotypic value is, in general, conserved.

Here, the large number of maintainers provides evidences for theories suggesting that a trade-off should exist in biological systems, between variability and robustness (Csete & Doyle [29], Kitano [58]).

The possibility for a mutant to increase or decrease its phenotypic variability has been recently studied by Weinberger et al. [148]. They modelled burst size by taking into account the mean and coefficient of variation of the distributions of proteins expression. In their analysis, they considered burst frequency as a factor allowing the regulation of mean expression and noise. With a careful screening of the chromatin related genes, they found complexes involved in noise regulation and more specifically in burst size and burst frequency regulation. The complexes are separated in two categories: the complexes increasing burst frequency and the complexes decreasing burst frequency.

In our study, we look at the complexes found by Weinberger et al. by calculating their respective GPV and GAPV scores. We then assess phenotypic variability of the complexes by their location on the GPV-GAPV space. Hence, we plot in Figure 4.8 the complexes and their relationship to one of the four categories. Here, we find similar results as in Weinberger et al. for the Swr1 and Rpd3(L). We see in our results that the Rpd3(L) genes are located in both the maintainer and inducer categories. This suggests that the Rpd3(L) complex can either increase or decrease global variability as it mainly belongs to the maintainer category. The second complex of interest is the Swr1 as it decreases burst frequency and therefore it increases phenotypic variability. Here, we see that the majority of the genes in the Swr1 complex are in the capacitor category. This observation agrees between the results of Weinberger et al. and our classification of knockout genes.

The other complexes are mainly located in the maintainer and neutral categories. However, the genes belonging in these complexes are in a few number in the SCMD, therefore not allowing to specifically determine their categories.

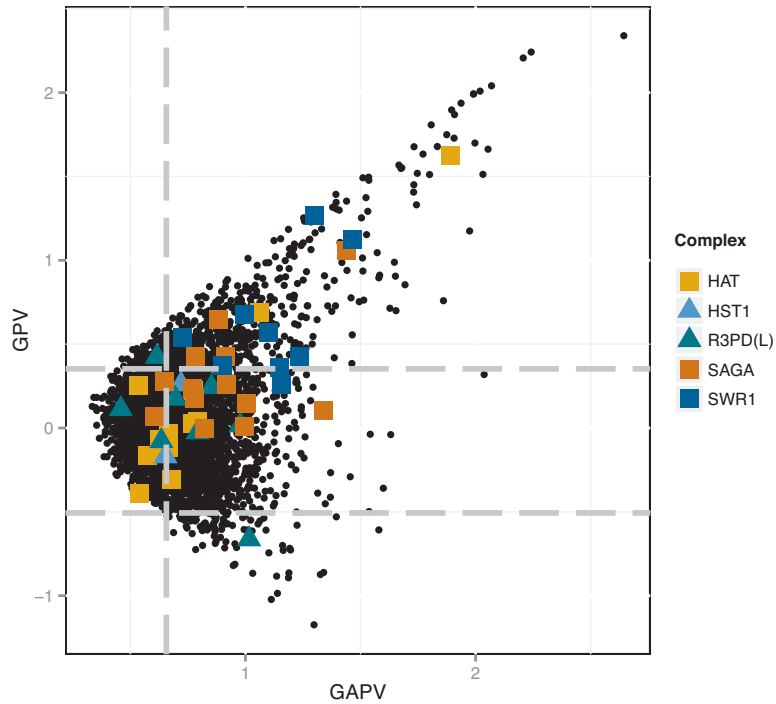


Figure 4.8: In their recent work, Weinberger et al. [148] found complexes increasing or decreasing burst frequency. Here, we plot these complexes in the GPV-GAPV space to visualise their position and assess their phenotypic variability. The complexes increasing variability in [148] are symbolised with a square. Conversely, complexes decreasing variability are indicated by a triangle symbol. Here, we see that the Swr1 complex is a member of the capacitor category, therefore indicating its capability to increase, as found in [148], phenotypic variability. We also show that the Rpd3(L) complex is shared between the maintainers and the inducers, thus decreasing variability.

4.4.2 Gene Ontology Annotations

We showed that the number of mutants in the four categories can range from 85 mutants for the inducers to 2206 for the maintainers. Here, we look at the Gene Ontology annotations of the four categories of mutants. To this end, we employ FunSpec (Robinson et al. [115]), a Gene Ontology search engine, which provides annotations for the genes being investigated. We use the FunSpec engine instead of the standard *Saccharaomyces* Genome Database (SGD, [19]) as FunSpec has an updated list of published analyses in order to find annotations, for instance, for the inducer category that possesses only a few number of mutants.

As the maintainer category has a large number of mutants, we do not describe its related GO annotations but we show them, as well as the annotations for the inducer and capacitor categories, in Figure 4.9.

There are only few annotations for the inducer category. Moreover, the number of mutants present in the category (labelled k) are usually in a low number compared to the total number of mutants in the annotation (labelled f). However the annotations do not provide specific information on phenotypic variability. The capacitor category shows more annotations when analysed with FunSpec. A large number of the annotations are related to DNA repair and DNA damage in general.

Inducers

GO Molecular Function	p-value	k	f
3'-5' exonuclease activity [GO:0008408]	0.003298	2	7
actin filament binding [GO:0051015]	0.008353	2	11
GO Biological Process			
proline transport [GO:0015824]	0.0001638	2	2
filamentous growth [GO:0030447]	0.001229	3	17
U5 snRNA 3'-end processing [GO:0034476]	0.00556	2	9
sphingolipid metabolic process [GO:0006665]	0.008353	2	11
GO Cellular Component			
vacuolar membrane [GO:0005774]	0.007364	6	134

Capacitors

GO Molecular Function	p-value	k	f
copper chaperone activity [GO:0016531]	0.0006922	3	3
minus-end-directed microtubule motor activity [GO:0008569]	0.0006922	3	3
tubulin binding [GO:0015631]	0.001717	4	7
telomeric DNA binding [GO:0042162]	0.001741	6	16
proton-transporting ATPase activity, rotational mechanism [GO:0046961]	0.0029	8	29
protein heterodimerization activity [GO:0046982]	0.003195	4	8
protein binding, bridging [GO:0030674]	0.004294	10	44
transcription coactivator activity [GO:0003713]	0.006202	6	20
ubiquitin-protein ligase activity [GO:0004842]	0.006993	14	77
hydrogen ion transmembrane transporter activity [GO:0015078]	0.007601	5	15
GO Biological Process			
response to DNA damage stimulus [GO:0006974]	4.37E-10	46	197
negative regulation of transposition, RNA-mediated [GO:0010526]	1.27E-06	7	9
mitotic sister chromatid cohesion [GO:0007064]	2.06E-06	11	24
chromatin remodeling [GO:0006338]	2.46E-06	16	49
double-strand break repair via single-strand annealing [GO:0045002]	3.06E-06	6	7
DNA repair [GO:0006281]	3.19E-06	36	183
telomere maintenance [GO:0000723]	9.00E-06	12	32
histone exchange [GO:0043486]	9.95E-06	7	11
transcription-coupled nucleotide-excision repair [GO:0006283]	2.21E-05	7	12
mRNA 3'-end processing [GO:0031124]	4.42E-05	7	13
GO Cellular Component			
Swr1 complex [GO:0000812]	4.42E-05	7	13
Ada2/Gcn5/Ada3 transcription activator complex [GO:0005671]	0.0002837	4	5
proton-transporting V-type ATPase, V0 domain [GO:0033179]	0.0002837	4	5
mitochondrial intermembrane space [GO:0005758]	0.0005112	11	40
telomerase holoenzyme complex [GO:0005697]	0.0006922	3	3
carboxy-terminal domain protein kinase complex [GO:0032806]	0.0006922	3	3
chromosome, telomeric region [GO:0000781]	0.001647	10	39
prefoldin complex [GO:0016272]	0.001717	4	7
vacuolar proton-transporting V-type ATPase, V0 domain [GO:0000220]	0.001717	4	7
proton-transporting two-sector ATPase complex, proton-transporting domain [GO:0033177]	0.001717	4	7

Maintainers

GO Molecular Function	p-value	k	f	GO Biological Process	p-value	k	f
structural constituent of ribosome [GO:0003735]	0.000131	98	218	metabolic process [GO:0008152]	3.30E-05	179	425
peptidyl-prolyl cis-trans isomerase activity [GO:0003755]	0.0017	12	17	fatty acid beta-oxidation [GO:0006635]	0.0003269	9	10
catalytic activity [GO:0003824]	0.005066	176	455	histone deacetylation [GO:0016575]	0.001352	17	27
glutathione peroxidase activity [GO:0004602]	0.006508	6	7	protein peptidyl-prolyl isomerization [GO:0000413]	0.0017	12	17
Rab GTPase activator activity [GO:0005097]	0.008273	8	11	post-Golgi vesicle-mediated transport [GO:0006892]	0.003929	5	5
isomerase activity [GO:0016853]	0.008422	32	67	mitochondrial genome maintenance [GO:0000002]	0.004369	20	36
GO Cellular Component				histone methylation [GO:0016571]	0.006508	6	7
ribosome [GO:0005840]	2.41E-06	751	2026	phosphatidylcholine biosynthetic process [GO:0006656]	0.007811	7	9
cytoplasm [GO:0005737]	0.000118	10	11	phospholipid metabolic process [GO:0006644]	0.007811	7	9
peroxisomal matrix [GO:0005782]	0.001153	398	1072	nucleoside metabolic process [GO:0009116]	0.008224	9	13
mitochondrion [GO:0005739]	0.001833	32	62				
cytosolic small ribosomal subunit [GO:0022627]	0.004379	23	43				
mitochondrial large ribosomal subunit [GO:0005762]	0.004422	150	381				

Figure 4.9: We look at the nature of the mutants belonging to the inducer, maintainer and capacitor categories by performing a Gene Ontology annotations analysis. The annotations are retrieved using the FunSpec [115] engine. The inducers do not show known and relevant annotations for phenotypic variability. However, the capacitors have multiple annotations, in general, related to DNA damage.

4.4.3 Analysis of Robustness and Phenotypic Variability

In the previous parts of this section, we created categories of mutants, according to their effect on phenotypic variability. Here we look at the robustness of the three main categories: the inducer, the maintainer and the capacitors, on the fitness of the cell. In our analysis, we define the fitness of a cell as its growth rate, i.e., the increase of size by unit of time.

To explain our interest in studying robustness and fitness of the cell using phenotypic variability, we employ the analogy between a formula one car and a tank. The tank is robust in its environment, meaning that its structure will resist if the tank drives into a wall. However, the tank will operate slowly due to its heavy weight (here considered as a "phenotype"). The formula one car, unlike the tank, will go fast but will certainly be destroyed if it encounters a wall or an obstacle on the road. Here, we investigate this analogy for the robustness of phenotypes and the fitness of the cell. We want to see if robust phenotypes have a better fitness or rather the opposite.

Cell fitness is acquired here by extracting the growth rates of two published databases and one database created by our collaborator Michael Cook at the University of Toronto. To ensure qualitative results and be sure that the growth rates are consistent with the experiments performed by Ohya et al., for the same set of mutants, we choose two databases which proceeded with the same experimental settings as in the SCMD, using the same medium (YPD here) and the same mutants. The two databases are: the *PROfiling of PHE-notypic Characteristics in Yeast* database [38] (referred to as *Prophecy* in this study) and the growth rate data collected by Qian et al. [109] (referred to as *GL* in this study). Our collaborator also followed the same experimental conditions for the acquisition of the growth rates. However, the unpublished growth rates dataset only contains 384 measurements.

The three databases and their respective growth rates are plotted in Figure 4.10. Here the GL growth rates are normalised by the wild-type growth rate. The units are in terms of optical density per hour. The surfaces representing the growth rates according to the GPV and GAPV values are smoothed using a two-dimensional Lowess regression (with a span

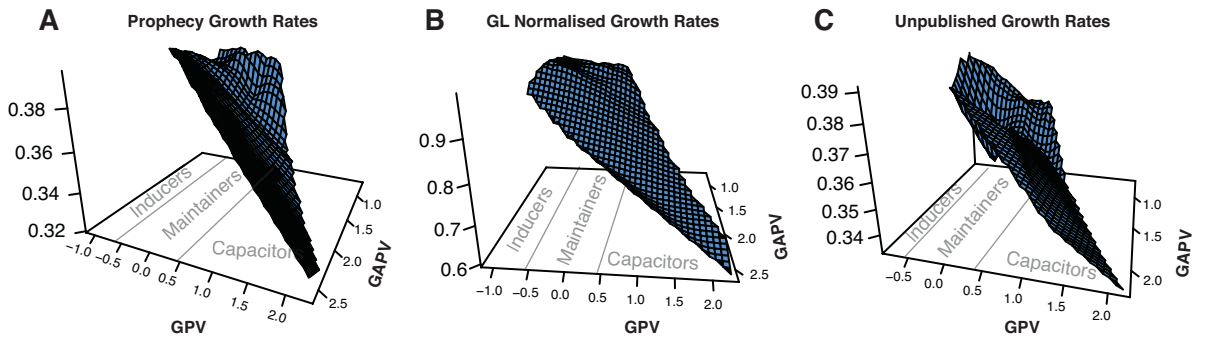


Figure 4.10: The mutants increasing variability, i.e., the mutants, are in terms of robustness, less robust than the inducers which decrease variability. Here, robustness is compared to fitness, i.e., the growth rate of a mutant, by plotting both and using a two-dimensional Lowess regression to smooth the surface. For this purpose, three databases of growth rates are plotted in the three panels. A The Prophecy database. B. The GL database and C. the unpublished database provided by our collaborator at the University of Toronto. In all three databases, we see the same trend. The inducers have a higher growth rate than the capacitors. In between, the maintainers seem to create a transition between the inducer and capacitor categories. Here, we observe a contradicting observation to our analogy between robustness and fitness. Inducers, more robust show a high growth rate and capacitors, less robust, show a lower growth rate.

of 0.2). Panel A represents the Prophecy experiment, panel B the GL experiment and panel C the unpublished experiment. Although the growth rate shape extracted from the unpublished database seems reduced due to the low number of available data points, all three shapes show the same trend. The inducers have a higher growth rate than the capacitors and, in between, the maintainers seem to function as an intermediary.

This observation goes against our analogy. The inducers, more robust to phenotypic variability seem to grow faster than the capacitors, which are less robust to phenotypic variability.

4.5 Control of the Methodology

Our methodology, based on the calculation of the residuals with the coefficients of variation of the clustered morphological phenotypes, could be biased by the regression. As opposed to Levy & Siegal [76] who calculated the residuals with a Lowess regression, our methodology employed the $L1$ -regression to capture mutant outliers. Although we showed that the scores using the $L1$ -regression are well correlated with other genetic measures of phenotypic variability, here we re-calculate the residuals with the average coefficient of variation of the wild types. In the previous section, the $L1$ -regression followed the mutants and wild-type trends. Now, we do not regard variability by its trend but as a difference of variation between the mutants and the wild types.

Here, the new matrix of residuals R is

$$R = (R)_{ij} = \log_2 \left(\frac{o_{ij}^m}{\langle cv_j^w \rangle} \right) \times \left[\text{SD} \left(\log_2 \left(\frac{o_{kj}^w}{\langle cv_j^w \rangle} \right) \right) \right]^{-1}, \quad (4.5)$$

$$i = 1, \dots, 4681, k = 1, \dots, 122, j = 1, \dots, 67,$$

where $\langle cv_j^w \rangle$ is the average value of the wild-type coefficients of variation, for the phenotype j . Here, the methodology is very similar as in Eq. (4.2). The difference comes from the ratio computed with the wild types. Here, instead of employing the distance between the data and the fit, we look at the average coefficient of variation across all wild types for a specific phenotype.

With the new residuals, we compute again the GPV and GAPV scores as in §4.4.1. The new GPV and GAPV scores based on the coefficient of variation show high correlations respectively ($r = 0.88, P < 0.001$) and $r = 0.83, P < 0.001$) with the original scores calculated using the $L1$ -regression.

We also compare the new scores with the phenotypic potentials devised by Levy & Siegal. As previously seen with the original scores, the new GPV and GAPV show high correlations with the phenotypic potentials as well ($r = 0.71, P < 0.001$ and respectively

$r = 0.57, P < 0.001$).

In the optic of analysing the Gene Ontology annotations of the new lists of inducers, maintainers and capacitors, we compare the number of mutants shared between the scores created with the $L1$ -regression and the scores created with the coefficient of variation. The shared proportions are showed in Table 4.4. For instance, 62% of the capacitors for the $L1$ -regression are shared with the capacitors from the CV analysis. Although both methodologies present a large number of common mutants, the GO annotations (see Appendix B.3) are slightly different. Again, the inducers categories shows only few annotations. The annotations for the capacitors category are also related to DNA damage as found previously.

Category \ Residuals	$L1$	CV
Capacitors	0.62	0.77
Maintainers	0.75	0.69
Inducers	0.42	0.53
Neutrals	0.69	0.71

Table 4.4: Proportion of shared ORFs amongst each category. The proportions indicate what is the percentage of ORFs that are shared between both analyses amongst the number of ORFs in each category.

We also investigate the effect of choosing the average coefficient of variation instead of the $L1$ -regression on the growth rates of the mutants. Similar trends as with the $L1$ -regression are observed (see Figure B.1 in appendix). The growth rates are higher for the inducers and slowly decrease in the maintainer and capacitor sections of the plot.

4.6 Discussion

A previous work from Lehner [71] demonstrated that mutants confer similar robustness to environmental, stochastic and genetic perturbations. In our study, we pursued this idea by considering robustness at the individual morphological level, suggesting that the most malleable phenotypes under genetic modifications, i.e., the phenotypes subject to change often, are also the most variable due to stochastic effects (see Ho & Jianzhi [48]). Such related observations have already been suggested to explain the relationship between gene expression noise and the malleability of phenotypes under environmental and genetic modifications (Pancaldi et al. [102]).

We looked at the stochastic and genetic variability in different environments to see whether a phenotype keeps its robustness or not. Here, we showed that genetic variability, seen as the coefficient of variation calculated on the mean phenotypic values of the mutants, is well correlated ($r = 0.59$, $P < 0.001$) with the stochastic variability, seen as the average of the wild-type coefficients of variation. Similarly, we also showed a high correlation ($r = 0.67$, $P < 0.001$) between the genetic variability of the 37 wild strains and the stochastic variability of the wild types. The high correlations indicate that small genetic variations produce small stochastic variations for a given phenotype.

To understand the nature of phenotypic variability at a global level, we looked at the residuals of the coefficients of variation of the mutants and wild strains. To account for technical error and also give a fold change in the residuals, we included the wild-type coefficients of variation in the calculation. In this setting, positive residuals greater than one suggested that the mutants significantly increased variability. Similarly, negative residuals less than minus one decreased variability.

Next, we investigated the residuals of the mutants at a global level by creating a score, the Global Phenotypic Variability (GPV) capable of attributing the mutants a quantitative measure of their effect on phenotypic variability. We validated the score by comparing it with existing measures of variability, such as the phenotypic potentials of Levy & Siegal

[76] and showed a high correlation ($r = 0.75$, $P < 0.001$) between the two measures. In addition and to look at the spread of the GPVs, we also considered the absolute median deviation of the residuals as a second score.

Using both scores, we segregated the mutants in four categories according to their effect on phenotypic variability by using the wild-type distributions of GPV and GAPV. Here the four categories are: phenotypic inducers which are mutants decreasing phenotypic variability, phenotypic capacitors which increase variability, phenotypic maintainers that act as both inducers and capacitors by maintaining the level of phenotypic variability and finally the neutral mutants, not significantly affecting variability. The majority of the mutants are in the neutral category but 85 mutants are found to be inducers, 586 capacitors and 2206 maintainers. The Gene Ontology annotations provided for the inducers only indicated few terms, not related to phenotypic variability. However, the annotations for the capacitors were related to DNA damage. In addition, we confounded the complexes found by Weinberger et al. [148] and showed that the Rpd3(L) complex acts as a maintainer and inducer and the Swr1 complex is a capacitor.

Our results on fitness showed a down trend from the inducers category to the capacitor category. In all three sets of growth rates, the inducers showed higher growth rates than the capacitors. This observation went against our idea that robustness is inversely related to fitness.

Finally, as an additional analysis and control, we calculated the residuals with another regression method. Instead of the $L1$ -regression, we employed the average coefficient of variation of the wild types. We showed that even with another measure of the residuals, it is still possible to find similar results and categories of mutants acting on phenotypic variability.

Chapter 5

Extraction and Analysis of Haploid Genes Modifying Phenotypic Variability and Cell Cycle Regulation in *S. cerevisiae*

5.1 Introduction

In Chapter 4, we introduced the notion of capacitor and inducer genes, capable of creating or suppressing variability in certain morphological phenotypes. The question that was raised was not only to understand why gene deletions can produce more or less variability for some phenotypes but also to see if this phenotypic variability affected the robustness and fitness of *S. cerevisiae* cells. In addition, we wanted to see if phenotypic variability provides meaningful and predictive insight into gene functions.

Understanding the mechanisms underlying cell size and more generally cell cycle is still unclear and subject to debate. A first insight was given by Wilson [149] who discovered that ploidy, the number of sets of chromosomes, was correlated with cell volume. Such observation was a fundamental starting point in the study of cell size control and cell cycle in all organisms. To illustrate the importance in understanding the cell cycle of unicellular

organisms, Jorgensen & Tyers [55] relate cytokinesis, the division of the cytoplasm of the mother cell into two daughters, to the cell mass. For example, during the growth of *S. cerevisiae*, cytokinesis is asymmetric with respect to the mass of the cell. In other words, at division, the mother cell is larger than the daughter cell.

Currently, and to our best knowledge, no studies have been carried out to examine the cell cycle by considering variability in the morphological phenotypes. Usually, current approaches either do not consider the complete genome of *S. cerevisiae* and therefore miss some deletion strains or the methods employed lack the investigation of the variability in morphological phenotypes. In our research we decided to examine the cell cycle and cell size of *S. cerevisiae* under genetic mutations by looking at variability in morphological phenotypes essentially related to cell and nuclear sizes. To this end, we employ as in Chapter 4 the *Saccharomyces Cerevisiae* Morphological Database, which provides both mean and variability of the phenotypes.

As one can see in Figure 5.1.A, it is possible to find mutants (here *mdh3* Δ , in green) with similar mean (of the mother cell size) to a wild-type strain (indicated in red). Strictly considering any statistical tests based on the mean value of the samples, the tests would suggest that the wild types and the mutants are in this case similar. However, their variances are different. If considering mean values alone, the *mdh3* Δ deletion strain is indistinguishable from the wild type; in contrast, as one can clearly see in Figure 5.1.A, the variances of the two strains are different. The variance of *mdh3* Δ is smaller for the mother cell, indicating an alteration in cell size coherence, albeit in the absence of a change in mean size. We see that variability is in this case useful to visualise the difference between two strains. However, simply considering variability would lead to erroneous interpretations, as for example, with the *avt5* Δ mutant. Although *avt5* Δ shows a similar variance as the wild type, its mean value, or mode, is shifted to the right compared to the wild type. Therefore, we employ both the mean and the standard deviation or coefficient of variation (CV) to analyse morphological and genetic properties of the mutants.

The mean and variability in mother cell size is shown in Figure 5.1.B for the complete

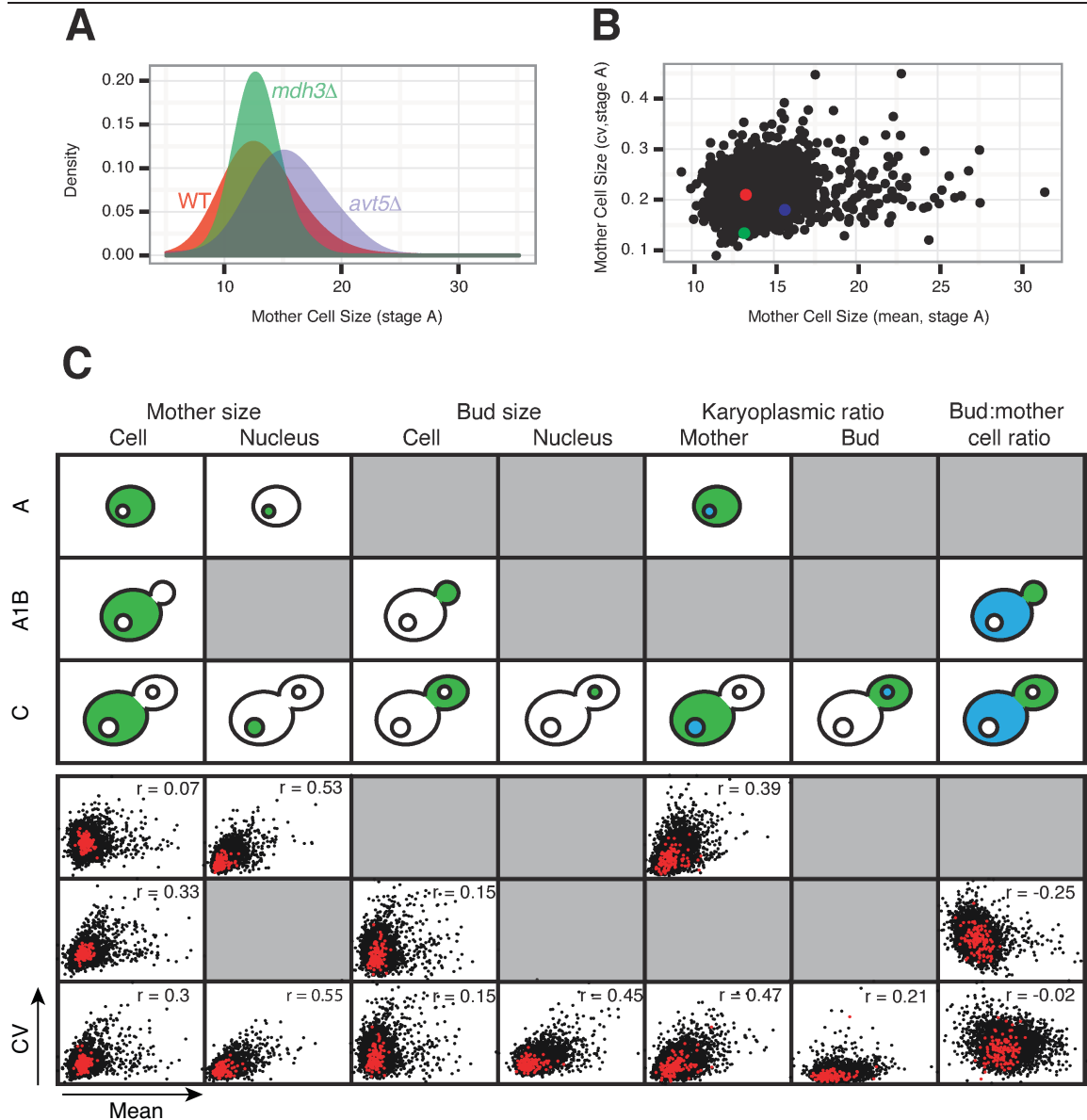


Figure 5.1: Phenotype variability can be genetically manipulated separately from phenotypic mean across many phenotypes. A. Yeast deletion mutants modify the mean or coefficient of variation (CV) of cell size in unbudded cells (stage A), separately or in combination. Wild type (*his3Δ*) and indicated in red and the mutants are indicated in green and purple. B. Overlay of the size and variability of mutants in (A) on a scatter plot for the haploid nonessential deletion collection. C. Phenotypes considered in this study. Green (and blue) indicate the cellular parameter(s) described in the scatter plots below for *his3Δ* (red) and deletion mutants (black). Correlation coefficients are indicated for each plot of CV (y-axis) versus mean (x-axis).

set of haploid non-essential yeast deletion mutants (SCMD, [99]). Again, the *mdh3* Δ and *avt5* Δ mutants and the wild type are indicated on the plot. As in the case of the prior example, some mutants have a large mean but a small coefficient of variation and vice versa. In this example, and in further examples provided below, mean and variability in morphological traits provide distinct information.

In this study, we are interested in finding groups of genes modifying the cell structure and more specifically cell size throughout the cell cycle. We want to determine whether some groups of genes are more subject to modify the cell structure during certain phases of the cell cycle and analyse their nature by looking at their Gene Ontology (GO) annotations. To this end, we look at phenotypes strictly related to the cell size and nuclear size, as shown in Figure 5.1.C. We employ two phenotypes related to the mother cell size and the bud cell size, as well as the karyoplasmic ratio, i.e., the ratio of the nuclear volume by the cytoplasmic volume, and the ratio of bud cell size to mother cell size. For each phenotype or ratio, we look at the mean phenotypic value, i.e., the mean value computed on single-cell measurements for a specific phenotype, and the coefficient of variation (CV) value, also calculated with single-cell measurements. The bottom table in panel C shows the mean phenotypic and CV values for the different phenotypes across the three cell cycle stages. As described in Chapter 4, the stages have been divided by the SCMD as: stage A (G1 phase), A1B (S and G2 phases) and finally stage C (late M phase).

The mean phenotypic values and the coefficients of variation of the mother cell size throughout the cell cycle stages are plotted in Figure 5.2. The left panel shows the evolution of the mean phenotypic cell size from stage A to A1B, stage A1B to C and from stage A to stage C. In all three comparisons, the mean phenotypic value of the mother cell size is consistent throughout the cell cycle. The Pearson's coefficients of correlation are $r = 0.93$, $P < 0.001$ from stage A to A1B, $r = 0.96$, $P < 0.001$ from A1B to C and $r = 0.92$, $P < 0.001$ from A to C. The correlations not only indicate that mother cell size increases throughout the cell cycle but also that, for most mutants, mean cell size is not perturbed as a function of cell cycle position. On the right panel of Figure 5.2, the coefficient of variation of the mother cell size is also plotted across the cell cycle. Here,

the Pearson's coefficients of correlation are: $r = 0.61$, $P < 0.001$ from stage A to A1B, $r = 0.72$, $P < 0.001$ from A1B to C and $r = 0.56$, $P < 0.001$ from A to C. In these three new comparisons we see that variability no longer follows the same linear trends as for the mean phenotypic cell size. Particularly, variability does not seem to be reported (or passed) from one cell cycle stage to another, depending on the mutants. Therefore, we employ phenotypic variability in particular cell cycle stages to provide information on gene function in that cell cycle stage.

In this chapter, we devise two methods to analyse the mutants causing variability in the cell cycle of budding yeast. The first method considers the mean phenotypic value and the CV of the mutants separately. We show that by simply looking at either the mean phenotypic value or the CV of certain cell and nuclear size phenotypes, it is already possible to find outlier mutants, compared to the wild types and thereafter relate their associated enriched GO annotations. However, we show that this method lacks statistical significance as the p-values obtained from hypergeometric tests on the GO annotations are not significant enough to be certain that the genes are effectively related to the GO annotations. To cope with this problem and allow a better statistical significance, we devise a second method that considers both the mean phenotypic value and the coefficient of variation of the mutants. Although both methods are different in their analyses of the GO annotations (see details in the next sections), the second method yields more significant p-values and still supports insights provided by the first method.

Our analysis is structured around the two extraction methods. The first section describes the morphological phenotypes analysed in this study and the creation of the scores employed by the first method. Results for the first method are also described at the end of the section and the significant GO annotations found are reported. The second section continues the analysis of the effect of gene deletions on phenotypic variability but this time utilising the second method based on both the mean phenotypic value and the coefficient of variation. Finally, the last section of this chapter summarises and concludes our analysis.

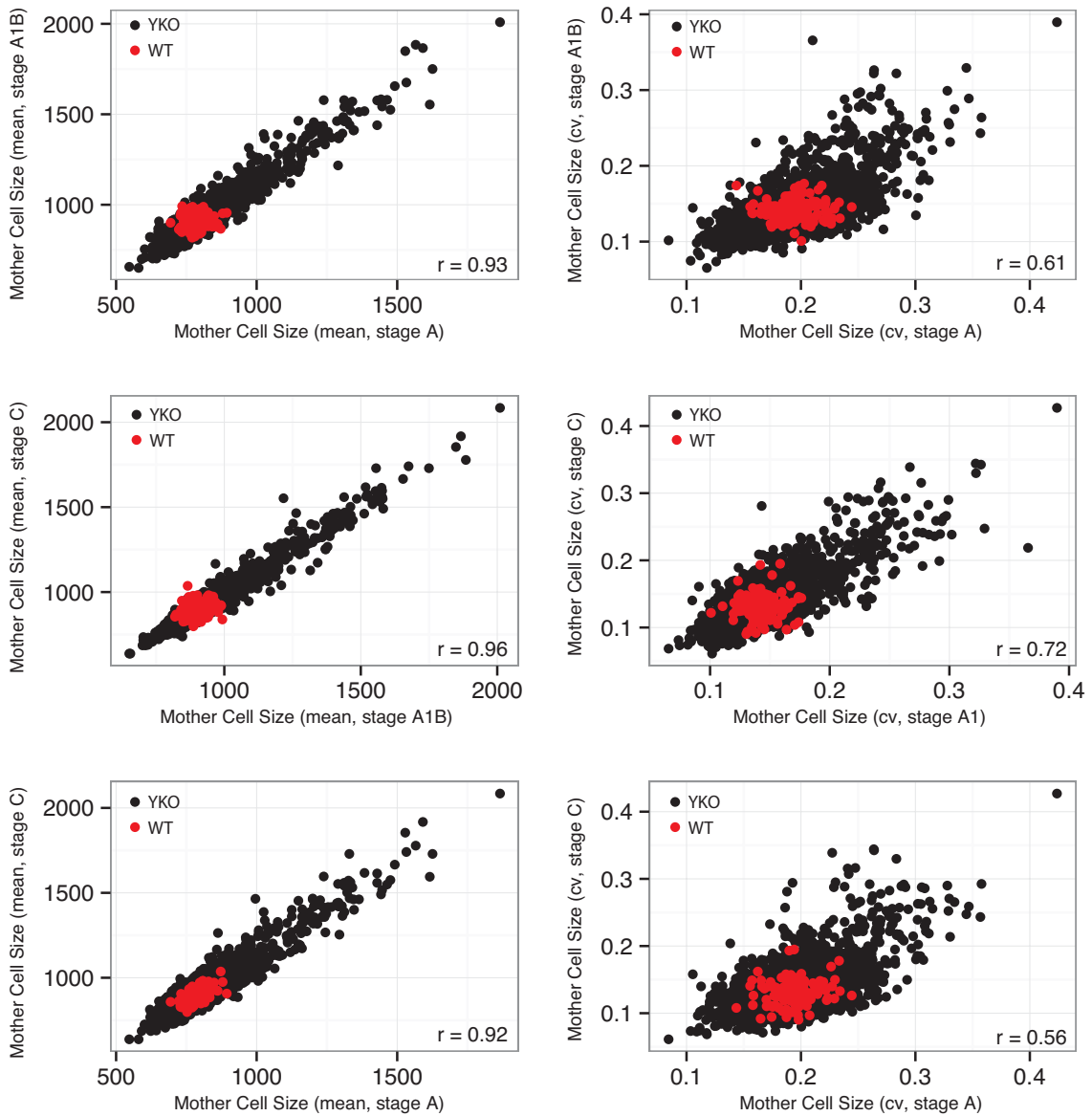


Figure 5.2: In contrast to mean size which keeps a high correlation between the cell cycle stage, variability in cell size provides additional information as a function of cell cycle position. All pair-wise scatter plots of cell size (left) and cell size variability (right) are shown for unbudded cells in G1 (stage A), pre-mitotic budded cells in S/G2/early M (stage A1B), and post-mitotic budded cells in late M phase (stage C). Wild-type control replicates are indicated (red). The correlations (Pearson's r) indicated on each panel are all significant ($P < 0.001$).

5.2 Gene Selection Based on Mean Distance and Residuals of the Wild types

In this section we investigate variability in the cell cycle of budding yeast by considering variability as the coefficient of variation of cell and nucleus related phenotypes.

We first motivate the analysis by explaining the type of phenotypes and how they are extracted from the SCMD. Our analysis is based on the phenotypes provided by the *Saccharomyces Cerevisiae* Morphological Database, which contains wild types (*his3* strain) and a collection of 4718 mutants. As in Chapter 4, the mutants dataset is filtered to remove gene deletion strains with less than 25 measurements to avoid poor quality estimates of variance in the analysis of phenotypic distributions. Although the entire SCMD contains 501 morphological phenotypes, we are only interested in the phenotypes related to cell size and nuclear size. In Figure 5.1.C, we depicted these phenotypes and the scatter plots corresponding to the mean phenotypic values and their respective coefficient of variations. The phenotypes analysed in this study as well as their descriptions are listed in Table 5.1.

The karyoplasmic ratio phenotypes (for the mother and the bud) are, by default, not available in the SCMD, due to the image acquisition process not being able to record the volume of the nucleus and the cell. Instead, we employ the raw data of the SCMD based on the single-cell measurements of the strains to compute a variant of the karyoplasmic ratio by considering the area, in pixels, of the nucleus and the cell. The same computation is carried out for the bud karyoplasmic ratio.

To address the question of genetic functionality between genes identified in our analyses, we regroup the mutants and analyse them as part as their gene ontology annotations (GO). To this end, we harness the *Saccharomyces* Genome Database (SGD,[19]) for our study.

Although the SCMD contains most of the genes that are nonessential in haploid cells with 4718 of the current 6275 discovered genes, we shrink the GO annotations to the gene

Phenotype	Description	Stage
C11-1_A	Mother cell size in nucleus	A
D14.1_A	Area of nucleus region in mother cell in nucleus	A
C11.1_A1B	Mother cell size in nucleus	A1B
C11.1_C	Mother cell size in nucleus	C
D14.1_C	Area of nucleus region in mother cell in nucleus	C
KR_A	Karyoplasmic ratio	A
KR_A1B	Karyoplasmic ratio	A1B
KR_C	Karyoplasmic ratio	C
BC_A1B	Bud cell size to mother cell size ratio	A1B
BC_C	Bud cell size to mother cell size ratio	C
C11.2_A1B	Area of daughter cell on nucleus	A1B
D14.2_C	Area of nucleus region in bud in nucleus	C
BKR_C	Bud karyoplasmic ratio	C
B_C	Bud cell size	C

Table 5.1: The SCMD provides phenotypes related to cell size, nuclear size and actin patches. For this analysis we are interested in 14 phenotypes, related to cell and nuclear sizes. Each phenotype possesses two values : the mean phenotypic value and a coefficient of variation. The mean phenotypic value corresponds to the mean value of the phenotype, computed on multiple single-cell measurements. The coefficient of variation is the standard deviation divided by the mean phenotypic value, computed using the same single-cell measurements. Here we show the phenotypes for their mean phenotypic values label.

deletion strains processed by the SCMD.

We then look at individual mutants and pairs (mean phenotypic value and CV) of phenotypes related to nucleus, cell and bud size. For each phenotype, we treat the mean phenotypic value and its corresponding coefficient of variation independently. Firstly, we devise a method to compare the mean phenotypic value of the mutants and the wild types for the same phenotype. Secondly, we devise a method that employs the wild types as a reference to compute residuals for the mutants.

Both separation methods provide mutants with scores for all the 14 phenotypes examined. The scores are separated in two categories: the score related to the mean difference (or called here mean distance) of the mutants compared to the wild types and the score based on the residuals obtained using the mean and coefficient of variation values of the mutants. The later score is exactly defined as in Chapter 4. The scores are qualitative measures of the dispersion of the mutants in both the mean and coefficient of variation values. At this

stage, it should also be mentioned that it is also possible to perform hypothesis testing on the mean phenotypic value and coefficient of variation separately. We have not performed such tests here as we want to consider the mean and coefficient of variation together and not independently.

By their quantitative nature, the two scores obtained from the two separation methods can be sorted to obtain mutants with low and high mean or coefficient of variation values. After sorting the genes in an increasing order (according to their scores), we select the bottom 50 and top 50 mutants to proceed with a Gene Ontology (SGD) analysis coupled with a Genemania [146] mapping. This first selection of 50 mutants is performed to give a first idea of what is happening for the mutants with high variability. It must be noted that this arbitrary value of 50 could definitely be extended to another value. However, here we are interested in a first time in selecting mutants with high phenotypic variability and seeing if we already obtain indications on the GO annotations.

The structure of this section is divided in four parts. The first part introduces the mean distance method that allows computing the dispersion of the mutants from the wild types when the mean phenotypic value is considered. The second part recalls the methodology already defined in Chapter 4 to compute the residuals of the mutants. The third part is dedicated to the permutation tests run to determine the validity of the methods to find GO annotations. Finally, the last part presents the results obtained with the two selection methods.

5.2.1 Mean Distance Method

As seen in Figure 5.1.B, mutants show a wide dispersion in their mean phenotypic value, i.e., the mean value of the single-cell measurements for a particular phenotype. Here, we want to assess this dispersion by comparing the mean phenotypic mutant value with the mean phenotypic wild-type value.

We recall that the wild types have 122 replicates of the mean phenotypic value for a particular phenotype. We denote $\bar{w}^j$, the average mean phenotypic value, calculated using the 122 replicates of a particular phenotype j , i.e.,

$$\bar{w}^j = \frac{1}{122} \sum_{i=1}^{122} w_i^j \quad (5.1)$$

where w_i^j is the wild-type mean phenotypic value of one biological replicate. We define m_i^j as the mean phenotypic value of mutant i for the phenotype j , i.e., its original mean phenotypic value.

In its essence, the mean distance score determines the spread of the mutants by computing the log-scaled difference of the mutants and wild types for their mean phenotypic values. To this end, we divide the mutant mean phenotypic m_i^j value by the average wild-type value $\bar{w}^j$ and we take the logarithm of this ratio to express a fold change. In addition, for all the wild-type replicates, denoted w_k (with k ranging from 1 to 122), we also measure the spread of the measurements by normalising the wild types by their average value and taking the standard deviation.

Finally, we define the score d_i^j for the mutant i and the phenotype j by calculating the ratio of the logarithm of the scaled mutants ratio by the standard deviation of the logarithm of the wild types to obtain a dispersion of the fold change compared to the wild-type fold change. The score d_i^j gives both the spread of the mutants compared to the wild types and the fold-change. The full equation for the score is given in Eq.(5.2).

$$d_i^j := \frac{\log_2 \left(\frac{m_i^j}{\bar{w}^j} \right)}{\text{SD} \left(\log_2 \left(\frac{w_k}{\bar{w}^j} \right) \right)} , \quad i = 1, \dots, 4681, j = 1, \dots, 14, k = 1, \dots, 122. \quad (5.2)$$

5.2.2 Residuals Based on the Wild Types

In the mean distance method, we looked at the spread of the mutants for the mean phenotypic value. Here, we are interested in the spread of the mutants for their phenotypic

variability. To this end, variability of the mutants is measured by comparing the coefficient of variation of the mutants with the coefficient of variation of the wild types.

We employ the exact same calculation as in Chapter 4, for the inducer and capacitor genes, by calculating the residuals coming from the L1-regression.

Once the L1-regression determined, we compare the fitted values of the mutants and wild types, i.e., f_{ij}^m for mutant i and phenotype j and f_{kj}^w for the wild type k and phenotype j , with the original mutant and wild-type values, denoted o_{ij}^m and o_{kj}^w for respectively the mutant i and wild type k for the phenotype j . The process is illustrated in Figure 4.3 of Chapter 4.

The residual score r_i , for mutant i , is then defined as in Eq.(5.3).

$$r_i := \log_2 \left(\frac{o_{ij}^m}{f_{ij}^m} \right) \times \left[\text{SD} \left(\log_2 \left(\frac{o_{kj}^w}{f_{kj}^w} \right) \right) \right]^{-1}, \quad i = 1, \dots, 4681, j = 1, \dots, 122, j = 1, \dots, 14. \quad (5.3)$$

5.2.3 Permutation Tests

The Gene Ontology analysis of the genes, based on their mean distance and residual scores, is performed by ranking the scores, for both categories, and picking the bottom and the top ranked 50 genes. The process is repeated for all the phenotypes employed in our study. Thereafter, we look at the associated gene ontology to draw conclusions on the nature of the bottom and top 50 genes.

Although the gene picking process seems simple and adequate for this study, it has some restrictions due to the nature of the ontology associations. The gene ontology can contain up to a thousand gene associations. Therefore, due to the small number of genes we select (50 here), it might happen that the ontologies represented by these genes may be due by chance.

To address this issue, for each phenotype analysed here, we devise a hypergeometric test to qualitatively determine the annotations found for the bottom and top 50 genes. The hypergeometric test measures the statistical significance of having a sample of s selected genes (amongst the 50) in a total of f genes (in the GO annotation) for which k genes should belong to the top/bottom 50 over the total number n of genes in the entire ontology. Instead of looking at the whole gene ontology, we employ the slim gene ontology. Hence, for the mean distance and residual scores, we order the scores from the smallest to the highest value to select the bottom and top 50 genes. For each GO slim annotation, we compute the hypergeometric p-value of having the bottom or top 50 genes within the annotation. For this, we find the number k of genes (amongst the bottom or top 50) that belong to a particular GO Slim category containing f genes in total and we compute the p-value as in the standard hypergeometric test. Thereafter, to compare the GO Slim an-

Algorithm : Hypergeometric Test on the Gene Ontology Annotations

1. Order the ORFs for the residuals.
2. Take the top 50 ORFs.
3. For the GO slim annotation "i", compute :
 - k : number of the top 50 ORFs belonging to GO Slim term i
 - f : total number of genes in GO Slim term i
 - s : number of top/bottom genes selected, in this case 50
 - n : number of genes in the SCMD
 - p : the p-value of the hypergeometric test
4. End For.
5. Repeat 1:4 for bottom 50 ORFs.
6. Repeat 1:5 across phenotypes.
7. Repeat 1:6 for mean values, instead of residuals.

Algorithm 5.2: Due to the small number of genes selected in the analysis (only 50), the process of finding the GO annotations might be biased and therefore not provide any significant information. To verify the accuracy of finding a GO annotation, we employ a hypergeometric test on the Slim GO annotations to see whether they are significant or not.

notations, we scale the p-values found for each annotation by taking their logarithm and plotting a clustered heatmap of the log-p-values. Here the heatmap is clustered by looking at the Euclidian distance of the average p-values.

The heatmap of the log-p-values, allows a visual inspection of the "chance" of really seeing the bottom or top 50 genes belonging to a particular GO Slim annotation. In addition, it shows whether some annotations are biologically related to each other.

The complete procedure is described in Algorithm 5.2.

5.2.4 Results

The ranked mean distance and residual scores for the bud nuclear size in stage C are shown in Figure 5.4.A. The black dots represent the mutants and the red dots the wild types. We see that for the mean distance score, the wild types are spread along the ranks almost as the mutants. Therefore, employing the mean phenotypic value is not necessarily the best method to extract information about the mutants. In contrary, the residual scores for the mutants show a clear difference compared to the wild types. As one can see, the wild types are located in the middle of the ranked plot, letting the mutants range from low to high variability. Therefore, the use of the mean distance or residuals gives different information on the gene deletion strains.

In panel B of Figure 5.4, we illustrate the physical interaction network of the mean distance and the residual scores. Here, we combine the top 50 genes from the ranked mean distance and residuals with 50 additional genes found to have physical interaction with the top 50 genes, by the Genemania engine. Genemania uses a large association dataset comprising pathways, protein and genetic interactions, protein domain similarity and co-localisation. In addition to the physical gene interaction networks, we plot in Figure 5.4.C a word cloud from words within the enriched GO annotations. The words related to the mean distance score are coloured in green and the words related to the residual score are

coloured in red. In general, the variability associated GO annotations are related to chromatin remodelling complexes, DNA damage and vacuolar pH.

In order to find enriched Gene Ontology annotations for all the 14 phenotypes, we combine the bottom and top 50 mean and residual scores and analysed them with the GO::Term Finder [14]. Here, we only show in Figure 5.3 the top 15 GO annotations for which the corrected p-values are the smallest. As in Chapter 4, the bottom residual ORFs do not show enriched GO annotations. The GO annotations found for the bottom mean scores are general common annotations, not necessarily indicating specific enrichment. The top mean and residual categories however show richer GO annotations as depicted in the word cloud of panel C.

Top Residual Annotations			Top Mean Annotations		
GO	Term	Corrected P-Value	GO	Term	Corrected P-Value
GO:0065007	biological regulation	5.79E-15	GO:0006974	cellular response to DNA damage stimulus	1.62E-14
GO:0006974	cellular response to DNA damage stimulus	3.79E-13	GO:0006259	DNA metabolic process	2.76E-12
GO:0050789	regulation of biological process	5.93E-11	GO:0065007	biological regulation	1.97E-10
GO:0016043	cellular component organization	5.35E-10	GO:0090304	nucleic acid metabolic process	2.45E-10
GO:0006281	DNA repair	7.08E-10	GO:0051276	chromosome organization	5.48E-10
GO:0050794	regulation of cellular process	1.19E-09	GO:0006281	DNA repair	1.36E-09
GO:0006996	organelle organization	7.27E-09	GO:0050794	regulation of cellular process	1.83E-08
GO:0051171	regulation of nitrogen compound metabolic process	1.30E-08	GO:0006139	nucleobase-containing compound metabolic process	1.87E-08
GO:0050896	response to stimulus	4.13E-08	GO:0050789	regulation of biological process	3.79E-08
GO:0044763	single-organism cellular process	4.32E-08	GO:0046483	heterocycle metabolic process	4.22E-08
GO:0006259	DNA metabolic process	5.23E-08	GO:0006725	cellular aromatic compound metabolic process	1.32E-07
GO:0060255	regulation of macromolecule metabolic process	8.87E-08	GO:0006950	response to stress	2.63E-07
GO:0051276	chromosome organization	9.77E-08	GO:0051171	regulation of nitrogen compound metabolic process	2.72E-07
GO:0019219	regulation of nucleobase-containing compound	1.15E-07	GO:1901360	organic cyclic compound metabolic process	2.73E-07
GO:0080090	regulation of primary metabolic process	1.21E-07	GO:0033554	cellular response to stress	2.75E-07
Bottom Residual Annotations			Bottom Mean Annotations		
GO	Term	Corrected P-Value	GO	Term	Corrected P-Value
No significant terms found			GO:0016043	cellular component organization	0.000180812
			GO:0065007	biological regulation	0.000382059
			GO:0008361	regulation of cell size	0.000606146
			GO:0007005	mitochondrion organization	0.00252558
			GO:0071840	cellular component organization or biogenesis	0.002817411
			GO:0044763	single-organism cellular process	0.003031453
			GO:0050789	regulation of biological process	0.003089927
			GO:0050896	response to stimulus	0.00324747
			GO:0006996	organelle organization	0.004132456
			GO:0051716	cellular response to stimulus	0.004720514
			GO:0007584	response to nutrient	0.007727622

Figure 5.3: To analyse the mean and residual scores, it is necessary to look at the bottom and top 50 genes for all 14 phenotypes separately, according to their class, i.e., either bottom mean or bottom residual and similarly for the top mean and top residual. Using the GO::Term Finder [14], we find the GO annotations related to each category of ORFs. Here we list the annotations (the first 15) with the smallest corrected p-values. The genes extracted for the bottom residuals do not show any significant terms, as in Chapter 4. The top genes are in general more enriched and show annotations related to DNA or stress response are involved in cell size.

To understand gene deletion strains with the lowest residual scores and find the reason of the non significance of the enriched GO terms among genes within these categories, we compute the hypergeometric test for all the GO Slim annotations. The p-values of the hypergeometric tests are represented in a heatmap (Figure 5.4.D). The majority of the p-values are larger than the acceptable threshold of 0.05, therefore indicating that the hypergeometric tests failed on some GO annotations.

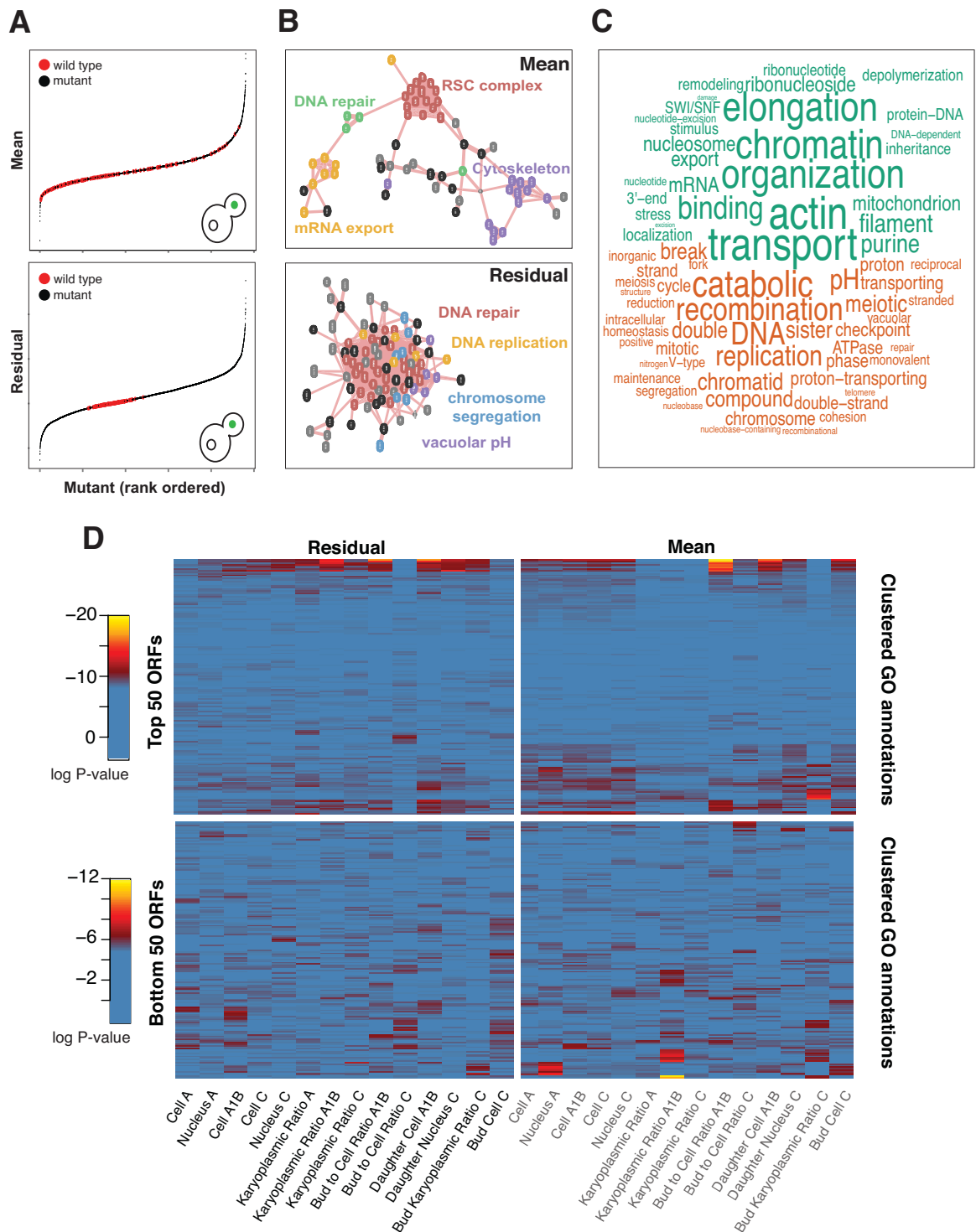


Figure 5.4: Outliers in phenotypic mean and variability provide unique biological information. A. Rank order plot for mean and variability, represented as residuals, of haploid nonessential deletion strains in the nuclear size of post-mitotic buds. Wild-type (*his3*) deletion strains are indicated in red. B. The top ranked genes were used to generate physical interaction networks in Genemania. Biological functions within the networks are indicated. C. Word cloud generated from words (excluding common words) within the significantly enriched GO terms from (B). D. Comparison of GO Slim enrichment for top and bottom ranked genes in mean or variability, across the phenotypes considered in this study. To compare the GO annotations a hypergeometric test is performed on each annotation.

5.3 Algorithms for Identification of Related Groups of Multi-gene Outliers in 2-dimensional Data Sets

In the last section we devised two scores based on the mean phenotypic value and the coefficient of variation of phenotypes related to the cell, the nucleus and the bud sizes. We then analysed the bottom and top 50 genes for all the 14 phenotypes and looked at the enriched gene ontology annotations for the bottom and top genes. We showed that the separate analysis of both scores gives a first insight to find mutants that increase or decrease variability during the cell cycle. However, we also showed that when the gene ontology annotations are compared via a hypergeometric test for the bottom and top genes, we could only see few statistically significant annotations emerging.

To identify groups of deletion mutants with significant effects on the mean and variance of the cell and nuclear sizes, we devise two related algorithms based on geometric separation between the mutants and the wild types. We apply the algorithm on each of the different gene ontology categories available. Here, we remove categories with less than five genes to consider annotations containing between 5 and 1583 genes.

In all cases, the separation between wild-type and mutant populations is assessed in 2-dimensional space, where the x - and y -dimensions are described by the mean phenotypic value and its associated variability (coefficient of variation; CV), respectively. In particular, we examine the same morphological phenotypes as in Section 5.2.

Through application of each algorithm, we identify groups (GO annotations) with significant and coherent separation between wild types and mutants, where significance is defined through application of each algorithm to randomly permuted gene sets.

5.3.1 Distance between Ellipsoids

In a simple geometric algorithm, wild-type and mutant populations are estimated as ellipses in the mean phenotypic value and coefficient of variation space, with centre and axes defined by the average value of the mean and standard deviation respectively of both the mean phenotypic value and the respective coefficient of variation.

Once the two ellipsoid centres (one for the mutants and one for the wild types) are found, we find the line intersecting the two centres and we calculate three distances. The first distance, labelled **A**, is the distance between the centre of the mutant ellipsoid and the intersection of the mutant ellipsoid with the line intersecting both centres. Similarly, the second distance, **B**, is the distance between the wild-type ellipsoid centre and the intersection of the ellipsoid with the line intersecting the two centres. Finally, the distance **C** is the distance between the two centres. An example of the distances is illustrated in Figure 5.6.A.

The separation score between each population is defined as the relative ratio of the dispersion of the two ellipses versus the distance between them as defined in Eq. (5.4).

For a particular GO annotation x and a particular phenotype j , the ellipsoid score E_x^j , based on the distances A, B and C, is formally defined as

$$E_x^j := \frac{A + B}{C} , \quad x = 1, \dots, 1100 , \quad j = 1, \dots, 14. \quad (5.4)$$

In this setting, a score larger than one indicates that the ellipsoids are geometrically separated.

The score is then computed for all the gene ontology annotations and the phenotypes related to the cell and nucleus. The full algorithm is described in Algorithm 5.3.

Algorithm : Ellipsoid separation

1. Import the mutants and wild-type datasets.
2. Import the GO annotations with references to the ORFs.
3. Normalise the mutant and wild-type phenotypes by subtracting the mean and dividing by the standard deviation of all samples of each phenotype.
4. Compute : M_{WT} and CV_{WT} the average mean and respectively CV phenotype values for the wild types.
5. Compute : $R_{1,WT}$ & $R_{2,WT}$ as

$$R_{1,WT} = 2 \times SD(X_{mean,WT}) ; R_{2,WT} = 2 \times SD(X_{cv,WT})$$
 where $X_{mean,WT}$ is the collection of data representing the mean value phenotype of interest for the wild types. The same applies for the CV.
6. Create the wild-type ellipsoid centred in (M_{WT}, CV_{WT}) with radii $R_{1,WT}$ and $R_{2,WT}$.
7. For each GO annotation :
 - a. Select mutants within GO annotation.
 - b. Compute : M_{MT} and CV_{MT} the average mean and respectively CV phenotype values for the mutants.
 - c. Compute : $R_{1,MT}$ & $R_{2,MT}$ as

$$R_{1,MT} = 2 \times SD(X_{mean,MT}) ; R_{2,MT} = 2 \times SD(X_{cv,WT})$$
 - d. Create the mutant ellipsoid centred in (M_{MT}, CV_{MT}) with radii $R_{1,MT}$ and $R_{2,MT}$.
 - e. Define the line L , connecting the pairs (M_{WT}, CV_{WT}) and (M_{MT}, CV_{MT}) .
 - f. Compute :
 - The distance A between (M_{WT}, CV_{WT}) and the intersection with the wild-type ellipsoid.
 - The distance B between (M_{MT}, CV_{MT}) and the intersection with the mutant ellipsoid.
 - The distance C between the centres of both ellipsoids.
 - g. Calculate

$$\text{Ellipsoid Score} := \frac{A + B}{C}$$
 - h. End For

Algorithm 5.3: Estimation of the 2D distance between the wild types and the mutants. The algorithm computes ellipses around the two populations and find whether the populations are well separated or not.

5.3.2 K-medoids clustering

In an alternative approach, we apply k-medoids clustering (with $k = 2$) to combined populations of wild-type and mutant groups. Separation efficiency and, by extension, phenotypic difference between wild-type and mutant groups is assessed by the error rate in attributing each respective group back to its original category.

The algorithm processes the two groups in four steps. First, the raw data containing both the wild types and mutants with their respective mean phenotypic value and coefficient of variation are imported for a specific gene ontology annotation x . The wild-type dataset is shrunk to the number of mutants in the specific annotations to avoid irregularities in the clustering process. Therefore, if the number m of mutants available in the gene ontology x is less than 25 (our statistical admittance here), a random sampling of the wild types is performed to produce m mean phenotypic values and CV's. Otherwise, if the number of mutants in the annotation is greater than 122 (the number of replicates for the wild types) the mutants are sampled with replacement to produce a list of 122 mutants. In the sampling process, the mean phenotypic value and its respective coefficient of variation are kept in pair. This step assures the parity in number of elements to analyse in the k-medoids algorithm. Second, the mutants and wild types are tagged to distinguish in which group they belong. A value of 1 is given to the wild types and a value of 2 for the mutants. The third step utilises the k-medoids algorithm, on the merged set of mutants and wild types. The idea is to cluster the complete dataset in two groups ($k = 2$) and look at the distribution of labels (1 or 2). This is to see whether a mutant or a wild type has moved and now belongs to the other group. The final step calculates the error rate r_x^j of mislabelled 1s and 2s for a particular phenotype j .

The final "k-medoids score" for the gene ontology annotation x and the phenotype j is the minimum labelling error given by

$$K_x^j := \min\{1 - r_x^j, r_x^j\} \quad , \quad x = 1, \dots, 1100 \quad , \quad j = 1, \dots, 14. \quad (5.5)$$

Algorithm : K-medoids separation

1. Import the mutants and wild-type datasets.
2. Import the GO annotations with references to the ORFs.
3. Normalise the mutants and wild-type phenotypes by subtracting the mean and dividing by the standard deviation of all samples of each phenotype.
4. For each GO annotation :
 - a. Select mutants within GO annotation.
 - b. If the number **m** of ORFs in the GO annotation is less than 25, create a random sampling of the wild-type data with **m** elements. if $m \geq 25$, sample mutant data with replacement 122 times.
 - c. Label the wild types with "1" and the mutants with "2".
 - d. Cluster the wild types and mutants with a permutation around medoids algorithm set for two clusters.
 - e. Compute the error r , consisting of the number of mislabelled "1" and "2" divided by the number of elements in the GO annotation. Given that group sizes are equal, the error can be no greater than 0.5.
 - f. Calculate the "K-medoids score"

$$K := \min\{1 - r, r\}$$
5. Apply the algorithm N times on each category to get an interval of scores and their respective means and standard deviations.

Algorithm 5.4: The application of a partitioning around medoids algorithm on labelled mutants and wild types allows discernment of whether a mutant could be erroneously classified as a wild type and vice versa. Once the mutants and wild types are clustered in two groups, it is possible to compute an error about the number of misplaced labels. This algorithm gives an idea about the separation of the mutants and wild types and therefore allows to determine whether a specific GO annotation differs from the wild-type trend.

5.3.3 Acceptance/Rejection Test

The number of genes present in the GO annotations varies from one annotation to another. Therefore, it is necessary to perform a specific test to accept or reject an annotation depending on the score it receives from both algorithms.

For each GO annotation, a pair of scores is created using the medoid and ellipsoid

Algorithm : Acceptance/Rejection test

1. Define, k = number of genes in GOx , construct the set of scores $S_x = \{\emptyset\}$, empty at first.
2. Randomly pick k mutants in the pairs of phenotypes (e.g., "C11-1_A" and "CCV11-1_A") .
3. Compute the medoid and ellipsoid scores on the k elements previously picked and update S_x with its new score.
4. Repeat points 2. & 3. N times.
5. Construct the 2D-kernel density estimator P_x using the values in S_x .
 - Accept score(GOx) if:

$$P_x[\text{score}(\text{GOx})] \leq p_{\text{val}}^2$$
 - Reject score(GOx) if:

$$P_x[\text{score}(\text{GOx})] > p_{\text{val}}^2$$

Algorithm 5.5: Based on the probability of randomly choosing mutants from the SCMD, the acceptance/rejection tests ensures that a GO annotation has a p-value small enough to be considered in the analysis.

algorithms. Thus, for the annotation GOx (with $x = 1, \dots, 1100$ annotations), we define

$$\text{score}(\text{GOx}) = (\text{Medoidx}, \text{Ellipsoidx}) \quad (5.6)$$

where Medoidx and Ellipsoidx are respectively the scores calculated using the medoid and the ellipsoid algorithms on the GO annotation x . The idea of the test is to calculate the p-value of choosing a GO annotation when its elements are chosen randomly.

Therefore, if $|M|$ is the total number of mutants, the probability " p " to pick the same original k elements from the GO annotation is

$$p = [C(k, M)]^{-1} \rightarrow 0, \quad M \gg 1$$

where $C(k, M)$ is the number of possible combinations of k elements within a set of M elements. Now if the test is repeated N times, each time permuting the elements, the overall

probability of choosing the original k elements is

$$p^N \rightarrow 0 \text{ as } N \gg 1 \quad (5.7)$$

If the number N of repetitions is large enough (here $N = 1000$), the original score for a particular GO annotation x can be assessed by computing its p-value. Hence, it is possible to discretise the two-dimensional score space, resulting from the N repetitions, into bins and compute the probability density P_x for the pairs of scores. Once P_x is calculated, the acceptance or rejection of the original GO annotation score can be thereafter decided

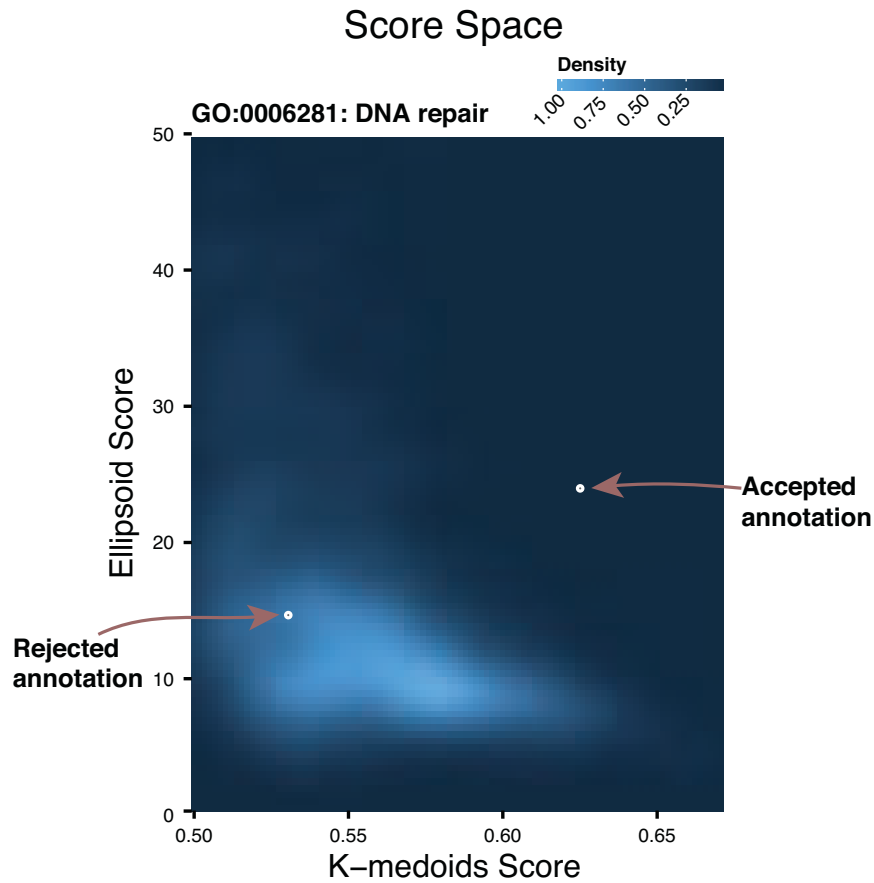


Figure 5.5: To accept or reject a particular GO annotation, we test whether the GO annotation could possess the same pair of scores as if the scores were found by randomly choosing mutants. Once the p-value of the test is set, we accept the GO annotation lying in the region of the score space where the probability is less than the p-value. Otherwise, the GO annotation is rejected

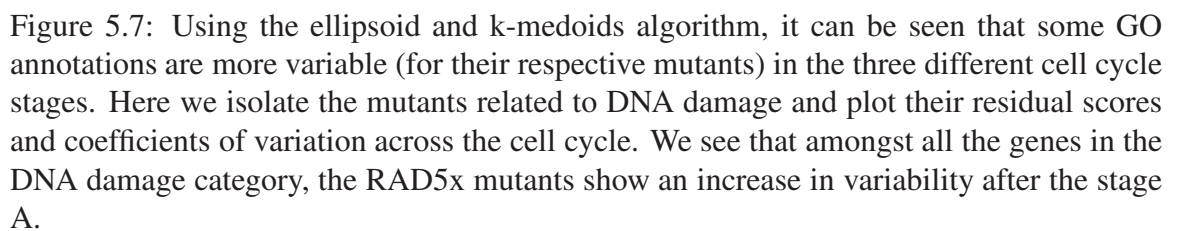
using a (squared) p-value, denoted p_{val} . If the original score lies in a region of the two-dimensional random score space with a probability smaller than $p_{\text{val}}^2 = 0.01^2$ (here, the probability is squared to consider both dimensions), then the score is accepted and the GO annotation is validated as illustrated in Figure 5.5.

5.4 Results

From the Acceptance/Rejection test, 158 enriched GO annotations amongst the 1100 are accepted (the full list is available in Appendix C). To investigate the effect of the annotations on the mutants and the wild types, we look at the fold change of the mean phenotypic value and the coefficient of variation. To this end, we select the ORFs belonging to a specific GO annotation and we compute the average value for the mean phenotypic value and also the coefficient of variation, for each of the 14 phenotypes analysed in this study. We repeat the same calculation for the wild types. Once the average value for each phenotype is found, we compute the logarithm (in base 2) of the ratio of the mutant and wild-type average values. The logarithm of the ratio indicates, in this setting, the fold change between the mutant and wild-type values. Three observations are then available. Firstly, if the log-ratio is zero, then the mutants are not significantly different from the wild types. Secondly, if the log-ratio is negative, the mutants show smaller values than the wild types and finally, when the log-ratio is positive, the mutants show larger values than the wild types.

We cluster in a heatmap, based on the average Euclidian distance, the log-ratio for the 14 phenotypes and the 158 GO annotations in Figure 5.6.C. The log-ratio values range from -0.4 to 1 . One can see for the phenotypes labelled with "CV" that some GO annotations do increase or decrease variability. The annotations decreasing variability are related to cell growth, such as those terms relating to tRNA, amino acids, rRNA processing and vesicle function. The annotations increasing variability are related to checkpoints of the cell cycle (such as those concerning DNA repair, mitotic spindle assembly, and chromosome segregation); chromatin remodelling, and maintenance of vacuolar pH. The GO annotations found here are similar to the annotations previously found using the mean and residual scores thus indicating that our second methodology is coherent with the first one.

In addition, we notice that there are GO annotations showing less or no variability in stage A compared to the stages A1B and C. In Figure 5.7, we isolate the mutants related to the DNA damage annotation and we plot their mean distance and residual scores in a heatmap. As one can see, some mutants become more variable once they reach stage A1B. The RAD5x category of mutants is amongst the genes showing such increase in variability. At the moment, it is still unclear why the RAD5x genes increase variability in stages A1B and C; however, in budding yeast, DNA damage causes cell cycle arrest in S phase (corresponding to stage A1B), without accompanying arrest of cell growth. Given the role of RAD5x genes in DNA repair, it is possible that these mutants are experiencing varying degree of damage and cell cycle arrest, perhaps explaining variability in cell size. Fitting with our hypothesis, in a recent work by Witkin et al. [150], they found that mutants belonging to the DNA repair and chromosome segregation categories, such as the RAD5x genes, produced an abnormal nuclear shape during the cell cycle. They showed the appearance of a nuclear extension, seen as an "arm" outgoing the nucleus, in the cytoplasm. These and our results suggest that the mutants cause a checkpoint-induced delay in the cell cycle. However, this supposition is still under consideration and requires further study.



5.5 Discussion

We showed in Chapter 4 that genes can increase or decrease variability in some morphological phenotypes, according to their function within the cell. Here, we decided to employ variability in cell and nucleus related phenotypes to better understand gene and cell cycle function in *S. cerevisiae*. We motivated the use of variability, seen as a coefficient of variation (CV), by looking at the CV of the mother cell size across the cell cycle of budding yeast. Unlike the mean phenotypic value of the mother cell size, i.e., the average mean of the cell for a specific stage, the CV, from one cell cycle stage to another, is less correlated indicating that some mutants are more variable in one cell cycle stage than the others. To identify which genes were involved in the deregulation of the cell cycle coherence, we decided to focus our research on 14 morphological phenotypes describing the cell and nucleus properties and devise two algorithms analysing phenotypic variability of the mutants.

The two methods we devised in this study first consider the mutants as single entities and second as groups of genes, belonging to gene ontology annotations. The first method consisted of extracting mutants that had a mean phenotypic value and a coefficient of variation larger than the wild types. To this end, we looked at the mean phenotypic value and the coefficient of variation separately. Then, we looked at the fold change compared to wild types and we ranked the resulting scores, based on fold change, to select the bottom and top 50 mutants showing a significant difference. Thereafter, we pursued the analysis by developing a second method that accounted for the mean phenotypic value and the coefficient of variation of the mutants as a unique parameter. The motivation for the second method was to investigate the mean phenotypic value and its coefficient of variation in a geometric manner, by looking at the location of the mutants compared to the wild types, in the mean phenotypic value-CV space.

With the first method, we showed that both the 50 mean distances and the residuals scores were able to provide sets of gene ontology annotations for the bottom and top ranked genes, compared to wild types. More specifically and using only 50 genes with high and low scores, the GO annotations that emerged from our analysis were related to DNA dam-

age, chromatin remodelling complexes and vacuolar pH. Although the annotations found here are of interest as they relate to cell cycle variability, we showed using a hypergeometric test that because of the small number of bottom and top genes we selected, the annotations could actually be an artefact of the random nature of the picking process.

The second method, based on the spatial analysis of the mean phenotypic value and its coefficient of variation provided a better and more statistically significant analysis of the gene ontology annotations. First, we were able to retrieve similar annotations found with the first method. Second, we showed that amongst the annotations found previously and the annotations found in the second method, there are GO terms that exhibit increase or decrease of variability during the cell cycle. The GO terms decreasing variability are related to cell growth, as for instance terms related to tRNA, amino acids, rRNA processing, and vesicle function. The GO annotations associated with an increase of variability are related to cell cycle checkpoints (DNA damage, spindle and chromosome segregation), chromatin remodelling complexes and vacuolar pH. These new results are currently under the investigation of our collaborators who are setting up new experiments to validate our analysis.

It must be noted that our two approaches differ from the standard hypothesis tests or classifications tests as we wanted to develop new procedures to segregate genes based on their phenotypic variability. Although hypothesis test such as t-test or hierarchical clustering could have been employed here, we decided to test novel approaches and see if we could get new results. As a follow up analysis, our collaborators will also extend our methodology to standard tests, i.e., hypothesis testing, and for example a two-dimensional paired Kolmogorov-Smirnov test to ensure the quality of our results.

We also noticed using both methods that there are categories of genes that do not show high variability at the beginning of the cell cycle in stage A but increase it in the remaining cell cycle stages. Here, we identified the RAD5x set of mutants as some of the genes increasing variability after the stage A.

To conclude, we showed in our analysis that although the mean phenotypic value of

phenotypes related to the cell and the nucleus can be employed to investigate the regulation of the cell cycle, variability can provide more information distinct from the mean phenotypic value.

Chapter 6

A Primer on Cell Size and Nuclear Size Regulation in *S. cerevisiae* using Bayesian Network Analysis

6.1 Introduction

In Chapter 4 we showed that the morphological phenotype distributions can be extracted to infer the importance of deleted genes on cell variability, as for instance with the inducer and capacitor genes, decreasing and increasing variability for some phenotypes respectively. In addition, in Chapter 5, we showed that it is possible to use phenotypic variability, seen as a the coefficient of variation (cv), together with the mean value of morphological phenotypes to extract GO annotations related to cell size. Although the mean and coefficient of variation are two standard and powerful tools to analyse genetic modifications, the mean and cv miss the true empirical nature of the morphological phenotypes. The distributions of the morphological phenotypes are considered independent in Chapter 5, as we treat each phenotype individually without assuming that if one phenotype changes, variations should be seen in one or many other phenotypes.

In this study, we regard the distributions of the morphological phenotypes as dependent variables, as a whole, meaning that a change in a phenotype might produce a change in

one or multiple other phenotypes. This assumption of dependence, as simple as it might sound, is sometimes not well understood in biology, more especially in the regulation of the cell size and the nucleus size during the cell cycle. Let us consider the following example, of a cell increasing its volume (cytoplasm) during its cell cycle. Does the nuclear volume increase alongside the cell volume or is there a mechanism that prevents the nucleus to increase its volume, faster or slower than the cell size? On the contrary, does the cell slow down or speed-up its growth to realise a coordinated growth with the nucleus? In this research, we give a first answer using mathematical modelling to see the dependence between the cell and the nucleus.

Geneticists have tried to understand the mechanism beneath the regulation or coordination of the cell and nuclear sizes. Studies from Neumann & Nurse [97], Jorgensen et al. [54] and Huber & Gerace [52], propose the idea of a possible control of the cell and nuclear sizes by showing that the karyoplasmic ratio, i.e., the ratio of the nuclear volume by the cell volume, is constant throughout the cell cycle of *S. pombe*. However, it is still unclear why the karyoplasmic ratio remains constant throughout the cell cycle. According to Marshall et al. [87, 86, 17], size control is seen as a function of the length of an organelle coordinated by the variation of a unique parameter. This hypothesis is now mainly agreed in biology and experiments focus on specific organelles or more generally on cellular entities.

However, due to the complexity of certain experiments, the genetic environment of the cell (such as other organelles) are often not taken into account in the process of examining cell and nuclear size regulation. Such "environmental" considerations were recently examined by Webster et al [147] in their commentary on nuclear size regulation, as they found a possible link between the nuclear envelope and the endoplasmic reticulum, indicating that the nucleus might, de facto, "communicate" with the outside of its membrane. This hypothesis has been verified by Levy & Heald [75] in their recent work on *Xenopus laevis* and *Xenopus tropicalis*. By analysing the concentrations of two transport factors, Importin α and Ntf2, they showed that nuclear transport mechanisms, e.g., cargoes flowing inside and outside the nucleus via the nuclear pore complexes, are physiological regulators of nuclear scaling. This recent discovery strongly suggests that in order to find a regulator

of the cell and nucleus sizes, it would be wise to consider the whole cell machinery more than specific cellular entities.

New techniques in cell imaging have nonetheless brought new insights in biology and the understanding of the whole cell machinery but also created new challenges to acquire and analyse the data gathered with biological experiments. The path from few phenotypes analysed by hand under the microscope to hundreds or even more phenotypes gathered with high-throughput screening requires the development of new tools. Straddling statistical inferences and computer science, machine learning algorithms can be trained to learn and work on such large amount of data. In order to classify, for example, phenotypes and their relationship to one another, machine learning algorithms such as Bayesian networks are well adapted.

For the past decade, new discoveries in gene regulation have been possible thanks to the use of Bayesian networks. Mainly employed to reconstruct gene regulatory networks (Vignes et al. [141], Markowitz [85]), Bayesian networks are becoming an important tool in Systems Biology to study the effect of genetic variations on morphological phenotypes (Nagano et al. [93], Lehner [70], Wang & Marcotte [145]). Bayesian networks allow capturing the essence of the biological experiment and the parameters being analysed by learning and constructing networks connecting the parameters hierarchically.

In our analysis, we employ Bayesian networks to reveal the existence of shared regulation between the cell size and the size of the nucleus in *S. cerevisiae*. By comparing the network topologies created using multiple morphological phenotypes, we show that wild types maintain a close relationship between the cell size and the nucleus size throughout the cell cycle. After breaking the cell cycle into three stages, namely: the stage A (corresponding to the G1 phase), A1B (corresponding to the S & G2 phases) and stage C (late M phase), we see that the wild types keep a strong directionality from the cell size to the nucleus size in stages A and A1B and a less strong in stage C. More specifically, the direction of the relationship indicates that the cell size dictates the nuclear size during the growth of *S. cerevisiae*. To find a potential explanation of the cell-nucleus regulation, we

investigate Bayesian networks generated on a collection of mutants available in the *Saccharomyces Cerevisiae* Morphological Database (SCMD). Both the wild types and mutants are available in the form of single-cell measurements, allowing to extract distributions for the morphological phenotypes.

This study is divided into four parts. In the first part we give a first insight in the cell-nucleus regulation problem, using wild-type data and computing a conditional score on them, without considering the use of Bayesian networks. The second part is related to the general analysis of the wild types using standard Bayesian networks and equivalence classes to demonstrate the relationship between the cell size and the nuclear size. The third part repeats the network analysis on a collection of mutants to find mutants disturbing or changing the directionality between the cell and nucleus. Finally, the last part concludes our study and suggests enhancements and further analyses.

6.2 Single-cell Measurements & Motivation

Understanding the relationship between the cell size and the nucleus size requires gathering and analysing morphological phenotypes related to the cell and its organelles as well as the nucleus. In this study, we employ single-cell measurements provided by the SCMD. This section introduces the available phenotypes, the nature of their distributions as well as a first insight in the regulation of the cell size and the nucleus size, by employing conditional probabilities on the phenotypic distributions.

6.2.1 Morphological Phenotypes and Single-cell Distributions

The *Saccharomyces Cerevisiae* Morphological Database provides sets of single-cell measurements for the wild types (*his3* strains) and 4718 mutants. After cleaning the data as performed in Chapter 4, we obtain a collection of raw morphological phenotypes. The raw nature of the phenotypes comes from the non-discretisation of the phenotypes into sub-phenotypes. For instance, a raw phenotype such as the mother cell size was subdivided into multiple phenotypes for the area or perimeter in the analyses of Chapters 4 & 5.

Phenotype \ Stage	A	A1B	C
Actin Region Ratio	x	x	x
Axis Ratio in Mother	x	x	x
Hip Nuclear	x	x	x
Nuclear Size in Mother	x	x	x
Mother Cell Center Nuclear	x	x	x
Mother Cell Size	x	x	x
Bud Actin Gravity Point		x	x
Bud Actin Ratio		x	x
Axis Ratio in Bud		x	x
Bud Cell Size		x	x
Bud Direction		x	x
Bud Ratio		x	x
Neck Nuclear in Mother		x	x
Neck Position		x	x
Neck Width		x	x
Nuclear Axis Ratio in Mother		x	x
Nuclear Size in Cell		x	x
Nuclear Axis Ratio in Bud			x
Nuclear Size in Bud			x
Dud Cell Center nuclear			x
Bud Top Nuclear			x
Length between Nucleus			x
Neck Nuclear in Bud			x

Table 6.1: Phenotypes available in the *Saccharomyces Cerevisiae* Morphological Database. In total 23 morphological phenotypes, chosen by Ohya et al. [99], are analysed but only 6 are expressed in stage A, 18 in stage A1B and 23 in stage C.

There are 23 morphological phenotypes in total. Some of the phenotypes are related to the cell, others are related to the bud or nucleus. Certain phenotypes possess distributions for the three cell cycle stages and some phenotypes only have distributions on a specific stage. This is mainly due to the expression of a specific phenotype at a certain stage. For example, it is unlikely to perceive bud related phenotypes in stage A as the cell just started growing. Hence, stage A has 6 phenotypes, stage A1B has 17 phenotypes and finally stage C is represented by 23 phenotypes (the whole available phenome). For clarity, we labelled the phenotypes with their availability in the three cell cycle stages in Table 6.1.

As an illustration, the distributions of the six morphological phenotypes in stage A for the wild types are shown in Figure 6.1. All the measurements are in the unit of pixels or square pixels for the areas. As one can see, the wild types have a considerable number of

Stage	Measurements	Phenotypes
A	16561	6
A1B	14953	18
C	7977	23

Table 6.2: Distributions of the 6 morphological phenotypes for the wild types in stage A based on the single-cell measurements. The frequencies show a large number of single-cell measurements available for the wild types.

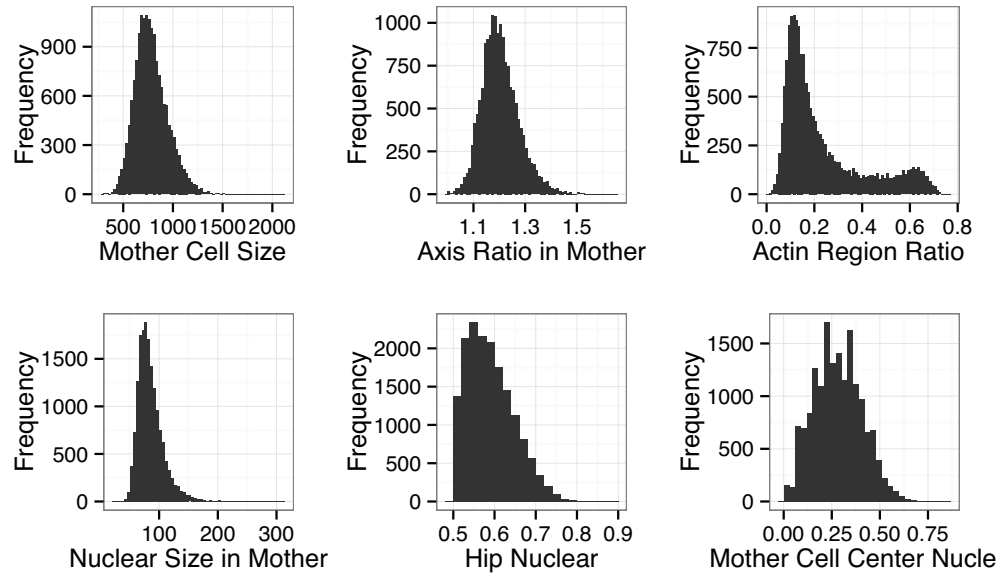


Figure 6.1: Phenotypic distributions of the wild types in stage A. The 6 available phenotypes in stage A have a large number of single-cell measurements, thus allowing to perform significant statistical analyses on them.

single-cell measurements for the six different phenotypes. The total number of phenotypes and their respective measurements are shown in Table 6.2.

Although all phenotypes have interesting features and properties, we mainly focus our analysis on two features: the mother cell size and the nuclear size in mother, which provide direct information on the cell (size) and nucleus (size). However, due to the nature of the phenotypes analysed in the SCMD, it sometimes happens that a phenotype is not expressed and has empty values for a specific stage. Hence, we only consider the phenotypes with non-empty values at a specific stage.

6.2.2 Conditional Dependence between the Mother Cell Size and the Nuclear Size in Mother

The single-cell measurements being in the form of n -tuples, where n is the number of phenotypes for a specific cell cycle stage, it is possible to perform standard statistics on the phenotypes. Here, we are specifically interested in the relationship between two phenotypes: the mother cell size and the nuclear size in mother. A first investigation, using the Pearson's coefficient of correlation, tells us that the mother cell size in stage A is well correlated with the nuclear size in mother ($r = 0.41$, $P < 0.001$). The correlation becomes less significant in stage A1B ($r = 0.32$, $P < 0.001$) and even less in stage C ($r = 0.26$, $P < 0.001$). As illustrated on the scatter plots (top rows) of Figure 6.2.A and B, the clouds representing the mother cell size and the nuclear size in mother tend to bend (from the stage A to the stage C), biasing the coefficient of correlation.

Although coefficients of correlation are a first good indication of the relationship between the mother cell size and the nuclear size, coefficients of correlation lack directionality, thus not informing which phenotype influences the other one. A recent work from Krishnaswamy et al. [64] has explored a new technique to elucidate the directionality between two features or phenotypes. The technique, called DREVI, is a conditional density-rescaled visualisation tool and is not based on the use of Bayesian networks.

DREVI works as a conditioner on the pair of variables that need to be analysed. When looking at a scatter plot, the number of data points (samples) can often become complicated to visualise, especially in the region where the majority of samples lie. To cope with this, DREVI first analyses the joint kernel density of the 2-tuples (x, y) , middle rows of Figure 6.2.A and B, where x 's and y 's are the realisations of two phenotypes X and Y . Then, the marginal density of X is determined in order to compute a conditional density estimate of Y given X . Finally, the conditional density estimate is rescaled by the maximum value given the ranges of y 's. The complete DREVI algorithm is given in Algorithm 6.3.

Algorithm : DREVI

1. For a pair of variables X and Y compute the joint density estimate $f_{X,Y}(x, y)$.
2. Determine numerically the marginal density of X , denoted $f_X(x)$ and the conditional density $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$ on a grid of points

$$G = \{(x_i, y_j), 1 < i < n, 1 < j < m\},$$

representing the range of samples attained by X and Y .

3. Rescale the conditional density estimation by defining

$$f^*(y_j|x_i) := \frac{f(y_j|x_i)}{\max_k \{f(y_k|x_i)\}}$$

4. DREVI is visualised by plotting $f^*(G)$ as a heatmap.

Algorithm 6.3: DREVI calculations of the conditional density in a pair of variables X and Y . The algorithm first calculates the joint densities of the two variables and thereafter the conditional density to examine the relationship between the two variables.

The conditional density estimate of a variable given another variable tells us whether or not one variable depends on the other. For instance, if variable X acts as a control on variable Y when X varies, Y should also vary (as X controls it). Therefore, when one looks at the conditional probability of Y given X , a defined variation in X produces the same defined variation in Y . In terms of probability, this means that the probability of seeing variations in Y given X should remain proportional to X , if X controls Y . If the probabilities do not remain proportional, this may indicate that X does not surely control Y . Therefore, in this case, the opposite analyse, i.e., Y controlling X , might be of interest.

A straightforward application of DREVI on the wild-type dataset for all three cell cycle stages is shown in Figure 6.2. Here, two scenarios are envisaged. The first one, in panel A, with the mother cell size controlling the nuclear size in mother. The second scenario, in panel B, is the opposite idea, with the nuclear size in mother controlling the mother cell size. The bottom rows of panels A and B are the final rescaled conditional densities. In panel A, we see high R^2 coefficients for all three cell cycle stages (0.97, 0.91 and 0.94 for the stages A, A1B and C respectively). Indeed, the conditional density of the nuclear

size in mother given the mother cell size shows linear trends when the mother cell size varies. In panel B, the trends are no longer as linear as in panel A. Variations in the mother cell size do not necessarily cause with the same probability, variations in the nuclear size in mother. However, the R^2 statistics associated with the second scenario are still high enough (0.7, 0.78 and 0.51 for the stages A, A1B and C respectively), therefore not clearly determining whether the first or the second scenario fits the model best.

Unlike Bayesian networks, DREVI functions with two phenotypes. This restriction probably causes DREVI to provide good R^2 scores but does not allow to decide whether or not the mother cell size controls the nuclear size in mother. To cope with this, in the next section we examine Bayesian networks on the wild-type dataset and we employ all the phenotypes (variables) available in the dataset.

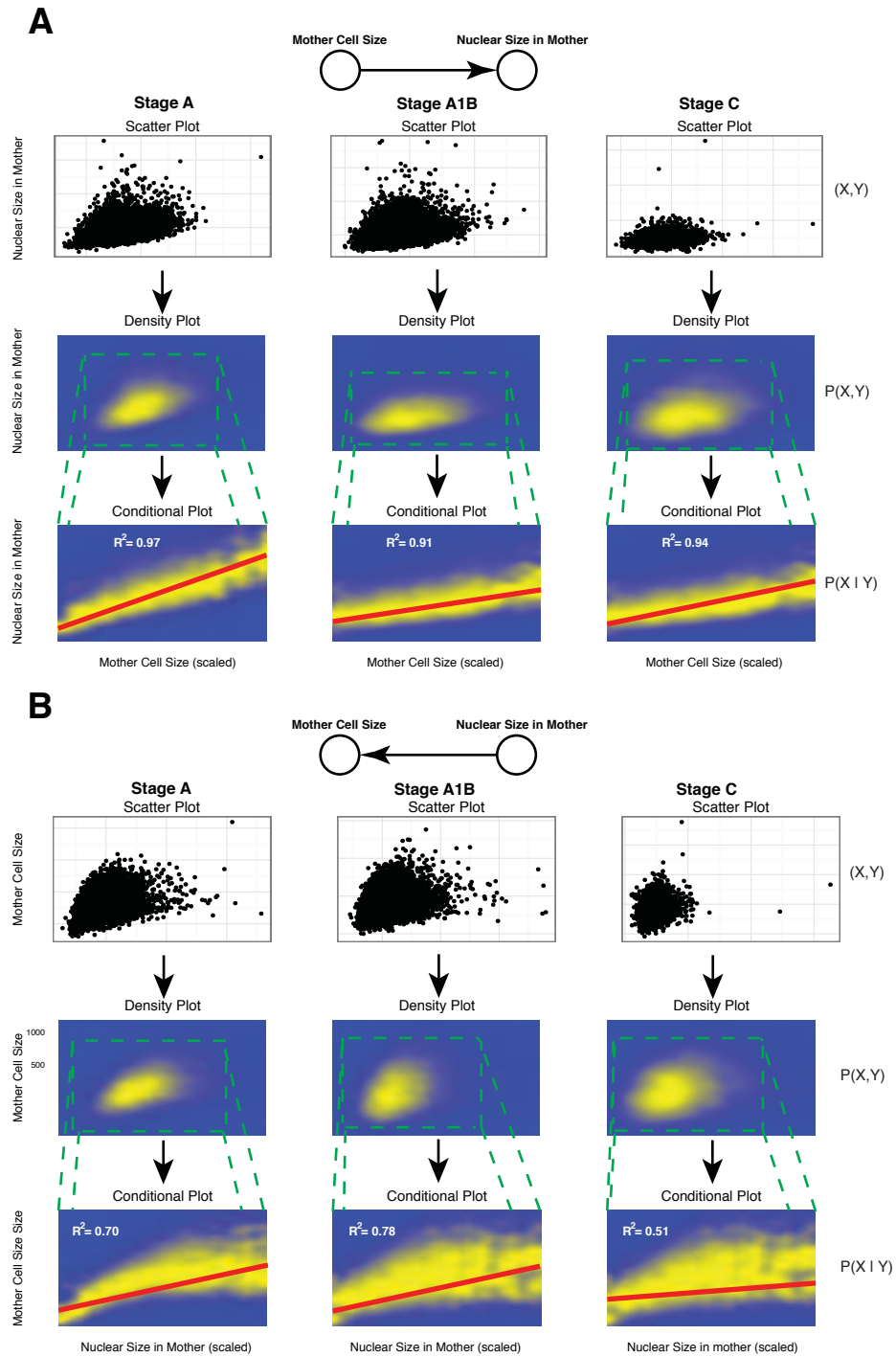


Figure 6.2: DREVI analysis of the relationship between the mother cell size and the nuclear size in mother for the three stages A, A1B and C for the wild types. From scatter plots of both the mother cell size and the nuclear size in cell measurements in the top panels of A and B, the DREVI score first computes the joint density of the measurements, middle panels, to finally produce the conditional probability of the measurements, bottom panel. The relationship "mother cell size" influences "nuclear size in mother" is analysed in A, whereas the opposite direction, i.e., "nuclear size in mother" influences "mother cell size" is shown in B. According to the R^2 scores, the first relationship, with the mother cell size influencing the nuclear size in mother seems to fit the model better. However, one needs to inspect if other morphological phenotypes also modify the behaviour of the cell and nucleus sizes.

6.3 Bayesian Networks Analysis of the Wild Types

As described previously in Table 6.2, the wild-type data has a large number of single-cell measurements for all 23 phenotypes. We recall that due to the biological nature of the phenomena, the cells acquired during the image processing have continuous distributions for their phenotypes. In order to process, analyse and learn the networks from the wild-type dataset we apply the Max-Min Hill Climbing (mmhc) algorithm [136].

The mmhc algorithm is a localised learning algorithm that constructs the topology of the network and performs a greedy hill-climbing search to find the directions of the edges. The mmhc belongs to the family of hybrid algorithms, i.e., algorithms combining both score-based and constraint-based searches.

In this section we present three different types of networks learned and constructed using the wild-type data. The first network is the network produced by the mmhc algorithm (called standard network). The second network is the equivalence class of the standard network, where the direction of the edges is significant only between equivalence classes. Finally, we give a bootstrapped version of the standard network, to visualise the strength and direction of the edges. All the analyses are performed using the `bnlearn` R package [122].

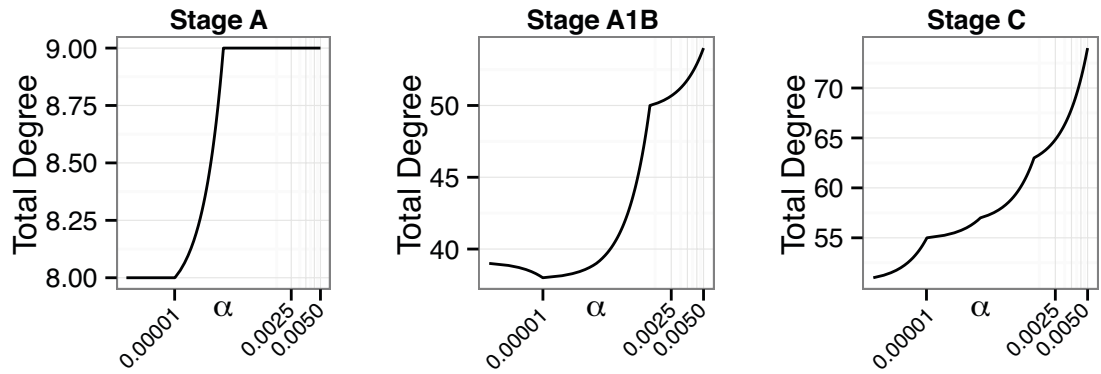


Figure 6.3: In its implementation, the Max-Min Hill Climbing algorithm allows to tune the nominal Type I error rate α . The plots represent the total number of directed arcs in the networks for decreasing values of α . In all three stages, the total number of directed arcs decreases as α decreases. For the purpose of our analysis, we set $\alpha = 1 \times 10^{-5}$ to guarantee strong causality between the nodes.

6.3.1 Cell Cycle Stage Networks and Equivalence Classes

The mmhc algorithm employs a G^2 statistics and an exact t-test (Spirtes et al. [127], Neapolitan [96]) to test the conditional independence and measure the associativity of a pair of two variables, i.e., if the two variables have an edge in the acyclic graph and what is the direction of the edge. In its implementation in the `bnlearn` package, the mmhc algorithm allows to tune the nominal Type I error rate α for the G^2 test.

By decreasing the value of α , the networks produced by the mmhc algorithm become finer and show stronger associativity between variables. To select an appropriate α , we look at the total number of directed arcs in the network for varied values of α (Figure 6.3), for the stages A, A1B and C. As one can see, the number of directed arcs diminishes as the value of α decreases. In this analysis we performed a sensitivity analysis on the nominal Type I error rate. We set the nominal Type I error to $\alpha = 1 \times 10^{-5}$ to guarantee strong causality between the phenotypes. We set the value of α as we can see in Figure 6.3 that the total number of directed arcs in stage A1B starts increasing again when α negatively overtakes 10^{-5} .

We also look at the equivalence class (as explained in the Background section) of the

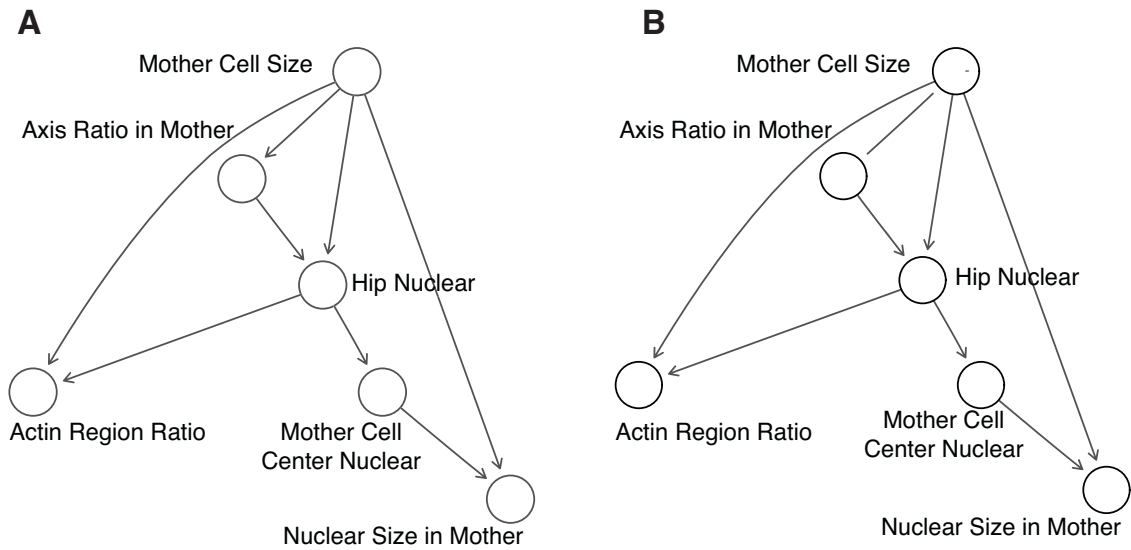


Figure 6.4: Bayesian networks learned and constructed for the wild types in stage A when the nominal Type I error rate α is set to 1×10^{-5} . A. Standard network of the 6 phenotypes. B. Equivalence class of the standard network. The mother cell size is the main node of the network, thus controlling the nuclear size in mother. This directed connection is seen in both networks.

standard network, i.e., the acyclic graph generated by the mmhc algorithm, to discern which nodes are members of the same class. The equivalence classes must be seen here as another criterion to decide whether or not two phenotypes are connected and if the direction between them is maintained.

The Bayesian network resulting from the mmhc algorithm on single-cell measurements of stage A (Figure 6.4.A) indicates that the mother cell size "controls" the nuclear size in mother. The nuclear size in mother in stage A is the actual size of the nucleus, as at this moment in the cell cycle, there is no bud. In addition, the mother cell size does not depend on any other phenotypes. The equivalence class of the network (Figure 6.4.B) also indicates that the mother cell size does influence the nuclear size in mother and is independent of the other phenotypes.

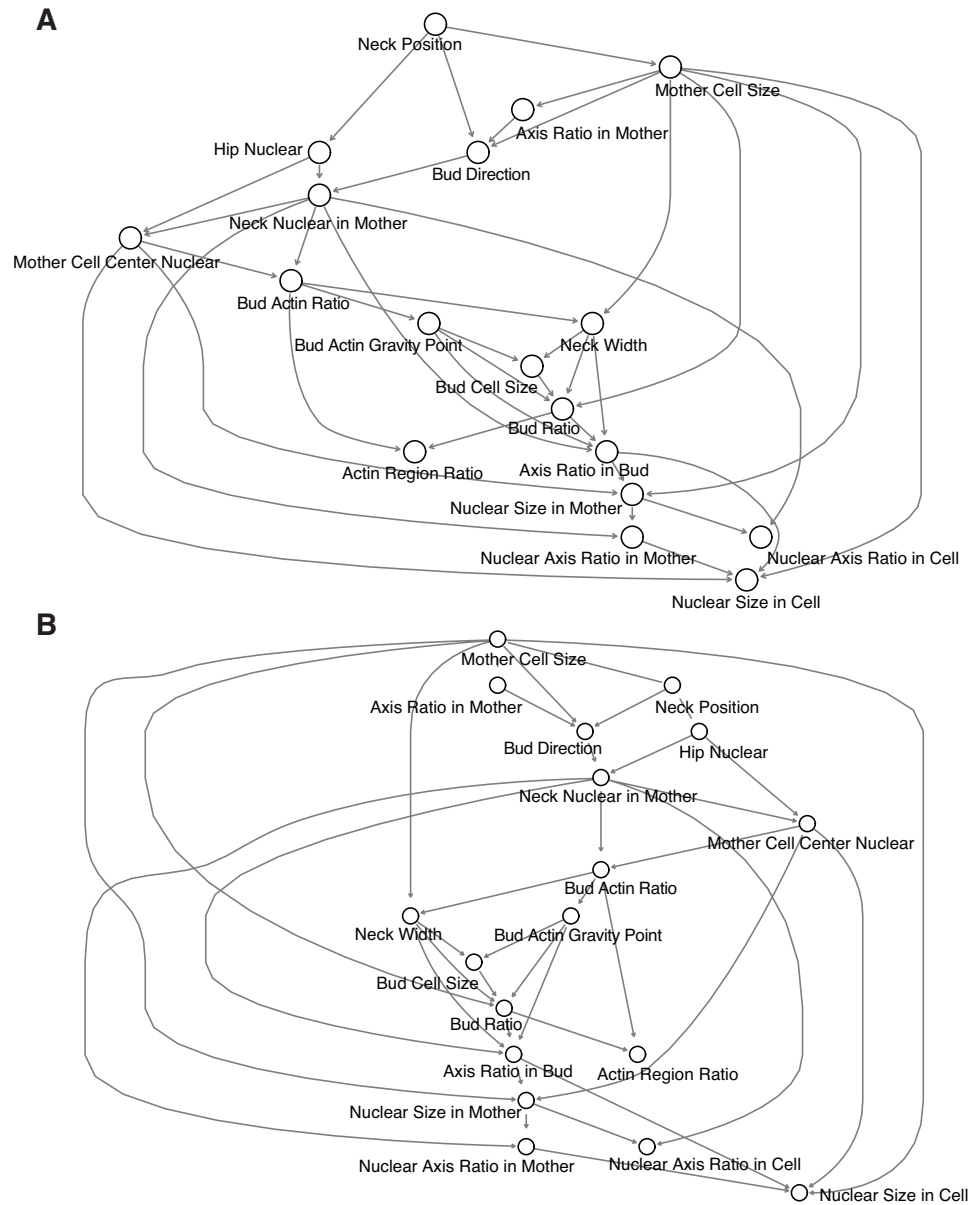


Figure 6.5: Bayesian networks learned and constructed for the wild types in stage A1B when the nominal Type I error rate α is set to 1×10^{-5} . A. Standard network of the 18 phenotypes. B. Equivalence class of the standard network. In the standard network in A., the mother cell size still controls the nuclear size in mother, however the mother cell size is controlled by the neck position. The control of the mother cell size by the neck nuclear disappears in the equivalence class network B., where the mother cell size becomes again the main node of the network.

The network for the stage A1B (Figure 6.5.A) also shows a relation between the mother cell size and the nuclear size in mother. In addition, the mother cell size now also controls the nuclear size in cell, as in stage A1B, there is the possibility of having a second nucleus in the bud.

Although the neck position influences the mother cell size in the standard network, we see that the causality disappears in the equivalence class (Figure 6.5.B), setting the mother cell size as the main principal (parent) node of the graph.

Finally, for stage C (Figure 6.6), the standard network is the same as its equivalence class. One can see here that the mother cell size is influenced by the nuclear size in mother but still influences the nuclear size in cell. We recall that at stage C, cells are at cytokinesis, which might cause the mother cell size to lose control of the nuclei.

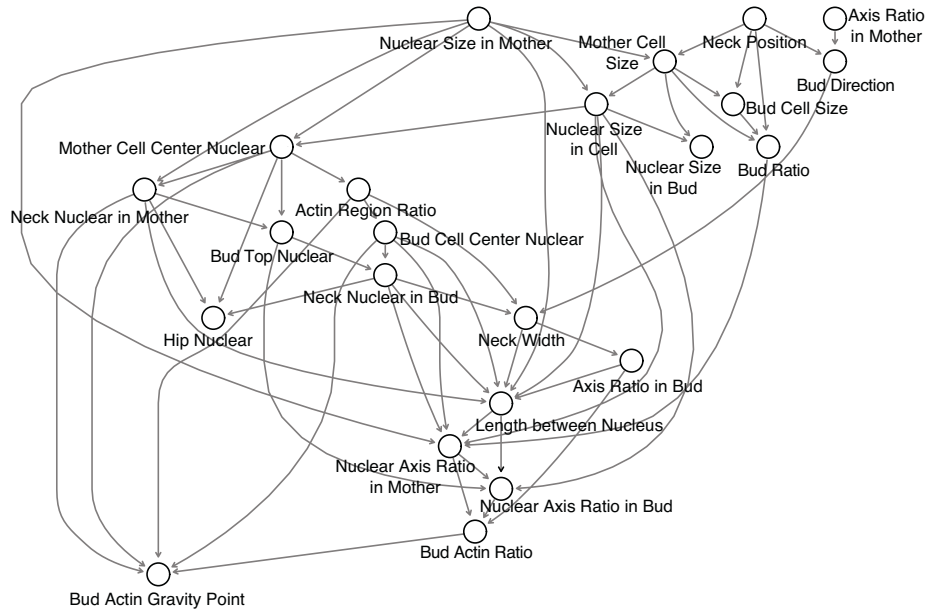


Figure 6.6: Bayesian networks learned and constructed for the wild types in stage C when the nominal Type I error rate α is set to 1×10^{-5} . Unlike the networks constructed for the stages A and A1B, the network for stage C is its own equivalence class, therefore not indicating whether some phenotypes belong to a class or not. Here, the mother cell size lost its directionality towards the nuclear size in mother. The direction is reversed, with the nuclear size in mother controlling the mother cell size. However, one can still see the mother cell size having a control on the nuclear size in cell. We recall that in stage C, the cells are at cytokinesis, which might probably cause the cell to lose control on its organelles and possibly changing the direction of the connectivity in the network.

6.3.2 Bootstrapping the Networks for all Cell Cycle Stages

We saw that the networks based on the wild types show a clear causation from the mother cell size to the nuclear size in stages A and A1B. In stage C, we saw an opposite direction between the mother cell size and the nuclear size in mother.

In this section, we investigate the strength and directionality of the edges in the networks constructed on the wild-type data. The strength and direction of the edges are measured by bootstrapping, i.e., sampling with replacement, the single-cell measurements and learning multiple networks (here 100) again. All newly constructed networks are then averaged to calculate a percentage of strength and direction for all edges in the network. We use the `boot.strength()` function in the `bnlearn` package to produce the bootstrapped networks and their percentages for the strength and direction.

The resulting bootstrapped networks are presented in Figure 6.7. The width of the arcs is scaled according to the strength between the nodes and a weight on the arc indicates the certainty of the direction. For clarity and because the networks in stage A1B and C are dense in arcs, the strength and direction from the mother cell size to all the other phenotypes are added in a table beside the networks. In addition, the top 10 phenotypes with the highest strength are added to the table, for informative purpose. The complete list of phenotypes and the strength and direction of their interactions is available in Appendix D.1.

The bootstrapped networks of Figure 6.7 clearly show a strong and well directed edge between the mother cell size and the nuclear size in mother for the stages A and A1B. For the stage A, a strength of 100% and a direction of 89% confirms our previous findings, on the causality between the mother cell size and the nuclear size in mother, presented with the standard network and its equivalence class. Similarly, for the stage A1B, the strength and direction have both 100% certitude, again confirming the standard and equivalence class analyses performed previously.

The bootstrapped network in stage C shows interesting properties for the mother cell

size and both, the nuclear size in mother and the nuclear size in cell. In the previous section, we showed that the nuclear size in mother controls the mother cell size and the latter is controlling the nuclear size in mother. A close look at the strength and direction from the mother cell size indicates that the mother cell size, although it has a high (93%) chance of being connected to the nuclear size in mother, has 52% chance of being controlled by the nuclear size in mother. Similarly, the strength and direction connecting the mother cell size and the nuclear size in cell are also small, possibly indicating that the mother cell size might not be influencing or being influenced by the nuclear size in cell.

The results provided by the bootstrapped version of the networks also indicate, as in the standard analysis of the wild types, that the mother cell size controls the nuclear size in mother for the stage A and A1B and the direction is lost in stage C. However, using the strength and direction of the arcs for the bootstrapped network in stage C, we see that the direction is not strong enough to clearly say that the control is reversed.

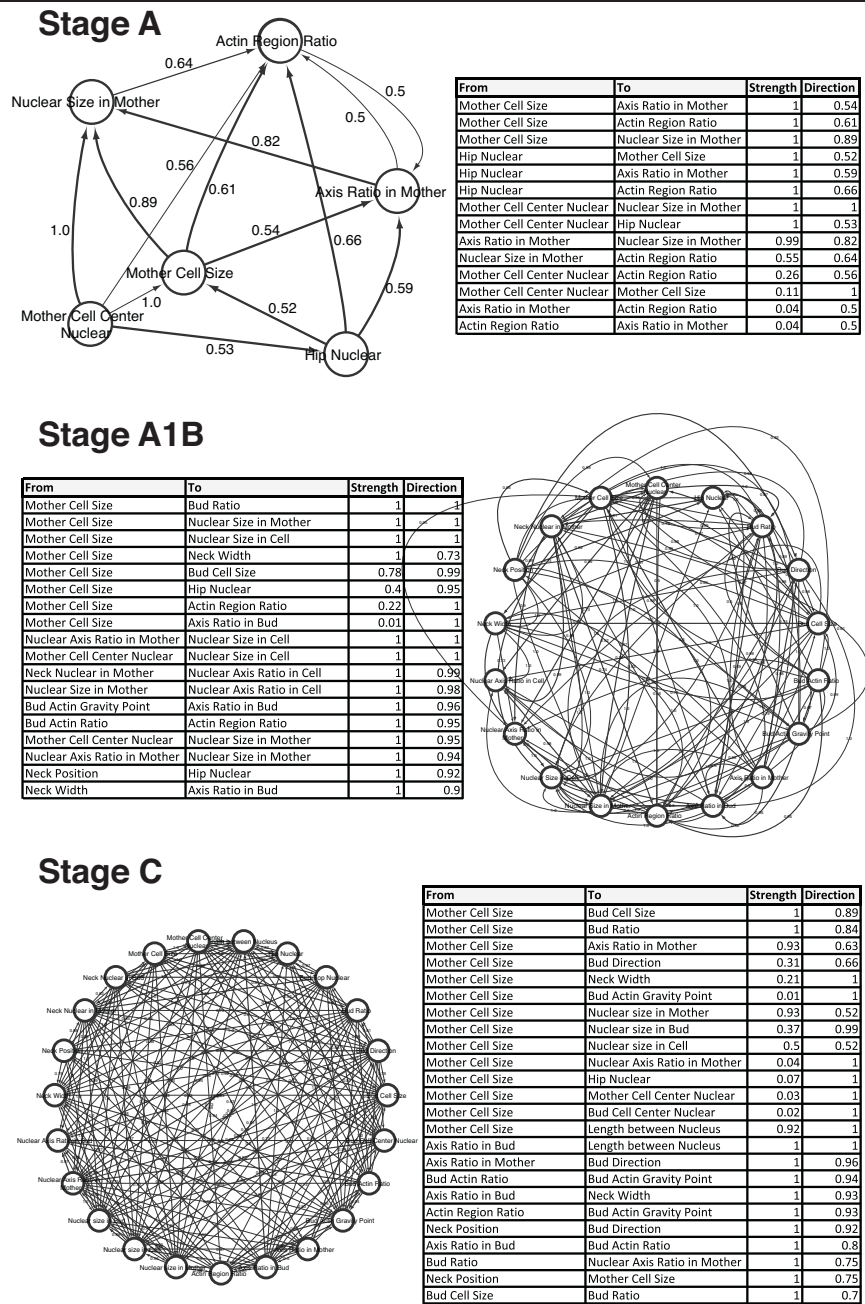


Figure 6.7: The strength and direction of the edges between the nodes (morphological phenotypes) is assessed by bootstrapping the single-cell measurements and analysing the quality of the relationships between nodes. The bootstrap analysis offers two measures, the strength of an arc, i.e., whether or not the arc should be between two nodes, and the direction of it. We bootstrapped the wild-type data for the three cell cycle stages. Each panel has a table with the strength and direction of the edges between the mother cell size and all the other phenotypes. The width of the edges in the networks are scaled to represent the strength between nodes and the weights on each nodes are the direction scores. Unlike the former standard networks in Figure 3, the bootstrapped networks in all three stages agree on the relationship between the mother cell size and the nuclear size in mother for the stages A and A1B. Although the direction between the mother cell size and the nuclear size in cell is reversed in stage C, the direction is 52%, signifying that the direction cannot be assessed with certainty.

6.4 Bayesian Networks Analysis of the Mutants

In the previous section, we analysed the wild-type dataset as a whole, meaning that there was only one strain to study, having a large number (Table 6.1) of single-cell measurements. Unlike the wild types, the mutants have a sparse number of single-cell measurements. During the acquisition process, no minimal threshold for the measurements was set, consequently implying some mutants to have only few single-cell measurements.

As depicted in Figure 6.8.A, there are more single-cell measurements for the stage A, fewer for the stage A1B and even fewer for the stage C. This situation complicates the analysis of the mutants dataset as it is not guaranteed that a mutant will have a sufficient number of measurements in each cell cycle stage. As a rule of thumb, it is accepted when using Bayesian networks that the number of measurements should be greater or equal the number of nodes (here, the phenotypes).

Although we want to keep the genetic degrees, i.e., the genetic association between the mutants to perform a gene ontology annotation test, we only keep the mutants having more single-cell measurements than the 95– percentile in their respective cell cycle stage when compared to the number of all single-cell measurements available. Once the mutants are selected for each stage, they are ”intersected”, i.e., their measurements throughout the three cell cycle stages are selected to form the final dataset of mutants. In total, 84 mutants are present in the final extraction.

As performed previously with the wild types, we explore the connectivity between the mother cell size and the nuclear size in mother. By computing the bootstrapped networks of the mutants, we look at the strength and direction between the mother cell size and the nuclear size in mother.

The mutants are sorted in the heatmap of Figure 6.8.B according to two classification values reported for the three cell cycle stages. The classification values are: the strength of an arc between the mother cell size and the nuclear size in mother and the direction from

the mother cell size to the nuclear size in mother. In this setting, it is possible to visually investigate clusters of mutants with similar network topologies.

In general, five categories of topologies appear in Figure 6.8.B (labelled from **I**, to **V**). The first category (**I**) has the mutants conserving the arc between the mother cell size and the nuclear size in mother. The second category (**II**) is similar to the first category but the strength and direction are less strong. The third category (**III**) shows mutants still having a strong arc between the mother cell size and the nuclear size in mother but the direction is not certain (around 50% for the direction). This observation is also reported for category **IV**. Finally, the last category (**V**) concerns the mutants where the arc between the mother cell size and the nuclear size in mother is not conserved.

An investigation of the GO annotations for the mutants in all did not provide any significant terms. This is certainly due to the small number of mutants present in each category. In addition, category (**V**) might actually be an artificial category or an artefact of the analysis, as it only has two mutants.

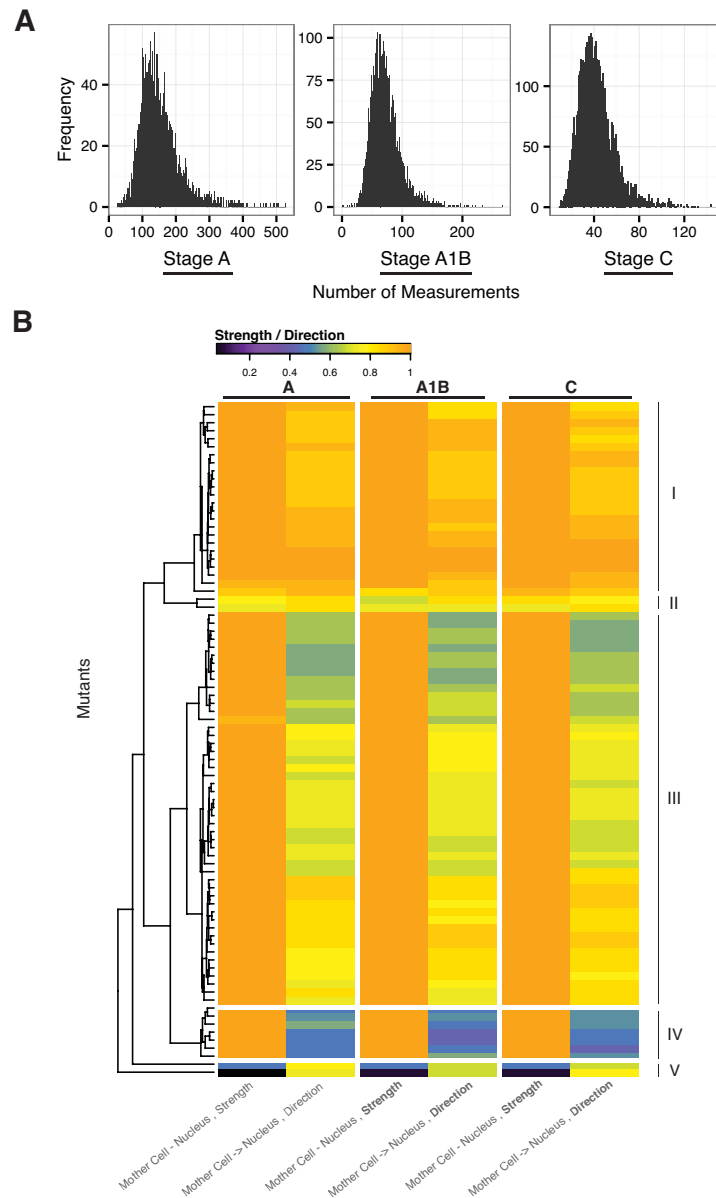


Figure 6.8: When the mutant phenotypes were acquired in the SCMD, the image acquisition process did not account for a minimum number of cells to measure. Therefore, the number of measurements available for each mutant in all three cell cycle stages varies, as shown in panel A. Such variation in the number of measurements requires a careful selection of the mutants. Using the empirical cumulative distribution of the three stages, we select 84 mutants and analyse their bootstrapped networks in a heatmap (panel B). The heatmap represents the strength and direction of the relationship between the mother cell size and the nuclear size in mother. The heatmap colour code represents the strength and direction (as a percentage) of an edge between the mother cell and the nucleus. One can distinguish five categories of causality, labelled from **I** to **V**. Mutants in category **I**, **II** and **III** are the mutants showing a strong and directed causality between the mother cell size and the nuclear size in mother, hence behaving like the wild types. The fourth category (**IV**) are the mutants where the direction of the arc between the mother cell size and the nuclear size in mother is not certain. Finally the fifth category (**V**) are the mutants not possessing an arc between the mother cell size and the nuclear size in mother.

6.5 Discussion

For decades now, biologists have tried to comprehend the biological phenomenon linking the growth of the cell and the nucleus. The interest is to understand if there is a possible regulation of the cell size and the nucleus size. In this type of setting, many scenarios are possible. In general, two accepted scenarios are: either the cell growth coordinates the nucleus size or oppositely, the growth of the nucleus coordinates the cell size.

In this research, we provide two methods to analyse the feasible scenarios. The first method employs a visual inspection of the cell and nucleus phenotypes by considering their conditional dependence on each other. The second method is based on the construction of Bayesian networks, learning the causality between the available phenotypes in the *Saccharomyces Cerevisiae* Morphological Database.

The visual inspection, provided by the DREVI calculation, showed that there is a good probability of seeing the first scenario, where the cell size controls the nucleus. However, we showed that although the second scenario with the nucleus controlling the cell has a lower R^2 statistics than the first scenario, it has a sufficiently good statistics fitting the model to induce or produce mistaken conclusions. We hypothesised that the DREVI score might be biased in this case because it only considers two phenotypes, i.e., the mother cell size and the nuclear size in mother, instead of considering the whole phenotypic landscape.

To account for all phenotypes and therefore consider the whole phenotypic landscape, we pursued the analysis of the wild types and mutants by learning Bayesian networks on their respective datasets. Once constructed, the networks offered a clear and visual inspection of the causality between the phenotypes.

For the Bayesian networks learned on the wild types, we showed that through the cell cycle, the mother cell size directly influences the nuclear size in mother in stage A (G1 phase) and A1B (S/G2 phases). This causation between the mother cell size and nuclear size is however lost in stage C (late M phase), possibly indicating that the cell loses its

control on the nucleus at division. We found similar results throughout the cell cycle, when equivalence classes and bootstrapped networks are constructed.

In the last part of the analysis, we focused on the mutants and their respective networks (for each strain). Unfortunately, due to the small number of measurements available for some strains, the available genome was not completely analysed. Only 84 mutants were studied. The mutants were clustered according to the strength and direction of their respective bootstrapped networks.

We showed that although there is a small number of selected mutants compared to the entire available genome (4718 in the *Saccharomyces Cerevisiae* Morphological Database), we can distinguish five categories of mutants. The first three categories contain the mutants having the same morphological connectivity as the wild types, i.e., the mother cell size influences the nuclear size in mother. The fourth category has mutants with an undecided direction between the mother cell size and the nuclear size in mother. Finally, the fifth category represents the mutants having no arc between the mother cell size and the nuclear size in mother.

A Gene Ontology inspection of the categories however did not produce any significant terms. We think that the number of mutants available in both categories is too small to compute any good cross-correlations with the entire genome. In addition, the fifth category of mutants might appear in our analysis as an artefact, as it contains only two mutants.

Chapter 7

Bacterial Population Growth and Variability in Cell Size and Division Process

7.1 Introduction

Colonies of bacteria are subject to environmental stresses and intrinsic modifications, both of which are considered as reasons for the cell cycle variability. Variability can be seen as noise at the level of gene expression (Elowitz et al. [34], Fraser & Kaern [39], Canela-Xandri [16], Aldridge et al. [4], Fudenberg & Imhof [41]). It is suggested (Boor [13], Dubnau & Losick [33]) that bacterial colonies need variability throughout the entirety of their growth as it creates different subcategories of bacteria and increases the chance of surviving of the colony. Although the subcategories of bacteria come from the same strain and therefore have the same genetic background, subtle phenotypic variations appear. Following this idea, phenotypic variability can be seen as a bet-hedging strategy which can provide a fitness advantage in fluctuating environments (Kussel & Leibler [66])

Recent studies (see, e.g., Guberman et al. [42], Wang et al. [144], Athale and Chaudhari [6]) that captured such variations using microfluidic devices also shed light on the competition for resources, such as nutrients or space, between different populations. Moreover,

these microfluidic measurements provide important quantitative data on the elongation rate of a single bacterium, its birth (i.e., its initial) length and its division (i.e., its final) length. The data is essential to correctly model and interpret the influence of variability on the bacterial cell cycle and the effect on a population of bacteria.

For decades biologists have thought of a possible cell size controller acting on the birth and division lengths, to allow the cell to sense its size and adapt its growth. This sensing effect, also named *cell-size homeostasis*, has largely been studied under the focus of two assumptions. The first one concerns a possible "timer", where cells grow for a determined amount of time before dividing. The second assumption, related to a "sizer", where the cells monitor their size and trigger their cell cycle once they have reached a specific size. The two assumptions have often been explored theoretically in the literature by Hill et al [46], Koch & Schaechter [59], Powell [107] and Amir [5] to name a few. Recent discoveries by Taheri-Araghi et al. [133], Jun & Taheri-Araghi [56] suggest however that instead of having a sizer or a timer, the cell must have an "adder". This adder acts as controller telling the cell how much mass it needs to divide depending on the cell birth length. This motivates developing models that also take into consideration such controllers in the modelling of bacterial growth.

Currently, models devised for bacterial growth can be separated into two categories. The first category often concerns the modelling of internal biochemical networks such as the formation of proteins (Hale et al. [43], Howard et al. [51], Howard and Ruttenberg [50]). The second category focuses on general aspects of the cell, such as the shape, the elongation rate, the volume, and the general aspects of a bacterial population (Tyson [138]). The theoretical and computational models typically rely on stochastic simulations (Bazin and Sanges d'Abadie [11], Charlebois and Kaern [18]), but agent-based models are also sometimes considered (Kreft et al. [63], Picioreanu [105]).

In this chapter we address the question of variability, seen as the variance in the birth and division lengths distributions of bacteria, to see if it is beneficial for the fitness of the whole population. A recent study from Schmidt et al. [121] modelled the effect of vari-

ability in the birth length. Their results showed that variability in the birth length does not affect the fitness of the population. As mentioned above, here however, we model variability by considering the distribution of birth and division lengths of bacteria together. That is, we want to revisit the problem analysed by Schmidt et al. [121] and study how variability in the division length affects the fitness of the population.

Specifically, we consider bacteria as one-dimensional objects whose growth follows an exponential growth process. Our main interest is then in the variability of the birth and division lengths and the effect of variability on the population of cells. To this end, we derive a simplified population balance equation (abbreviated PBE in the literature, see Ramkrishna [111], Hortsø[47]) applied to a growing colony of bacteria. The PBE indicates that the growth rate of the population has two main properties: the growth rate is time-dependent and the rate depends on the cell size distribution. To investigate the time-dependent nature of the growth rate, we devise an agent-based model, to enable stochastic simulations of the population growth. Then, to analyse the relation between the growth rate and the cell size distribution, we derive an analytical formulation for the growth rate at steady state. Thereafter, we add an additional parameter in the model to account for the relationship between the birth and division lengths. In other words, we verify the aforementioned hypotheses of a controller of the cell cycle. Finally, and in order to investigate the resistance to stress (Tyson and Diekmann [139], Pierrucci [106]), we integrate morphological statistics provided by the "mother machine" (by Jun Lab [144]) on genetically modified strains of *E. coli*.

The outcome of this framework is precisely that the fitness of the population strongly depends on the variance of the distributions of the birth and division lengths and also the cell distribution. According to the initial cell distributions, the growth rate will converge faster to its steady-state (fixed) rate if the cell distribution is uniform across all cell lengths and not simply centred around one particular length. After a certain time, as the cell distribution has reached its steady state, the growth rate of the population converges to a fixed rate. The time needed to converge depends again on the initial cell distribution. At steady state, we show that the variances of the birth and division lengths are still impor-

tant as they affect the growth rate. More specifically, we show that the growth rate of the population can be related to the cell cycle time (also called the doubling time) of the cells.

The first section introduces the population balance equation (PBE) and the derivation of the time-dependent growth rate. In addition, the algorithm for the agent-based model is introduced. The second section then introduces the calculations to obtain the growth rate of the population at steady state, based on the distribution of cell cycle times. The third section is a continuation of the second section, where the derivation of the distribution of the cell cycle times, when the birth and division lengths are linearly connected. The fourth section provides an algorithm to extract and analyse experimental data. The data are then compared to the theoretical models. Finally, the last section offers a summary and discussion of the obtained results.

7.2 Time-dependent Growth Rate of the Population

7.2.1 Elongation of a Single Cell

Quantification of variability in the cell cycle is possible when looking at general properties of bacterial elongation as well as at the growth of bacterial colonies. The latter is often described by different phases such as the lag, exponential, stationary and decay phases (Monod [90], Turchin [137]). Our main interest is on the exponential phase as in this phase cells grow and divide in a regular manner. The exponential phase is often represented by a standard Malthusian kinetics [80]. Specifically, defining $N(t)$ as the number of bacteria in the population at time t , the Malthusian growth is referred to a linear ordinary differential equation

$$\frac{d}{dt}N(t) = KN(t) \quad , \quad N(0) = n_0, \quad (7.1)$$

where K is called the growth rate of the population and n_0 is the initial number of bacteria at time $t = 0$.

The rate K in Eq. (7.1) does not provide any information on the cell cycle variability of a single bacterium which implies that we also need to consider the growth of a single bacterium as an additional component of our model. Therefore and to be consistent with

our notation, we define the elongation of an individual bacterium by

$$\frac{d}{dt}L(t) = r(L(t)), \quad (7.2)$$

with $L(t)$ the length of a bacterium at time t and $r(L) = r(L(t))$ its growth process (also called the reaction rate). In both eukaryotic and prokaryotic organisms, it is generally agreed (Cooper [25], Leslie [74]) that the elongation of a cell is either exponential or linear. Recent advances in microfluidic microchemostat [98] have shown that a bilinear growth, i.e., a mixture of exponential and linear elongation is also possible as for instance in *S. pombe*. However, for simplicity here, we only consider the exponential and linear elongations. The linear and exponential elongations can respectively be rewritten as

$$r(L) = \lambda_c \quad \text{or} \quad r(L) = \lambda_e L. \quad (7.3)$$

where λ_c and λ_e are the elongation rates of the bacterium for the linear and exponential growth respectively. In each case, it is important to note that λ_c and λ_e have different units. For the rest of our analysis, we only consider the exponential elongation. This choice is motivated by the data provided by the "mother machine" [144], the microfluidic device, which shows evidence on the exponential elongation of the cell. A recent study by Osella et al. [100] also analysed the data from the mother machine and showed that bacteria elongate exponentially.

In the exponential growth case and when there is no variability on the cell size, the bacterium will grow, for example, from a birth length μ and will divide at a division length 2μ (see Figure 7.1.A). In this setting, it can be shown for the exponential elongation that the colony and the individual bacterium both possess the same growth and elongation, i.e., $\lambda_e = K$.

However, the ideal case with no variability in the birth and division lengths is not realistic. Therefore, instead of having fixed birth and division lengths, we assume the lengths to follow probability distributions. Hence, with both lengths represented by distributions, we denote the parameter α as the standard deviation of the birth length and the parameter β as

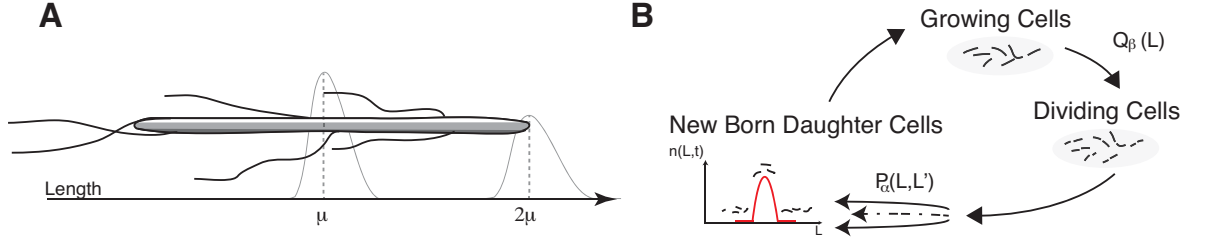


Figure 7.1: Description of the morphological parameters and its dependencies with the probability distribution functions. A. The growth of the bacterium can either be considered without variability and in this case the bacterium will grow from a length μ and will divide at a length 2μ . On the contrary, there is variability in the lengths, the bacterium will grow from a length distributed around μ and will divide at a length distributed around 2μ (B) Representation of the global growth and division of the system.

the standard deviation of the division length of the bacterium. Specifically, let $P_\alpha(L, L')$ be the probability density function that a mother of length L' will give birth to a daughter of length L and a daughter of length $L' - L$ and let $Q_\beta(L)$ be the "division rate" density that a bacterium that has not divided below L will divide at length in $(L, L + \Delta L)$. To be coherent, we set $P_\alpha(L, L')$ to zero when $L \geq L'$ so that the daughter cell cannot have a larger length than the mother cell at division. The growth and division process is sketched in Figure 7.1.B

In what follows, we employ the following division rate $Q_\beta(L)$:

$$Q_\beta(L) := \frac{r(L)h_\beta(L)}{\int_L^\infty h_\beta(l)dl}, \quad (7.4)$$

where $h_\beta(L)$ is the density of lengths at division, i.e. the density of lengths in the population when the cells divide. In other words, the division rate is the probability of having a cell dividing at a length L times the rate to reach this length. The rate is then normalised by the probability of having lengths greater than L . See [47, 23, 30] for more information on the derivation of the division rate as well as alternative approaches.

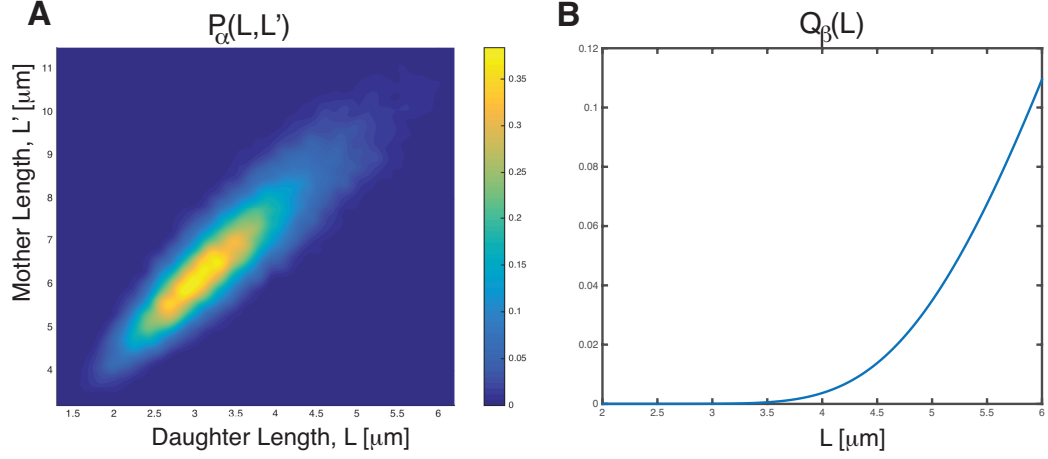


Figure 7.2: Description of the probability distributions employed in the population balance equation. A. Probability of a mother with length L' to give birth to a daughter of length L . B. Division rate of the cells according to their length. In both panels, the average birth length is $\mu = 3.3$ and the theoretical average division length is twice the birth length, i.e., 6.6.

We set the distributions of birth and division lengths $P_\alpha(L, L')$ and $h_\beta(L)$ respectively as log-normal distributions. This choice is motivated by the observation that a Gaussian distribution would produce cells with negative lengths, depending on the mean and variance of the distributions. In addition, the log-normal distribution is known to fit experimental data well, e.g., as shown in [144] (and also available, only for the strain MG1655, in the Appendix in Figure E.1). The distribution $P_\alpha(L, L')$ and the division rate $Q_\beta(L)$ are illustrated in Figure 7.2 in panels A and B respectively.

7.2.2 The Population Balance Equation

Due to probabilistic changes induced by the variations in length of a single bacterium, a population balance equation ([111], [47]) is suitable to describe bacterial growth. Population balance equations are a special type of integro-partial differential equation defining how a population of distinct entities evolve over time. More specifically, PBE are based on the birth, growth and death of the entities in the population.

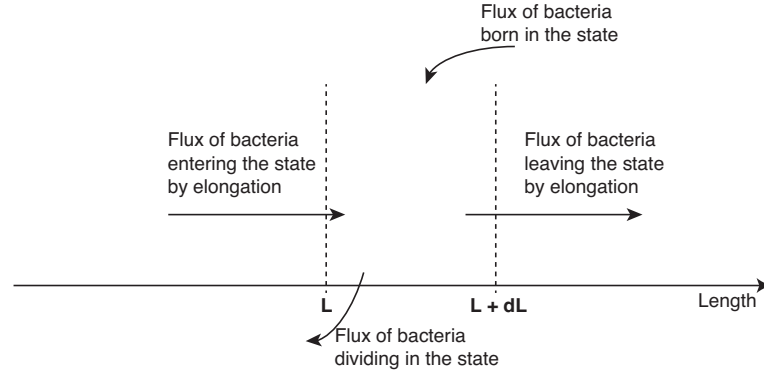


Figure 7.3: Description of the accumulation rates suggested to model the population balance equation. The four fluxes are the bacteria entering and leaving the interval $[L, L + dL]$ by growth and the bacteria being born or that divided in the interval.

We first describe the number $N(L, t)$ of bacteria with a specific length L at time t by

$$N : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}. \quad (7.5)$$

To obtain a mathematical formulation of $N(L, t)$, we follow the work of Hjortsø [47]. That is by considering that bacteria will grow within a certain time interval dt say from a length L to $L + dL$, in a closed environment, i.e., in a locally closed subspace of the environment where the bacteria do not die and are not separated/flushed from the colony.

The rate of bacterial accumulation in this interval can be phrased as:

$$\begin{aligned} \text{Rate of accumulation of bacteria} = & \text{rate of birth of bacteria} \\ & + \text{flux of bacteria entering the interval} \\ & - \text{flux of bacteria leaving the interval} \\ & - \text{rate of bacterial division,} \end{aligned}$$

which is also illustrated in Figure 7.3. The rate of accumulation of bacteria is the time derivative of $N(L, t)$ multiplied by dL :

$$\text{Rate of accumulation of bacteria} = \frac{\partial N(L, t)}{\partial t} dL. \quad (7.6)$$

Moreover, the flux of bacteria entering and leaving the interval can be represented using the reaction rate $r(L)$ of a single bacterium as:

$$\text{Flux of bacteria entering - leaving the interval} = r(L)N(L, t) - r(L + dL)N(L + dL, t). \quad (7.7)$$

Similarly, the rate of bacterial division is the division rate $Q_\beta(L)$ times the number of cells with a certain length L at time t in the population:

$$\text{Rate of bacterial division} = Q_\beta(L)N(L, t)dL. \quad (7.8)$$

Note that the rate of birth from division of bacteria with length L' is proportional to $Q_\beta(L')N(L', t)dL'$. Also, the fraction of bacteria born is proportional to $P_\alpha(L, L')dL$. Thus, it is necessary to integrate the product over all possible lengths, i.e.,

$$\text{Rate of birth of bacteria} = 2 \int_L^\infty Q_\beta(L')N(L', t)dL'P_\alpha(L, L')dL. \quad (7.9)$$

The factor of two appears here because each bacterium produces, at division, two new daughters in the colony and also due to $P_\alpha(L, L') = P_\alpha(L' - L, L')$.

Combining these rates, dividing by dL and taking the limit of dL to zero, eventually yields to the standard PBE

$$\frac{\partial N(L, t)}{\partial t} + \underbrace{\frac{\partial}{\partial L} [r(L)N(L, t)]}_{\text{Growth}} + \underbrace{Q_\beta(L)N(L, t)}_{\text{Division}} = 2 \underbrace{\int_L^\infty Q_\beta(L')P_\alpha(L, L')N(L', t)dL'}_{\text{Birth}}, \quad (7.10)$$

with the containment conditions (CC)

$$\text{(CC)} \quad r(0)N(0, t) = 0 = r(\infty)N(\infty, t) \quad , \quad t \in \mathbb{R}_{\geq 0}. \quad (7.11)$$

The containment condition assures that cells cannot be smaller than a length 0 and larger than ∞ .

In Eq.(7.10), the PBE relates in fact three processes. Firstly, the bacterial growth, allowing bacteria to grow according to their reaction rate $r(L)$. Secondly, the division, also called "breakage" in the literature, of bacteria into two daughters, depending on the division rate $Q_\beta(L)$ and finally the birth of new bacteria, which acts as a source term.

Notice that Eq. (7.10) is an integro-differential equation and the total number of cells at time t , denoted $N(t)$ is

$$N(t) := \int_0^\infty N(L, t) dL. \quad (7.12)$$

Normalising $N(L, t)$ by $N(t)$, i.e.,

$$n(L, t) = \frac{N(L, t)}{N(t)}, \quad (7.13)$$

it is possible to express $N(t)$ as a differential equation having the division rate $Q_\beta(L)$ and the density $n(L, t)$ as parameters.

Integrating Eq.(7.10) over all possible lengths and substituting Eq.(7.12) and Eq.(7.13) in it, yields

$$\begin{aligned}
 & \frac{\partial}{\partial t} \int_0^{\infty} N(L, t) dL + \underbrace{\frac{\partial}{\partial L} \int_0^{\infty} r(L) N(L, t) dL}_{= 0 \text{ due to (CC)}} + \int_0^{\infty} Q_{\beta}(L) N(L, t) dL \\
 & \qquad \qquad \qquad = 2 \int_0^{\infty} \int_L^{\infty} Q_{\beta}(L') P_{\alpha}(L, L') N(L', t) dL' dL. \\
 \Rightarrow & \frac{\partial}{\partial t} \int_0^{\infty} N(L, t) dL + \int_0^{\infty} Q_{\beta}(L) N(L, t) dL \\
 & \qquad \qquad \qquad = 2 \int_0^{\infty} P_{\alpha}(L, L') dL \int_L^{\infty} Q_{\beta}(L') N(L', t) dL'. \quad (7.14) \\
 \Rightarrow & \frac{\partial}{\partial t} \left(N(t) \underbrace{\int_0^{\infty} n(L, t) dL}_{=1} \right) + N(t) \int_0^{\infty} Q_{\beta}(L) n(L, t) dL \\
 & \qquad \qquad \qquad = 2N(t) \int_0^{\infty} Q_{\beta}(L) n(L, t) dL. \\
 \Rightarrow & \frac{d}{dt} N(t) = N(t) \int_0^{\infty} Q_{\beta}(L) n(L, t) dL.
 \end{aligned}$$

Let

$$K(t) := \int_0^{\infty} Q_{\beta}(L) n(L, t) dL, \quad (7.15)$$

be the rate at which the population size increases. Eq.(7.15) firstly indicates that the population is a time-dependent process. The growth rate $K(t)$ is given by both the division rate of a bacterium and the distribution $n(L, t)$ of the bacterial lengths. Consequently, this indicates that the growth rate of the population will depend on both parameters α and β , and the density of bacterial lengths. Next, we derive an integro-differential equation for the density $n(L, t)$ by straight forward algebra.

To this end, by substituting Eq.(7.13) and Eq.(7.14) in Eq.(7.10) yields:

$$\begin{aligned} \frac{\partial}{\partial t} (n(L, t)N(t)) + \frac{\partial}{\partial t} [r(L)n(L, t)N(t)] + Q_\beta(L)n(L, t)N(t) \\ = 2 \int_L^\infty Q_\beta(L')P_\alpha(L, L')n(L', t)N(t)dL'. \end{aligned}$$

An application of the chain rule gives

$$\begin{aligned} \Rightarrow \frac{\partial n(L, t)}{\partial t} N(t) + n(L, t) \frac{d}{dt} N(t) + N(t) \frac{\partial}{\partial t} [r(L)Q_\beta(L)] + Q_\beta(L)n(L, t)N(t) \\ = 2N(t) \int_L^\infty Q_\beta(L')P_\alpha(L, L')n(L', t)dL'. \\ \Rightarrow \frac{\partial n(L, t)}{\partial t} N(t) + n(L, t)N(t) \int_0^\infty Q_\beta(L)n(L, t)dL \\ + N(t) \frac{\partial}{\partial t} [r(L)Q_\beta(L)] + Q_\beta(L)n(L, t)N(t) = 2N(t) \int_L^\infty Q_\beta(L')P_\alpha(L, L')n(L', t)dL'. \end{aligned}$$

As $N(t) > 0$, one finally obtains

$$\begin{aligned} \frac{\partial n(L, t)}{\partial t} + \frac{\partial}{\partial L} [r(L)n(L, t)] + Q_\beta(L)n(L, t) = 2 \int_L^\infty Q_\beta(L')P_\alpha(L, L')n(L', t)dL \\ - n(L, t) \int_0^\infty Q_\beta(L)n(L, t)dL. \end{aligned} \quad (7.16)$$

In this density-based PBE (7.16), we recognise the growth, division and birth terms, that are also present in the standard PBE form. The last term in the right hand side of Eq.(7.16) is a normalisation term. To the best of our knowledge, no studies have provided any analytical solutions of the PBE in Eq. (7.16). It can be shown (Langlais [67]) that when the growth process $r(L)$ is constant, there exist a solution to the PBE for simple case when the variables $N(L, t)$, P_α or Q_β are either independent of time or independent of the length. In other situations, the existence of a solution for the PBE is not guaranteed. In terms of numerical

simulations, in a non-trivial case with bacteria evolving in space, finite element and spectral methods have been employed (Rigopoulos and Jones [113], Mahoney and Ramkrishna [79], Marchisio and Fox [84], Mantzaris et al. [81, 82, 83]). In this study, we decided to compute the growth rate $K(t)$ in Eq. (7.15) by simulating the behaviour of the PBE through an agent-based model (ABM). Here, the agent-based model allows us to recreate bacterial growths without having to physically and genetically modify the cells. Thus, the ABM allows us to run fast and reliable virtual experiments where morphological parameters can be modified in an easy manner and where the growth, division and birth of cells can be simulated.

7.2.3 The Agent-Based Model

In the last section, we introduced the PBE and we derived an equation for the growth rate of the population, $K(t)$. In order to approximate the growth rate, we develop an agent-based model that simulates the growth of the population at the individual cell level. Each cell is represented by an agent growing and dividing independently from the other cells. The simulations provided by the model allow us to capture the cell distribution $n(L, t)$ at each time thus permitting to compute the growth rate $K(t)$ according to the parameters α and β .

The model is entirely based on the elongation of a bacterium and the creation of the two new daughter cells at the time of division. Each bacterium is added to a stack of bacteria and is treated independently from its neighbours. All bacteria grow according to their respective growth process $r(L)$. When a new bacterium is born, it is attributed a random birth length L , according to its mother length L' and the distribution $P_\alpha(L, L')$. We set $P_\alpha(L, L')$ as a log-normal distribution,

$$\ln \mathcal{N}(L'/2, \alpha^2). \quad (7.17)$$

where α^2 is the variance around the average birth length, $L'/2$.

At the same time, the model already sets the division lengths of the bacterium by sampling a random number taken from the distribution $h_\beta(L)$, which is also assumed to be a log-

normal distribution

$$\ln \mathcal{N}(2\mu, \beta^2). \quad (7.18)$$

where β^2 is the variance around the division length and μ is, in this setting, the average birth length of the cells in the population.

Initially, the model starts at time $t = 0$ with n_0 agents in the population. The distribution $n(L, 0)$ is set by the user (see Eq. 7.20). The model stops when the maximum number $N_{\max}$ of cells is reached in the population.

After each time step dt , the model produces statistics on the cell distributions, i.e., the distribution $n(L, t)$, and also compute the growth rate $K(t)$. The full algorithm employed in the agent-based model is shown in the schematic of Figure 7.4 and all the parameters are described in Table 7.1.

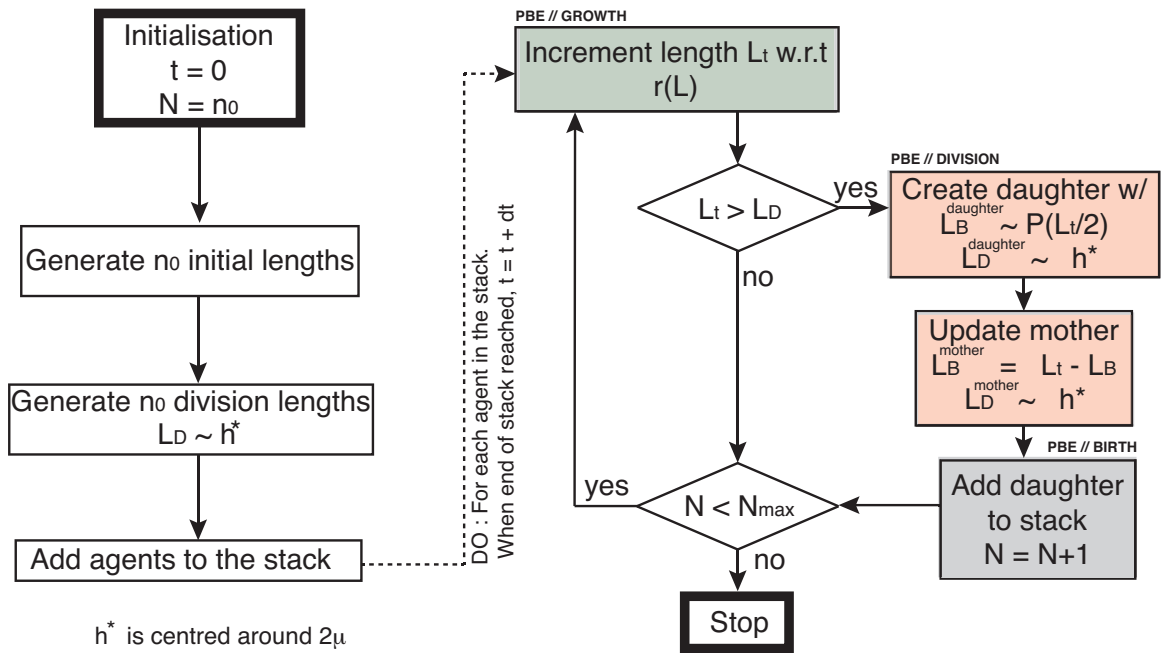


Figure 7.4: Procedure of the agent-based model. The green, red and grey compartments represent respectively the growth, division and birth as defined in the population balance equation. The agent-based model considers that cells grow independently and, that each cell creates new daughters at the division time.

Symbol	Units	Description
N	[1]	Current number of bacteria in the population
n_0	[1]	Initial number of bacteria in the population
$N_{\max}$	[1]	Maximum number of bacteria in the population
T	[min]	Current time in the simulation
dt	[min]	Time step
L	$[\mu m]$	Current length of the bacterium
L_0	$[\mu m]$	Initial length of the bacterium
L_{final}	$[\mu m]$	Final length of the bacterium
μ	$[\mu m]$	Average length of the bacterium at birth
α, β	[1]	Variance around the birth and the division lengths respectively
$r(L)$	$[\mu m \cdot \text{min}^{-1}]$	Elongation rate of the bacterium

Table 7.1: Description of the parameters and variables employed in the agent-based model. The symbols and units are set to reflect values found in real-life experiments.

7.2.4 Computation of the Growth Rate $K(t)$

We recall that the time-dependent growth rate of the population, $K(t)$, is given by

$$K(t) = \int_0^{\infty} Q_{\beta}(L)n(L, t)dL. \quad (7.19)$$

The integrand indicates that the growth rate depends on both the value of the division rate $Q_{\beta}(L)$ and the distribution of cells $n(L, t)$, as we have emphasised earlier already.

In order to see the influence of the initial cell distribution, $n(L, 0)$ we test three different scenarios. The first scenario (**I**) sets the initial cell distribution to be centred in μ according to a log-normal distribution. The second scenario (**II**) samples the cell distribution over the interval $[\mu, 2\mu]$ and finally, the third scenario (**III**) considers the steady-state/equilibrium distribution of $n(L, t)$ as an initial cell distribution. The three scenarios are summarised below:

$$\begin{aligned}
 \text{(I)} \quad n(L, 0) &= \ln \mathcal{N}(\mu, \alpha^2), \quad \forall L, \\
 \text{(II)} \quad n(L, 0) &= \mathcal{U}(\mu, 2\mu), \quad \forall L, \\
 \text{(III)} \quad n(L, 0) &= \varphi, \quad \forall L,
 \end{aligned} \quad (7.20)$$

where φ denotes the steady state of $n(L, t)|_{t=\infty}$.

In order to distribute the cells according to the steady state φ , the agent-based model is first run for a sufficiently long final time T , where the absolute difference between $n(L, T)$ and $n(L, T - \Delta t)$ is smaller than 10^{-5} . Then, the agent-based model is restarted and the cells

are sampled according to the empirical distribution of $n(L, T)$.

The agent-based model initially starts with $N(0) = 2^{10}$ cells and continues until the total number of cells in the population reaches $N_{\max} = 2^{15}$. As we are interested in the behaviour of $K(t)$, we want to look at its value on the long run, to see if $K(t)$ converges to a steady-state value. To this end, when the agent-based model has reached the maximum number of cells in the population, cells are randomly removed to decrease the population to 2^{12} cells, to retain a certain level of accuracy of the density $n(L, t)$, but also to run the agent-based model for longer times. We opt for this approach to also avoid reaching memory boundaries. The agent-based model is written in C++ and uses the computer memory to store the agents, i.e., the cells. During our preliminary test of the agent-based model, we have noticed that when the number of cells exceeds a value greater than 2^{15} , the operating system (here, Linux and OS X) crashes. Therefore, we have come with the solution of restricting the total number of cells to 2^{12} .

To investigate the role of variability in the model, as suggested by the division rate appearing in the time-dependent growth rate of the population (c.f., Eq. (7.15)), we perform the agent-based model simulations for different values of α and β . Here, we set the pairs as :

$$(\alpha, \beta) = \{(0.1, 0.1), (0.1, 0.2), (0.2, 0.1)\}, \quad (7.21)$$

to encompass various values of standard deviation and have a preliminary idea of the behaviour of $K(t)$.

The results for the three initial cell distributions and the (α, β) parameters are shown in Figure 7.5. We see that for the log-normal and uniform scenarios, the growth rate $K(t)$ oscillates for a certain period of time and slowly starts to converge to a constant, regardless of the choice of (α, β) . This is explained by the nature of the growth rate. Indeed, the cell distribution within the integrand of Eq.(7.19) evolves with time, therefore forcing the growth rate to evolve as well. For the log-normal and uniform scenarios, the growth rates converge to their constant value after 180 and 100 minutes (in the simulation) respectively.

To explain the fact that the log-normal distribution requires more time to reach the steady-state growth rate, one needs to consider the shape of $n(L, t)$. Indeed, the distribution of cells is transient throughout the time and slowly becomes wider while cells are dividing. Unlike the log-normal distribution, the uniform distribution already has cells in their division stage, therefore modifying the shape of $n(L, t)$. Hence, it takes more time for the log-normal distribution to reach the growth rate steady-state value. For the third scenario (III), i.e., starting with $n(L, 0)$ sampled from $\varphi(L)$, the growth rate $K(t)$ has already converged to a constant rate. This is explained similarly as the convergence for the log-normal and uniform distributions. At steady state, the distribution $n(L, t)$ which then equals $\varphi(L)$ is constant for all length values. Thus, the integrand for $K(t)$ is constant as the division rate $Q_\beta(L)$ and $n(L, t)$ do not depend on the time.

Although numerical integration creates variability in the computation of the growth rate, we see (black, blue and red trajectories in Figure 7.5) that the α and β values influence the value of the growth rate. For instance, we see that when the variance around the birth length is 0.1 and the variance around the division length is 0.2, the growth rate $K(t)$ is larger than for the set of variances (0.1, 0.1) and (0.2, 0.1).

In all settings of variability and for all initial cell distributions, we see that the growth rate $K(t)$ converges to a fixed growth value when the cell distribution converges to its steady state, i.e., when $n(L, t)$ does not depend anymore on the time t . In the next section we give an analytical formulation for the steady-state value of $K(t)$ by considering a heuristic based on the cell cycle time, i.e. the time needed for the cells to grow and divide.

7.3 Growth Rate of the Population at Steady State

We have shown that the growth rate of the population, denoted by $K(t)$, strongly depends on variability around the birth and division lengths, denoted by α and β respectively during the exponential growth phase. We also saw that the rate $K(t)$ converges to a constant/stationary value when the cell distribution $n(L, t)$ reaches its steady state. In this section, we are now interested in understanding further properties of the growth rate of the

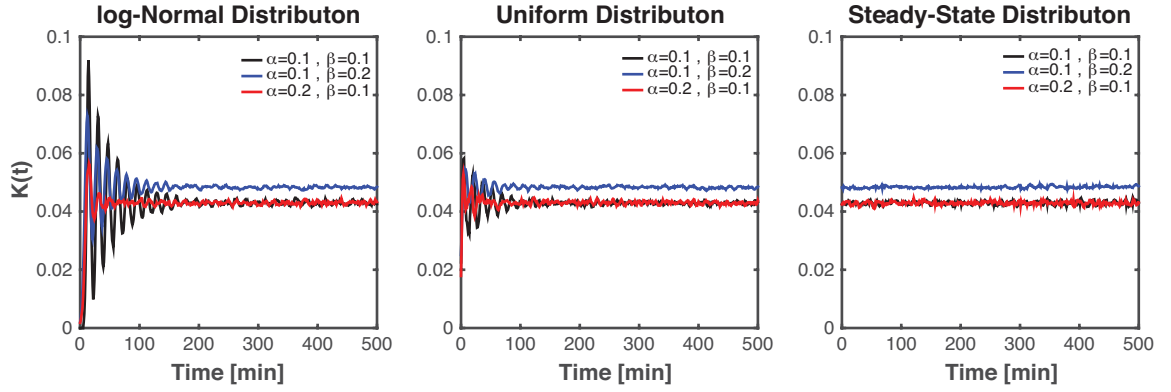


Figure 7.5: Sample trajectories produced by the agent-based model for the value of $K(t)$ with three initial distributions $n(L, 0)$. The first distribution is log-normal, centred around μ with variance α^2 . The second distribution is uniform on $[\mu, 2\mu]$ and the third distribution is the steady state of $n(L, t)$. The uniform distribution seems to accelerate the convergence of the growth rate $K(t)$ when compared to the log-normal distribution (see explanation in text). One can also see that the steady-state distribution does not change the value of $K(t)$, thus indicating that the growth rate already converged to its steady-state (fixed) value. For the log-normal and uniform distributions, large oscillations appear at the beginning of the simulation due to the transient nature of the density $n(L, t)$. Then, as the number of cells in the population increases, the density $n(L, t)$ stabilises and converges to its steady stage $\varphi(L)$.

population at its steady state. During the steady-state phase, the density $n(L, t)$ becomes constant. As shown in Figure 7.5, the time-dependent growth rate $K(t)$ converges to a constant. Here we denote this constant growth rate by K . In this section, we investigate the growth rate of the population at steady state, based on an analytical relationship between the growth rate K of the population and the distribution of the cell cycle time (or also called "time until division") of each cell (see Painter & Marr [108]). Here we denote by $f_\tau(t)$ the density function of the cell cycle times. According to the results of Painter & Marr, the growth rate relates to the cell cycle time through the identity:

$$\int_0^\infty 2e^{-Kt} f_\tau(t) dt = 1. \quad (7.22)$$

To derive the relation of Eq.(7.22), let us consider a population consisting of $N(0)$ cells at steady-state growth. If the initial number of cells is large enough, one can write $N(t)$, the

total number of cells

$$N(t) = N(0)e^{Kt}. \quad (7.23)$$

Here, one can count the total number of cells in the population at time $T > 0$ with a portion R of cells present at time 0 that have not divided yet and a remaining portion of cells produced by division in the interval $[0, T]$. At any time t , with $0 \leq t \leq T$, the rate of production of cells (in the population) that will be still growing and not dividing at time T is

$$2KN(0)e^{Kt} \int_{T-t}^{\infty} f_{\tau}(s)ds. \quad (7.24)$$

where $f_{\tau}(t)$ is the density function representing the cell cycle times of the cells. The meaning of Eq.(7.24) as to be understood as the probability of having cells with a cell cycle time greater than $T - t$, times the number of cells at time t multiplied by two times the rate of production of cells in the population.

Thus, the total number of new cells formed in the interval $[0, T]$ is

$$\int_0^T \int_{T-t}^{\infty} 2KN(0)e^{Kt} f_{\tau}(t)dt ds. \quad (7.25)$$

Substituting $x = T - t$ in Eq.(7.25) yields

$$\int_0^T \int_x^{\infty} 2KN(0)e^{K(T-x)} f_{\tau}(t)dt dx, \quad (7.26)$$

so that the total number of cells at time T is the portion R of cells that have not divided yet and the remaining portion found in Eq. (7.26), i.e.,

$$\underbrace{N(0)e^{KT}}_{= N(t)} = R + N(0)e^{KT} \int_0^T \int_x^{\infty} 2Ke^{-Kx} f_{\tau}(t)dt dx. \quad (7.27)$$

Dividing by $N(0)e^{KT}$ gives

$$\int_0^T \int_x^\infty 2Ke^{-Kx} f_\tau(t) dt dx = 1 - \frac{Re^{-KT}}{N(0)}. \quad (7.28)$$

When T tends to infinity, the portion R of cells that have not divided will decrease more slowly than e^{-KT} forcing the fraction $\frac{Re^{-KT}}{N(0)}$ to approach zero. Therefore, can rewrite Eq.(7.28) as

$$\int_0^\infty \int_x^\infty 2Ke^{-Kx} f_\tau(t) dt dx = 1. \quad (7.29)$$

We recall that this calculation is valid for any time t .

Thus, interchanging the order of integration yields:

$$\begin{aligned} & \int_0^\infty \int_0^t Ke^{-Kx} 2f_\tau(t) dx dt = 1 \\ \iff & \int_0^\infty \left(\int_0^t Ke^{-Kx} dx \right) 2f_\tau(t) dt = 1 \\ \iff & \int_0^\infty (-e^{-Kt} + 1) 2f_\tau(t) dt = -1 \\ \iff & - \int_0^\infty 2e^{-Kt} f_\tau(t) dt + \int_0^\infty 2f_\tau(t) dt = 1 \\ \iff & \int_0^\infty 2e^{-Kt} f_\tau(t) dt = 1 \end{aligned} \quad (7.30)$$

Finally, we also mention that alternative derivations, including branching processes and renewal theory [37, 44, 12], finds the same equation as in Eq.(7.22), which links the growth rate K at steady state and the distribution of cell cycle times $f_\tau(t)$.

In the next section, we describe and derive an analytical distribution for the cell cycle times varying according to the parameters α and β .

7.3.1 Computation of the Cell Cycle Time Distribution

The cell cycle time of a single cell can be computed using the exponential growth of its length by solving the ordinary differential equation

$$\frac{dL(t)}{dt} = r(L(t)) = \lambda L(t), \quad (7.31)$$

where λ is the elongation rate of the cell. That is

$$L(t) = L(0)e^{\lambda t}. \quad (7.32)$$

If we moreover assume that the cell starts growing from a birth length L_B and divides at a division length L_D , then its cell cycle time τ is calculated as

$$\tau = \frac{1}{\lambda} \log \left(\frac{L_D}{L_B} \right). \quad (7.33)$$

For the theoretical case where no variability is present around the birth and division length, the ratio L_D/L_B is simply 2 and thus the distribution of cell cycle times $f_\tau(t)$ is a Dirac delta function, with mass in $\frac{1}{\lambda} \log(2)$. In the case where we have variability for the birth and division lengths, the ratio L_D/L_B is no longer deterministic and therefore we need to compute the distribution of cell cycle times, according to the parameters α and β . To emphasise this parametric dependency, we will use the notation $f_\tau^{\alpha,\beta}(t) = f_\tau(t)$ in what follows.

In the next paragraphs, we present an analytical derivation of $f_\tau^{\alpha,\beta}(t)$ by considering the creation of new cells, originating from the division of mother cells. To compute the distribution of cell cycle times, we hypothesise the creation of new cells as follows (**Heuristic**):

1. A cell ready to divide possesses a division length L_D^1 that follows (according to our experimental data) a log-normal distribution centred at 2μ with variance β^2 :

$$L_D^1 \sim \ln \mathcal{N}(2\mu, \beta^2).$$

2. Once the cell has divided, a new daughter cell is created according to a log-normal

distribution centred around half the mother length, with variance α^2 :

$$L_B \sim \ln\mathcal{N}(L_D^1/2, \alpha^2).$$

3. The cell cycle time is computed as

$$\tau = \frac{1}{\lambda} \log \left(\frac{L_D^2}{L_B} \right),$$

where the new division length is (already) set as

$$L_D^2 \sim \ln\mathcal{N}(2\mu, \beta^2)$$

In other words, the cell cycle time τ can be computed by considering the creation of daughter cells by the division of the mother cells, for which the new division length is then computed.

To compute the distribution of τ , one thus has to compute the distribution of the ratio

$$Z = L_D^2/L_B \tag{7.34}$$

in Eq. (7.33). To this end, we use the result by Curtiss [31] on the distribution of a ratio of random variables, given by:

$$f_Z(z) = \int_{-\infty}^{\infty} |\xi| f_{L_B, L_D^2}(\xi, z\xi) d\xi, \tag{7.35}$$

where $f_{L_B, L_D^2}(l_B, l_D)$ is the joint density of the two random variables L_B and L_D^2 , which is found by recursively computing the conditional densities of L_D^1 and L_B , according to our heuristic.

Thus, by considering the birth and division lengths as log-normal random variables, as

defined before (in **Heuristic**), we can compute the distribution of L_D^1 as

$$f_{L_D^1}(l_D^1) = \frac{1}{2} \frac{\sqrt{2} e^{-\frac{1}{2} \frac{(\log(l_D^1) - \log(2\mu))^2}{\beta^2}}}{l_D^1 \sqrt{\pi} \beta}. \quad (7.36)$$

Using this probability density function of L_D^1 , we then compute the conditional density L_B given L_D^1 as

$$f_{L_B|L_D^1}(l_B|l_D^1) = \frac{1}{2} \frac{\sqrt{2} e^{-\frac{1}{2} \frac{(\log(l_B) - \log(l_D^1/2))^2}{\alpha^2}}}{l_B \sqrt{\pi} \alpha}. \quad (7.37)$$

This, allows us to compute the joint density of L_B and L_D^1 using Bayes' rule:

$$\begin{aligned} f_{L_B, L_D^1}(l_B, l_D^1) &= f_{L_B|L_D^1}(l_B|l_D^1) f_{L_D^1}(l_D^1) \\ &= \frac{1}{2} \frac{e^{-\frac{1}{2} \frac{(\log(l_B) - \log(l_D^1/2))^2}{\alpha^2}} e^{-\frac{1}{2} \frac{(\log(l_D^1) - \log(2\mu))^2}{\beta^2}}}{l_D^1 \pi \alpha \beta l_B}. \end{aligned} \quad (7.38)$$

From this joint distribution, the marginal density of L_B can be obtained by integrating Eq.(7.38) with respect to l_D^1 :

$$\begin{aligned} f_{L_B}(l_B) &= \int_0^\infty \frac{1}{2} \frac{e^{-\frac{1}{2} \frac{(\log(l_B) - \log(l_D^1/2))^2}{\alpha^2}} e^{-\frac{1}{2} \frac{(\log(l_D^1) - \log(2\mu))^2}{\beta^2}}}{l_D^1 \pi \alpha \beta l_B} dl_D^1 \\ &= \frac{1}{2} \frac{\sqrt{2} l_B^{\frac{-\alpha^2 - \beta^2 + \log(\mu)}{\alpha^2 + \beta^2}} e^{-\frac{1}{2} \frac{\log(l_B)^2 + \log(\mu)^2}{\alpha^2 + \beta^2}}}{\sqrt{\pi} \sqrt{\alpha^2 + \beta^2}}. \end{aligned} \quad (7.39)$$

Furthermore, with the density f_{L_B} of l_B , it is possible to compute the conditional density of L_D^2 given L_B as

$$f_{L_D^2|L_B}(l_D^2, l_B) = \frac{1}{2} \frac{\sqrt{2} e^{-\frac{1}{2} \frac{(\log(l_D^2) - \log(2\mu))^2}{\beta^2}}}{l_D^2 \sqrt{\pi} \beta}, \quad (7.40)$$

as well as the joint law of L_D^2 and L_B as follow:

$$\begin{aligned} f_{L_D^2, L_B}(l_D^2, l_B) &= f_{L_D^2|L_B}(l_D^2|l_B) f_{L_B}(l_B) \\ &= \frac{1}{2} \frac{e^{-\frac{1}{2} \frac{(\log(l_D^2) - \log(2\mu))^2}{\beta^2}} l_B^{\frac{-\alpha^2 - \beta^2 + \log(\mu)}{\alpha^2 + \beta^2}} e^{-\frac{\log(l_B)^2 - \log(\mu)^2}{2\alpha^2 + 2\beta^2}}}{l_D^2 \pi \beta \sqrt{\alpha^2 + \beta^2}}. \end{aligned} \quad (7.41)$$

Finally, we compute the the distribution of Z using Eq.(7.35), so that we find

$$\begin{aligned} f_Z(z) &= \int_0^\infty |\xi| f_{L_D^2, L_B}(\xi, \xi z) d\xi \\ &= \int_0^\infty \frac{1}{2} \frac{e^{-\frac{1}{2} \frac{(\log(z\xi) - \log(2\mu))^2}{\beta^2}} \xi^{\frac{-\alpha^2 - \beta^2 + \log(\mu)}{\alpha^2 + \beta^2}} e^{-\frac{1}{2} \frac{\log(\xi)^2 + \log(\mu)^2}{\alpha^2 + \beta^2}}}{z \pi \beta \sqrt{\alpha^2 + \beta^2}} d\xi. \end{aligned} \quad (7.42)$$

To find the cell cycle time distribution, we can now simply compute the cumulative distribution of τ and derive it to find the probability distribution function $f_\tau^{\alpha, \beta}$. More specifically, recall that

$$\tau = \frac{1}{\lambda} \log(Z), \quad (7.43)$$

so that the cumulative distribution function of the cell cycle times, denoted by $F_\tau^{\alpha, \beta}$, is

$$F_\tau^{\alpha, \beta}(t) = P(\tau \leq t) = P\left(\frac{1}{\lambda} \log(Z) \leq t\right) = F_Z(\exp(\lambda t)) \quad (7.44)$$

where F_Z denotes the cumulative distribution function of the ratio Z with density f_Z .

The cell cycle time distribution is eventually obtained by computing the derivative of $F_\tau^{\alpha, \beta}$ with respect to t :

$$f_\tau^{\alpha, \beta}(t) = \frac{d}{dt} F_\tau^{\alpha, \beta}(t) = \lambda \exp(\lambda t) f_Z(\exp(\lambda t)). \quad (7.45)$$

Next, we wish to investigate the accuracy of the analytical solution for the cell cycle time distribution, i.e., the formula in Eq.(7.45). To this end, using our agent-based model, we simulate the growth and division of cells under the hypothesis of log-normality for the lengths as in the heuristic. In Figure 7.6.A, we compare the analytical distribution $f_\tau^{\alpha, \beta}$

with the distribution produced by the agent-based model. From the results, we see that the analytical densities fits the numerical densities perfectly. We also notice that the variational parameters α and β significantly change the behaviour of the cell cycle time distributions as opposed to the model devised by Schmidt et al. [121]. This preliminary result strongly suggests that variability in both the birth and division lengths has an important role in bacterial growth. Therefore, we pursue our investigation in the next sections by linking the growth rate and the average cell cycle time and we provide a new model accepting a possible cell size regulator, to see whether a regulator can modify the population growth or not.

7.3.2 Relationship between the Growth Rate at Steady State and the Average Cell Cycle Time

We mentioned earlier that the growth rate of the population at steady state has to satisfy Eq.(7.22). However, the evaluation of the integral can be challenging to solve analytically for some values of the parameters α and β . Therefore, we employ an approximation of Eq.(7.22) which has been presented in [108].

By approximating the exponential of Eq. (7.22) with its Taylor series and centring the cell cycle time to zero (i.e., by subtracting the mean), one can approximate $f_{\tau}^{\alpha,\beta}(t)$, Eq.(7.22) by

$$\int_0^{\infty} e^{-K(t-\bar{t})} f_{\tau}^{\alpha,\beta}(t) dt = 1 + \sum_{i=1}^{\infty} \frac{m_i(-K)^i}{i!} \quad (7.46)$$

where m_i is i^{th} central moment of $f_{\tau}^{\alpha,\beta}(t)$ and $\bar{t}$ is the average cell cycle time.

Thus, expanding the series to its second order in K and using the integration, we find that Eq. (7.46) reduces to

$$2e^{-K\bar{t}} \left(1 - m_1 K + \frac{m_2 K^2}{2} \right) \approx 1. \quad (7.47)$$

Since $m_1 = 0$ (as we centred it by subtracting the average cell cycle time $\bar{t}$) and $m_2 = \sigma^2$, the variance of $f_{\tau}^{\alpha,\beta}(t)$, we obtain

$$1 + \frac{\sigma^2 K^2}{2} \approx \frac{e^{K\bar{t}}}{2}. \quad (7.48)$$

Assuming $\frac{\sigma^2 K^2}{2}$ to be small enough, we take the logarithm on both sides and use its Taylor expansion to conclude that

$$\bar{t} \approx \frac{\log(2)}{K} + \frac{K\sigma^2}{2}. \quad (7.49)$$

Since the variance and the mean of the cell cycle times distribution $f_{\tau}^{\alpha, \beta}$ can be found analytically, we plot in Figure 7.6.B the growth rate of the population K for different values of α and β . It is evident that the growth rate K increases as variability around the birth and division lengths increases. Although the shape of the surface calculated for K appears to be symmetric, one easily notices that for small values of α , increasing β seems to increase the growth rate. This may be explained upon recalling the definition of the growth rate as defined in the PBE. Therein, the growth rate was shown to depend on the integration of both the density of cells and the division rate.

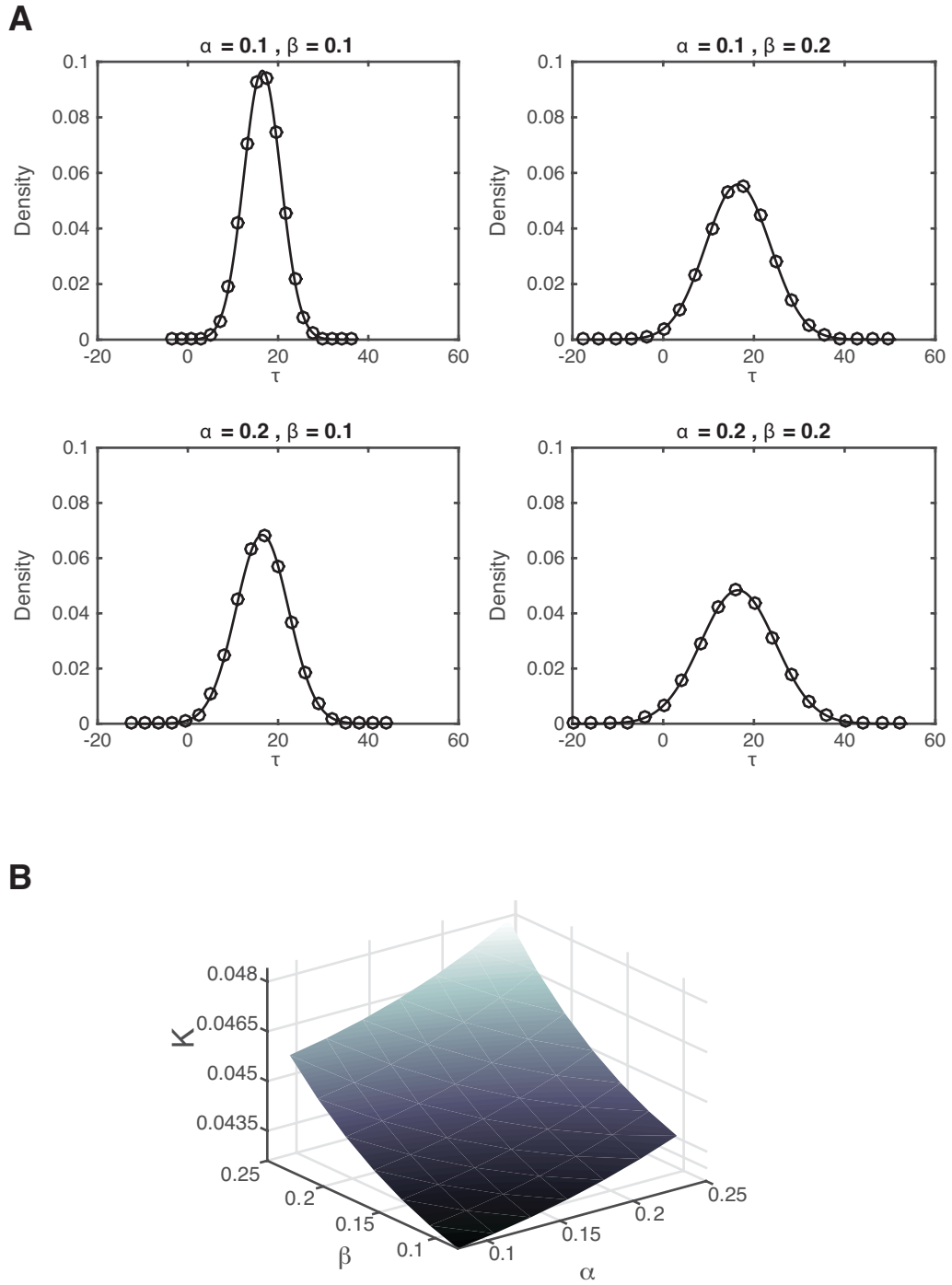


Figure 7.6: The analytical distributions for the cell cycle time, i.e., the times until division, allows to accurately compute distributions for different values of variability in α and β . In addition, it is possible, using the simplification by Painter & Marr [108], to estimate the growth rate K of the population at steady state. Different analytical and numerical cell cycle time distributions are presented in panel A. for different values α and β . The dots represent the distributions coming from the numerical simulations and the line are the distributions from the analytical solution. In panel B. the growth rate K is approximated for values α and β ranging from 0.1 to 0.25 respectively.

7.4 Cell Size Regulation

For a long time, cell size regulation has been an unanswered question in biology. Why cells can trigger their cycle and divide at the "right" time is still unclear and debated. Hypotheses on possible controllers of the birth and division length have been expressed in the past. In general, it is thought that cells either possess a "timer", a "sizer", or, as advanced more recently, an "adder" [133, 56] that controls their cell cycle. The three hypotheses can be summarised as follow: in the first case, the timer lets the cell wait a certain amount of time before dividing. In the second case, the sizer waits until the cell reaches a certain size to trigger the cell cycle. Finally, for the third scenario, the adder senses the birth length of the cell and adds to it a fixed amount of "size" to determine the division size of the cell. Until now we have assumed that the birth and division lengths were independent.

In this section, we model a possible control between the birth and division length by combining all three ideas. Specifically, we use both the timer and the sizer combined with the adder by considering a linear dependence between the birth and division lengths. We model the division length of a cell by restraining the division length to a linear function of the birth length. In other words, if L_B is the birth length of a cell, the division length L_D will follow a log-normal distribution centred in $2((1-s)\mu + sL_B)$, i.e.,

$$L_D \sim \ln\mathcal{N}(2((1-s)\mu + sL_B), \beta^2), \quad (7.50)$$

with $s \in [0, 1]$.

The role of the parameter s in our model is crucial as it allows to switch between the sizer model, i.e., no dependence between the birth and division length when $s = 0$ and the adder model, when $s = 1/2$, when the division length depends on the birth length plus an added length and finally the timer model where $s = 1$ and the division length is completely dependent on the birth length. To illustrate this relationships between the birth and division lengths when the parameter s varies, we sketch the difference, denoted Δ , between the division and birth lengths against the birth length in Figure 7.7. The sizer model is not

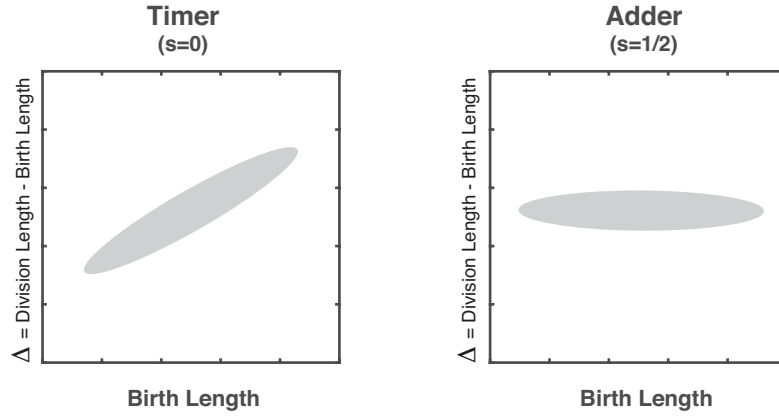


Figure 7.7: The parameter s in our model, allows to switch between the three models: timer, sizer and adder. By setting $s = 0$, we model a sizer. When $s = 1/2$, the adder is chosen and finally when $s = 1$, the timer model is considered.

shown in the figure as it does not depend on the birth length. It must be noted that as the parameter s lies in $[0, 1]$, we can set it to perform in between the three models, to see how the growth rate changes. Such linear dependence between the birth and division lengths has recently been explored by Tanouchi et al. [134] to reveal the nature of transient oscillations in cell size.

Here, we are interested in varying the parameter s and see how the average cell cycle time, its variance, and the associated growth rate change accordingly. As we have seen in the section about the time-dependent growth rate $K(t)$, the growth rate tends to converge to a fixed growth K , where the speed and the value of K itself depends on the parameters α and β . Here we employ a similar approach as in the last section to derive an analytical formulation of the cell cycle times distribution, based on the new parameter s . Furthermore, we give an extended point of view by computing the growth rates of the population, according to the values of s .

In addition, and to confirm our model, we devise an algorithm that extracts morphological parameters, such as α , β and s , from experimental data.

7.4.1 Linear Dependence between Birth and Division Lengths

To account of the new parametric dependence parameter, labelled s in the analytical distribution of the cell cycle times, we derive again the distribution of the cell cycle times τ ,

denoted $f_{\tau}^{\alpha,\beta,s}$. To this end, we modify the third point of the heuristic devised previously, and we set the division length to be

$$L_D^{2,s} \sim \text{ln}\mathcal{N}(\log(2(1-s)\mu + sL_B), \beta^2), \quad s \in [0, 1]. \quad (7.51)$$

In this setting, s "continuously" links the birth and division lengths.

To find the distribution of the cell cycle times τ , we derive again the distribution for the new ratio Z^s , i.e., $Z^s := L_D^{2,s}/L_B$ and find

$$\begin{aligned} f_{Z^s}(z) &= \int_0^\infty |\xi| f_{L_D^{2,s}, L_B}(\xi, \xi z) d\xi \\ &= \int_{-\infty}^\infty \frac{1}{2} \frac{e^{-\frac{1}{2} \frac{(\log(z\xi) - \log(2(1-s)\mu + 2sL_B))^2}{\beta^2}} \xi^{\frac{-\alpha^2 - \beta^2 + \log(\mu)}{\alpha^2 + \beta^2}} e^{-\frac{1}{2} \frac{\log(\xi)^2 + \log(\mu)^2}{\alpha^2 + \beta^2}}}{z\pi\beta\sqrt{\alpha^2 + \beta^2}} d\xi. \end{aligned} \quad (7.52)$$

In this new formulation for the distribution of the ratio, we clearly see that the parameter s appears inside the integrand in a non-trivial way.

Once the new density f_Z of the ratio Z is defined, we derive the cumulative distribution function of the cell cycle times, $F_{\tau}^{\alpha,\beta,s}$ as

$$F_{\tau}^{\alpha,\beta,s}(t) = P(\tau \leq t) = P\left(\frac{1}{\lambda} \log(Z^s) \leq t\right) = F_{Z^s}(\exp(\lambda t)). \quad (7.53)$$

Finally, the probability density function for the cell cycle times, depending on the parameter s , is found as

$$f_{\tau}^{\alpha,\beta,s}(t) = \frac{d}{dt} F_{\tau}^{\alpha,\beta,s}(t) = \lambda \exp(\lambda t) f_{Z^s}(\exp(\lambda t)), \quad (7.54)$$

accordingly. To test the accuracy of the analytical distribution $f_{\tau}^{\alpha,\beta,s}$, we incorporate the new parameter s into the agent-based model. Simulations based on the modified agent-based model (shown in Figure 7.8) indicate that the analytical density fits the numerical approximation well for different values of the parameter s and standard deviations α and β .

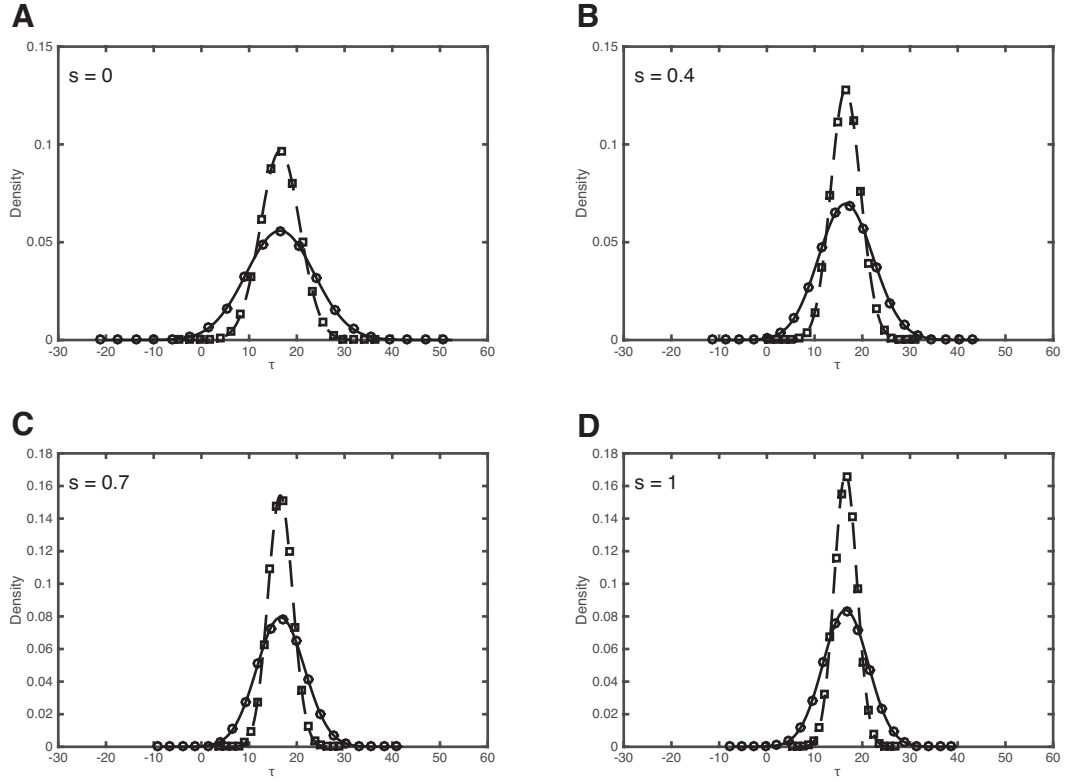


Figure 7.8: Comparison between the analytical and numerical distributions of the cell cycle times. The thick and dashed lines represent the analytical densities for $(\alpha = 0.1, \beta = 0.2)$ and $(\alpha = 0.1, \beta = 0.1)$ respectively. The circles and squares represent the numerical densities for respectively $(\alpha = 0.1, \beta = 0.2)$ and $(\alpha = 0.1, \beta = 0.1)$. Each panel shows the densities for different values of s . A. No dependence between the birth and division lengths, i.e., $s = 0$. B. Moderate dependence between the birth and division lengths, i.e., $s = 0.4$. C. Strong dependence between the birth and division lengths, i.e., $s = 0.7$. D. Complete dependence between the birth and division lengths when $s = 1$.

To investigate the dependency between the birth and division lengths with respect to the new parameter s , we plot the growth rate of the population as defined in Eq. (7.49) as well as the average cell cycle time $\bar{t}$ and σ^2 in Figure 7.9.A. The average cell cycle time $\bar{t}$ seems to depend only on the variation of the parameter s as the quadratic shape has always a maximum in $s = 0.5$, independently of the values of α and β . However, the variance does not show any global extremums, therefore not indicating whether a specific value s would be more advantageous for the cells than others.

Although one might suppose that for a faster population growth a lower cell size variability could be beneficial, we find that increasing the parameter s decreases the variability of the cell cycle time and also the growth rate of the population. A possible answer to this phenomenon is that the cells need to spend a certain amount of time in their cycle to grow and divide without slowing down the population growth. If the cell cycle time has too much variability, i.e., the variance here, then the cells will spend more time growing and dividing. In contrary, if cell cycle time variability is too short, then the cells will most likely not grow enough and thus divide without attaining their correct size. Consequently, there exists a trade-off for the dependency between the birth and division lengths, which is what we see in Figure 7.8. Figure 7.9.B displays the growth rate of the population at steady state, plotted for different values of the parameter s . As one can see, the growth rate is larger when there is no dependence ($s = 0$) between the birth and division lengths. Although the maximum growth rate seems to decrease when s increases, the fluctuations for the maximum growth rate seem to vary only marginally when s is larger than 0.4.

To verify the hypothesis of a trade-off in the birth and division length dependence, in what follows we devise an algorithm capable of extracting morphological statistics on real experimental data.

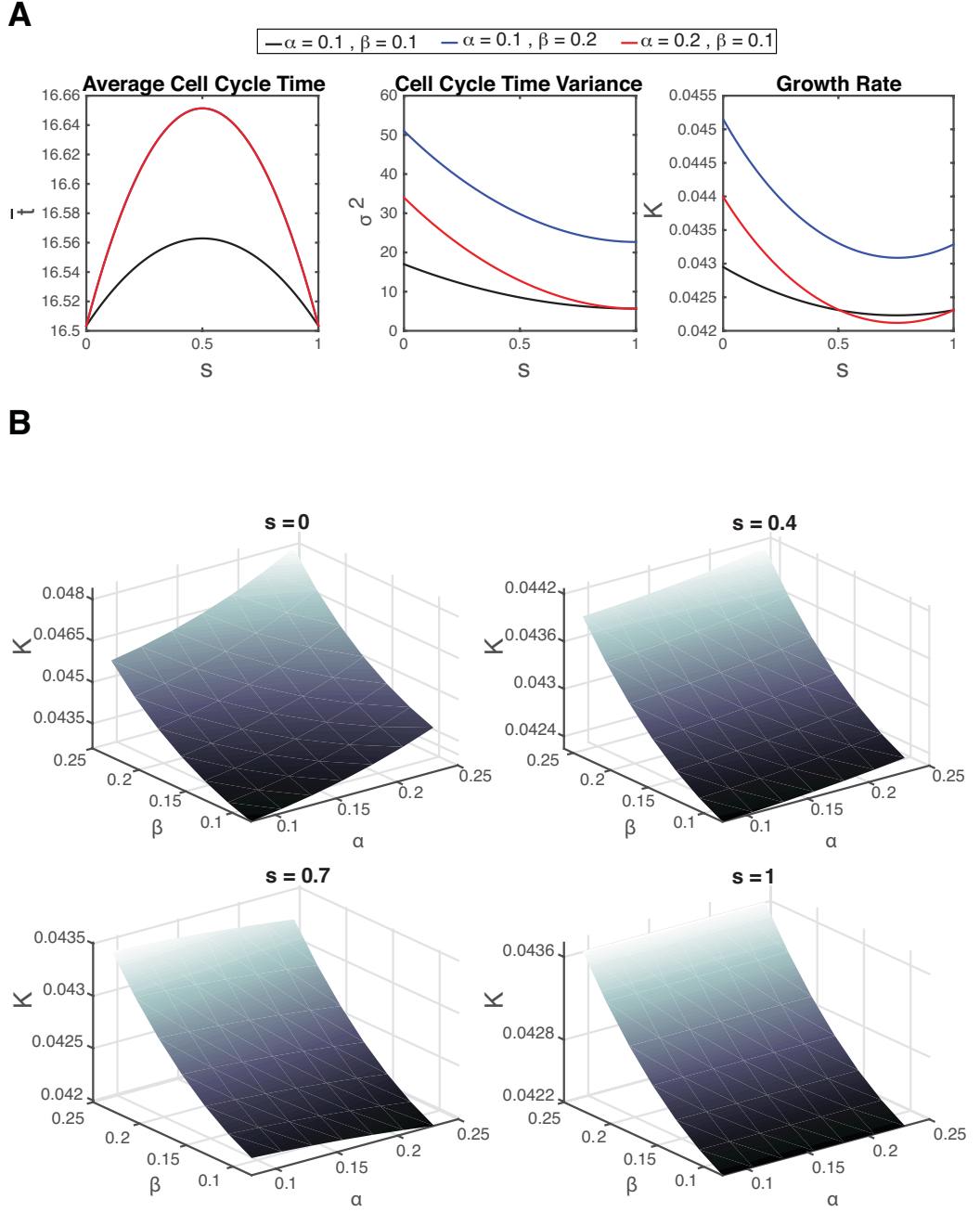


Figure 7.9: The analytical distribution of the cell cycle times, according to the parameter s allows to conveniently compute the average cell cycle time, the standard deviation, and the growth rate of the population. In panel A. the average cell cycle time, the variance and the growth rate for $(\alpha, \beta) = \{(0.1, 0.1), (0.1, 0.2), (0.2, 0.1)\}$ are illustrated. B. Growth rates of the population for different values s , i.e., with no dependence between the birth and division lengths ($s = 0$) and complete dependence ($s = 1$). As the parameter s increases, the shape of the surface representing the growth rate K evolves.

7.4.2 Data Analysis of Biological Experiments on *E. coli*

To support and verify our conjecture on the existence a possible trade-off between the birth and division lengths, we analyse morphological data provided by the "mother machine" [144]. The mother machine consists of a microfluidic chemostat having multiple channels allowing the bacteria to grow and divide continuously, while maintaining their (fixed) number (in a refreshed medium). The strains analysed in the mother machine are *E. coli* MG1655, MG1655 *lexA3*, B/r SJ108 and B/r SJ119. The mother machine allows us to capture morphological parameters such as the length, the width, the area, the intensity, and the centre of mass of each bacterium. The data provided by the device are of the form of time series. For each time point, an additional Boolean parameter is included with value 0 when no division occurs and 1 when there is a division. The Boolean parameter allows to separate the time series into blocks of single cell growths and consider cell cycles independently (see Table 7.2 for an extract of such times series).

Time	Division	Length	Width	Area	Intensity	CMx	CMy
0	0	42.0483	15.1492	637	116.761	52.4019	472.477
1	0	41.0898	13.6043	559	122.655	51.5456	471.256
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
31	0	98.2098	18.6132	1828	111.803	50.2511	429.8
32	1	51.2416	10.109	518	120.409	51.1892	459.479
33	0	53.2435	11.6822	622	117.822	51.7797	454.637

Table 7.2: Morphological parameters available via the mother machine device. The measurements are in pixels. The time series are decomposed into growth and division (indicated by a 1 in the division column).

Although the mother machine provides statistics on the morphological traits of the strains analysed, the statistics are biased due to the discrete nature of the time series. In the mother machine experiment, morphological features are measured every minute. The algorithm employed in the mother machine evaluates, using image processing techniques, the length of the cell and decides whether or not the bacterium has divided by comparing the length of the bacterium at time t and the length at time $t - 1$. Since the growth process is continuous, the mother machine might miss the exact division time of the cell and might set it to the next minute instead (as illustrated in Figure 7.10.A). This error induces a bias,

which can be quantified by calculating the ratio of half the mother cell length by the daughter length. In a non-noisy case where there is no variability, the ratio is one. However, in a variable context with variability around the birth and division lengths, the ratio of half the mother length by the daughter length should be centred around one. As we can see in Figure 7.10.B, the ratio is slightly shifted to the left, therefore indicating that the time series gathered by the mother machine lacks the resolution required to capture the division process.

However, it is possible to extract the real nature of the time series gathered in the mother machine by simulating the elongation of bacteria. Supposing an exponential elongation of the bacteria, one can reproduce the time series of the mother machine by simulating the growth according to the parameters we want to examine, i.e., the variability around the birth and division length and the dependence parameter s . To this end, we devise an algorithm that simulates the growth of the bacteria according to five parameters:

- μ , the average birth length of the cell
- λ , the elongation rate of the cell
- α , the variability in the birth length
- β , the variability in the division length
- s , the birth and division dependence parameter

all describing the morphology of the cells.

In a first time, the algorithm sets the simulated birth length $L_{\text{sim}}(0)$ to the initial birth length value, l_0 , available from the mother machine dataset. In the meantime, the algorithm already sets the division length, $L_{\text{sim}}^{\text{max}}$ as a random variable following a log-normal distribution with mean $2((1-s)\mu + sl_0)$ and standard deviation β . Then, the current simulated length, $L_{\text{sim}}(t)$, is calculated by solving the ordinary differential equation of the standard exponential growth of Eq. (7.32). Here, we employ a forward Euler scheme with a discretised time step $\Delta t = 1 \times 10^{-5}$ to numerically solve the differential equation. At each time step, the algorithm verifies if the current length obtained with the ordinary differential equation is

Algorithm : Simulated Exponential Elongation $L(t) \equiv L(t, \mu, \alpha, \beta, \lambda, s)$

Input : $\mu, \alpha, \beta, \lambda, s$

1. Set $t = 0$ and $L_{\text{sim}}(0) = l_0, N = 1$.
2. Define $L_{\text{sim}}^{\text{max}} \sim \ln\mathcal{N}(2((1-s)\mu + sl_0), \beta^2)$ as the division length.
3. Elongate the cell according to

$$L_{\text{sim}}(t + \Delta t) = (1 + \lambda)L_{\text{sim}}(t)\Delta t$$
 until $L_{\text{sim}}^{\text{max}}$ is reached. Record the length L for each time step, using the values of L_{sim} .
4. If $L_{\text{sim}}(t) \geq L_{\text{sim}}^{\text{max}}$ "divide" the cell: set $L_{\text{sim}}(0) \sim \ln\mathcal{N}(\frac{L_{\text{sim}}^{\text{max}}}{2}, \alpha^2), N = N + 1$
5. Repeat 2 – 4 until $N = N_d$

Output : $L(t)$

Algorithm 7.3: Simulation of the exponential elongation

greater or equal the division length. If $L_{\text{sim}}(t)$ is less than $L_{\text{sim}}^{\text{max}}$, the algorithm continues elongating the cell. Otherwise, if the current length has reached or overtaken the division length, then the algorithm creates a new daughter by setting the current length to follow a log-normal distribution with mean $L_{\text{sim}}^{\text{max}}/2$ and standard deviation α . The elongation and division process is repeated N_d times, where N_d is the actual number of divisions in the mother machine data. The complete process is summarised in Algorithm 7.3.

Once the exponential elongations are computed, we compare the simulated trajectories with the real trajectories coming from the mother machine. The first step is to extract the birth and division lengths from the original time series. Then, a log-normal distribution is fitted for both measurements. Hereby, it is possible to acquire the original (fitted) α and β values. In addition, the elongation rate λ is fitted to the original trajectory. Moreover, the dependence parameter s , between the birth and division lengths, is also measured. The extraction of the morphological features then consists of simulating the elongation of the bacteria and fitting the simulated time series (using Algorithm 7.3) to the original time series. The input parameters for the simulation of the elongation are randomly chosen in a

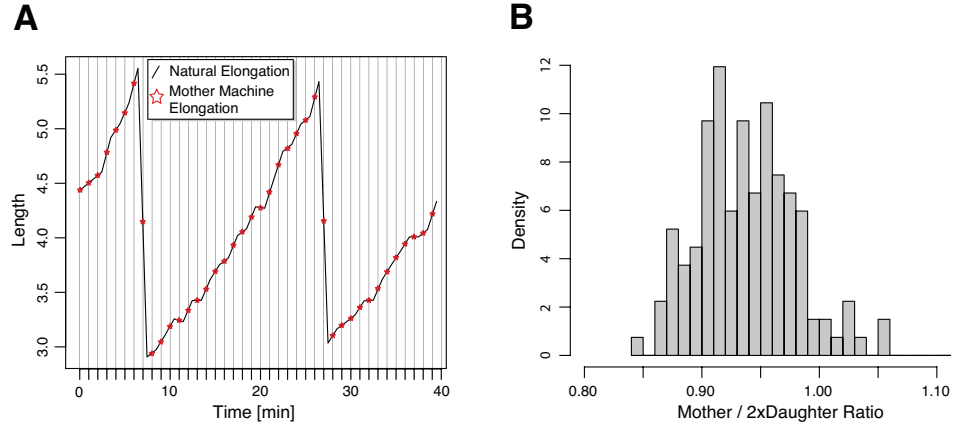


Figure 7.10: Time series recorded by the mother machine microfluidic device and ratio between half of the mother length and the daughter length. (A) The time series recorded by the mother machine (red stars) sometimes misses the division points and birth length (black line). (B) Due to the missing points, the ratio of half the mother length by the daughter length is shifted (here to the left), indicating that the time series do not account for the real division time points.

range of known parameters. The distributions for the parameters described in Eq. (7.55).

$$\begin{aligned} \mu &\sim \mathcal{U}(1, 10) \quad , \quad \lambda \sim \mathcal{U}(0.01, 0.1) \quad , \quad \alpha \sim \mathcal{U}(0, 1) \\ \beta &\sim \mathcal{U}(0, 1) \quad , \quad s \sim \mathcal{U}(0, 1). \end{aligned} \quad (7.55)$$

For small time steps Δt , the Algorithm 7.3 provides accurate values of the simulated cell length elongation, which is needed to ensure that the right division times in the time series are obtained. Once the lengths have been simulated we discretise the entire elongation processes (for multiple N_d) by rounding the time to minutes. In other words, we define

$$T := \{1, 2, \dots, t_{\text{end}}\},$$

as the time interval of the simulated time series, where t_{end} is the final time of the series (for example in Table 7.2 the value of t_{end} would be 33).

Then, for each time point we identify the division points, labelled by a boolean true (i.e., 1) and we calculate the birth and division lengths. As for the original time series, we obtain the α and β parameters by fitting a log-normal distribution to the birth and division lengths

previously extracted. For each simulation we define the quadratic error function

$$e(\mu, \lambda, \alpha, \beta, s) := \sum_{t \in T} |S_{\text{param}}(t) - L_{\text{param}}(t)|^2, \quad (7.56)$$

where S_{param} are the fitted morphological parameters of the original time series and L_{param} are the fitted parameters of the simulated exponential elongation. As in Chapter 3, we use the *Limited Memory Broyden-Fletcher-Goldfarb-Shanno* (L-BFGS-B) algorithm to minimise the error function and find the least squares fit of the parameters for the simulation. However, due to the stochasticity of the elongation process and for statistical purpose, we create the error vector

$$E := (e_i)_{i=1, \dots, n_{\text{sample}}}, \quad (7.57)$$

where e_i is the error function defined in Eq. (7.56) based on sample i where the number n_{sample} of repetitions is large enough to compute the average and standard deviation of the fitted parameters confidently.

The mother machine was employed to extract morphological features of four *E. coli* strains: B/r SJ108, B/r SJ119, MG1655 and a mutant of it, MG1655lexA3. The distributions of the extracted features are shown in Figure 7.11. Unlike the MG1655 strain and its mutant, the B/r strains have fewer measurements. This is seen on all panels of Figure 7.11 by the small histograms for the B/r SJ108 and SJ119 strains.

We see in Figure 7.11.A that the distributions of birth lengths follow similar trends as described in the original paper of Wang et al. [144] (and as provided in the appendix). This first result indicates that the simulated elongation process coupled to the minimisation algorithm performed well and in a natural way. However, our result for the elongation rate λ significantly differs from the original analysis as the distribution of rates are here shifted to the left, and thus are smaller on average. We can explain this phenomenon by again considering the machinery behind the image acquisition process of the mother machine, where the size of the bacteria is measured every minute, therefore possibly missing the real division length of the cells. Thus, for the same amount of time, the cells acquired in the

mother machine data spend less time to grow and divide, which causes the elongation rate to be faster than the ones presented in Wang et al. [144]. In panels C and D, we see that the standard deviation for the birth length distribution is centred around 10% as firstly noticed by Schaechter et al. [120]. This phenomenon has also recently been confirmed again by Jun & Taheri-Araghi [56]. In addition, the standard deviation for the division length is centred around 20%. Our results and fits strongly indicate that our model and consequently the numerical simulations run on it, are good representation of the biological phenomenon happening in the mother machine. More than finding similar results, in terms of variation in lengths, our model described the possibility of having a cell cycle regulator. We comment on the dependences between the birth and division lengths, i.e., the parameter s in our model. We see here that the distribution of s is centred around 0.5 and varies between 0.25 and 0.75. This tells us that, as shown in the theoretical model, the cell senses its size and adapts it in order to help the population grow at the optimum rate. We recall that for a value of $s = 0.5$ we obtain the maximum average cell cycle time in the theoretical model. As we see variations in the distributions due to the parameter s and also that the optimal parameter of 0.5 is found theoretically and experimentally is a good indicator of a possible cell size regulator in the cell cycle of *E. coli* as recently discovered by Taheri-Araghi et al. [133] and Jun & Taheri-Araghi [56]. Indeed, for a value of $s = 0.5$, we illustrated (in Figure 7.7) that the model that best describes the growth is the adder model. It must be noted that this result of $s = 0.5$ has to be considered carefully, as suggested by Tanouchi et al. [134] who showed that the adder model might not be applicable in a general fashion to all organisms. Tanouchi et al. suggest that the adder model is in fact dependent on the organism being under study and the growth conditions.

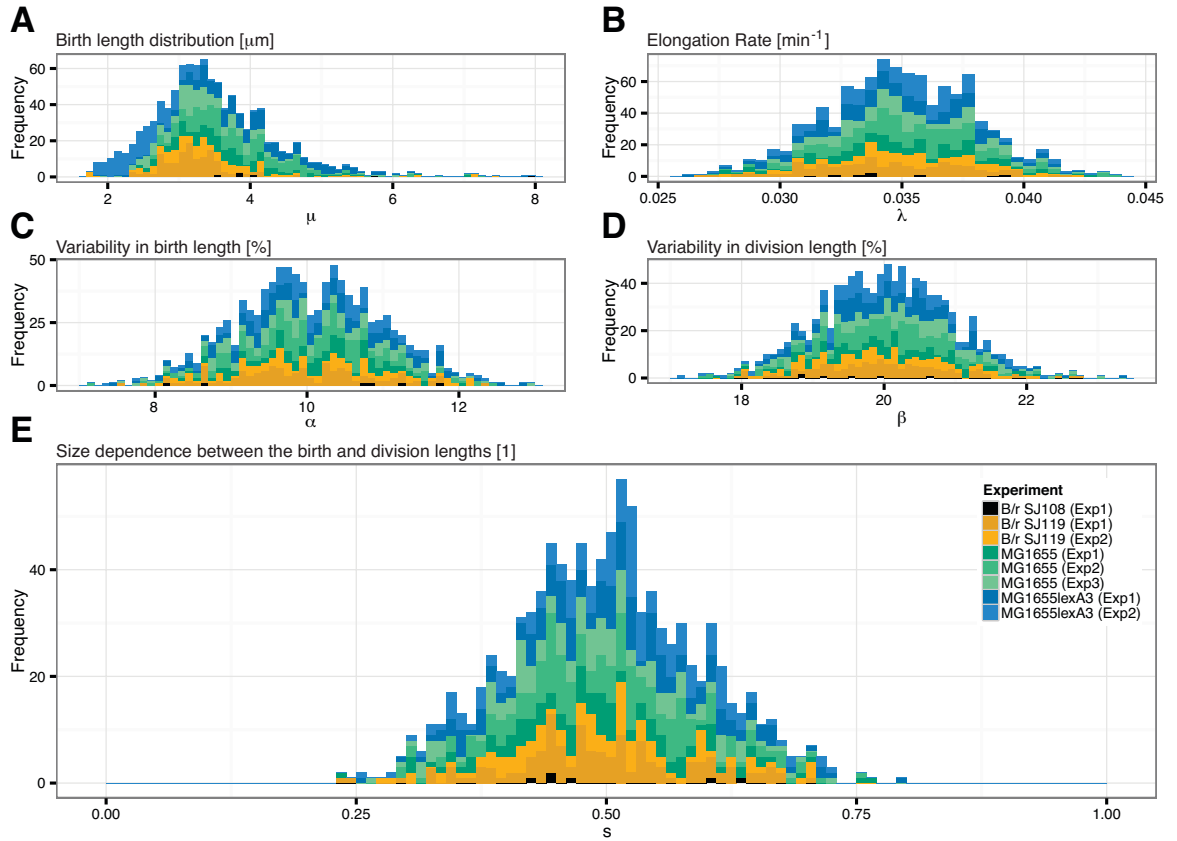


Figure 7.11: As discussed in the main text, the mother machine lacks the possibility of capturing the real morphological phenotypes of the cells as the images are sampled every minute. We solve this problem by simulating exponential elongations of the cells and fitted the simulations to the original data. The simulations allow to extract better estimations of the morphological parameters. A. the average birth length μ B. the elongation rate C. variability in the birth length and D. variability in the division length. Finally, the algorithm also allows us to extract the dependence parameter s between the birth and division lengths in panel D. The simulations seem to indicate that the linear relationship between the birth and division lengths is centred around $s = 0.5$. This results was shown previously as the theoretical models also shows that the maximum average cell cycle time $\bar{t}$ has a maximum in $s = 0.5$.

7.5 Discussion

In a recent study by Schmidt et al. [121], no correlation between variability around the birth length and the fitness (or growth rate) of the population was found. However, in their analysis, the authors did not account for the possibility of having variability in the division length. They also did not include any type of cell size control, as suggested in [133, 56]. Because of these shortcomings, we decided to generalise the analysis from Schmidt et al. with the aim to include additional sources of morphological variability, specifically by considering both the variability in the birth and division lengths and a possible cell cycle regulator.

We modelled variability for the birth and division lengths by deriving a special form of population balance equation (PBE), adapted to the analysis of the cell cycle of *E. coli*. The population balance equation offers the advantage of considering morphological parameters with their probability distributions. In fact, in our analysis, it was also possible to adapt the population balance equation to real experimental data.

Once adjusted to the binary fission of *E. coli*, we reproduced the PBE by simulating the growth of bacteria with an agent-based model. Based on the simulations, we were able to draw two main conclusions. Firstly, the growth rate of the population is time-dependent and relates to the cell distribution as well as the division rate of the cells. Depending on the initial cell distribution in the colony, the growth rate shows damped oscillations until the cell distribution reaches its steady state. Moreover, the convergence of the growth rate to a fixed value did not depend on the cell distribution. The speed of convergence however did depend on the cell distribution. Secondly, the value of the growth rate at steady state appeared to crucially depend on the variances of the birth and division length distributions. This observation is contrary to the results given in Schmidt et al. [121].

To examine the growth rate of the population at steady state, we derived an analytical formulation of the cell cycle time distribution based on a heuristic for the growth and division processes of the bacteria and computed the growth rate of the population according to Painter & Marr [108]. Similarly to the time-dependent growth rate, the growth rate at

steady state crucially depends on the variances of the birth and division length distributions. In this scope, we decided to implement a cell cycle regulator to see whether the growth rate at steady state would increase or decrease if the cell cycle regulator was on or off.

We have also mentioned that cell cycle regulators have been categorised in three: the "timer", the "sizer" and the "adder". In our study, we employed a mixture of these three regulators by arguing that the division length of the bacterium should be linearly dependent on the birth length with a factor s between 0 and 1. In this setting, we revisited the analytical formulation of the cell cycle times distribution and recomputed an updated version of the growth rate of the population at steady state. With the linear dependence between the birth and division length, we showed that the average cell cycle time reached a maximum at $s = 0.5$. We explained that this could be due to the fact that a cell should not spend too much time growing nor grow too fast in order to avoid missing essential molecular properties. We also showed that the growth rate of the population at steady state is influenced by the variances of the birth and division length distributions and the value of the dependence parameter s .

In order to verify our mathematical models, we extracted morphological data from the "mother machine" [144] and compared the *in vivo* data to the *in silico* data. We found a good agreement between our simulated data and the data from the mother machine. In addition, we showed that there is approximately 10% variability in the birth length and 20% variability in the division length. Moreover, and as predicted by the analytical model, the bacteria in the mother machine showed a strong dependence parameter centred around $s = 0.5$. We showed that when $s = 0.5$, the model that best describes the growth of the cells is the adder model devised by Taheri-Araghi et al. [133] and Jun & Taheri-Araghi [56].

In conclusion, our analysis shed light on the dependence between variability in the birth and division lengths by introducing a theoretical model combined with numerical simulations provided by an agent-based model. We showed that a population of bacteria requires variability in the birth and division lengths to adapt its growth. We also showed that it is possible to capture experimental data using the ABM. By the nature of the agent-based

model, we think that the ABM could easily be used to simulate and analysis other organisms, such as *S. cerevisia* or *S. pombe*. Finally, our model on the growth of the population linked to a cell cycle regulator showed that the hypothesis of an adder is plausible. Still in the idea of expanding the ABM to other organisms, we think that it would be interesting to approach the aforementioned organisms and analyse their cell cycle to see if an adder is also present.

Chapter 8

Conclusion and Outlook

The thesis was designed to answer biological questions related to phenotypic variability in three different unicellular organisms: *S. pombe*, *S. cerevisiae* and *E. coli*. We wanted to assess the source of phenotypic variability and understand the consequences of such variability on the organisms. We looked at the robustness and the fitness of the organisms themselves and when possible, the entire population. For each organism we developed specific mathematical models, combined with numerical simulations and/or analytical solutions, to study the effect of phenotypic variability.

Here, we conclude this thesis with a summary of each chapter and discuss further extensions of our work and the outlook.

8.1 Stochastic Lag-time and Fractional Growth of the Population

In this chapter we presented a model for the dynamic transition from quiescence to proliferation in *S. pombe*. We showed that the "lag", i.e., the amount of time the cells wait before starting growing again, can be modelled either using a probabilistic model which considers the lag as a random variable or with a fractional growth of the population, where only some cells survived quiescence.

Although both models correctly reproduced the growth curves obtained from the biological experiments, the Akaike Information Criterion of both models was not strong

enough to discern the best model. Therefore, we suggest a set of experiments at the single-cell level to quantify viability at the end of the quiescence phase. This is to observe the fraction of revived cells. However, it must be noted that both models agreed on a set of mutants being outliers compared to the wild types. The Set1C complex proved to diverge from the wild-type trend, therefore suggesting more experiments on this complex.

8.2 Inducer, Maintainer and Capacitor Mutants

The second project aimed at reproducing the work of Levy & Siegal [76] in *S. cerevisiae* and enhance their results by considering a wider methodology. Here, instead of using the Lowess regression as in their work, we employed an $L1$ -regression, more robust to the outliers. We analysed the mutants both increasing and decreasing variability, seen as the coefficient of variation of clustered morphological phenotypes. To this end, we devised a score quantifying the global effect of a gene deletion on the set of phenotypes: the "Global Phenotypic Variability" score. After validating the score with existing measures/scores available in the literature, we categorised the mutants according to their effect on phenotypic variability.

Our methodology allowed us to categorise the mutants into four groups regarding their effect on phenotypic variability. The first group concerned the mutants that increase variability. This group was highly correlated with the phenotypic capacitors of Levy & Siegal. The second group contained the mutants decreasing phenotypic variability. We named these genes: "inducers". The third group showed to have mutants acting as both capacitors and inducers. Interestingly, we compared this third category of mutants with a collection of 37 wild strains. In this setting, the wild strains seemed to belong to the maintainer genes category, possibly suggesting that although phenotypic variability might be shuffled, the absolute phenotypic value is conserved. Finally, our analysis and selection process indicated that robustness is inversely related to fitness. The inducers, decreasing variability and therefore being more robust, showed to have larger growth rates than the capacitors, increasing variability.

We note that it would be interesting to generalise our results to other phenotypes, such as gene expression, for example by using RNA-seq at the single-cell level.

8.3 Phenotypic Variability and Cell Size Regulation in *S. cerevisiae*

Chapter 5 continued the analysis of phenotypic variability in *S. cerevisiae* by considering only specific phenotypes related to cell, nucleus and bud size. We focused this chapter on the understanding of mutants affecting the cell size of budding yeast during its cell cycle. To this end, we devised two approaches, based on the dimensionality of the morphological phenotypes. Firstly, we analysed the mean phenotypic value and its coefficient of variation independently. Here, we employed the same idea as in Chapter 5 by considering the residuals of the coefficients of variation and taking the bottom and top ranked 50 mutants. Secondly, we created two algorithms working on the geometry of the Gene Ontology annotations of the mutants, compared to the wild types for a given phenotype. The first algorithm, separated the mutants and the wild types using ellipses based on the standard deviation of the phenotypes. The second algorithm employed a partitioning around medoids to label the mutants and the wild types and thereafter cluster them to get the ratio of wrongly clustered and labelled mutants.

Both methodologies indicated that the mutants increasing variability during the cell cycle of budding yeast were in general related to DNA damage, spindle and chromosome segregation, chromatin remodelling and vacuolar pH. On the contrary, mutants decreasing variability were related to amino acids, vesicle function, tRNA and rRNA processing.

Our collaborators are currently setting up experiments to analyse some of the extracted mutants related to vacuolar pH and DNA damage.

8.4 Cell Size and Nuclear Size Regulation with Bayesian Networks in *S. cerevisiae*

The fourth project gave a first answer to the question of the control of the nucleus by the cell. In this project we analysed the topology of directed acyclic graphs constructed with the morphological phenotypes of *S. cerevisiae*. We constructed the graphs using a Bayesian network approach. This approach allowed to create a first set of networks for three different cell cycle stages, namely: the stage "A" (G1 phase), the stage "A1B" (S/G2 phases) and the stage "C" (M phase). We then looked at the equivalence class of the first network to identify strong nodes (phenotypes). Finally, we bootstrapped the networks to get the strength and the direction of the arcs between the different phenotypes.

The networks created using the wild-type data showed a strong directed arc from the cell to the nucleus, indicating that cell size dictates nuclear size. However, the arc was strong only for the stages A and A1B. In stage C, the arc disappeared and although it was still possible to find a path from the cell to the nucleus, the path included multiple arcs in between the cell and the nucleus phenotypes. In addition, we showed that the mutants could be separated into five categories. The first three categories represent the mutants possessing the same connectivity as the wild types, i.e., the mutants showing an arc from the cell to the nucleus. The fourth category concerns the mutants with an arc from the mother cell size to the nucleus size but the direction is not decided. Finally, the last category contains the mutants having no arcs between the mother cell size and the nucleus. For the last category, we noted that the small number of mutants belonging in this category might actually be an artefact due to the poor number of single-cell measurements available for the distributions of morphological phenotypes.

As a further investigation, we think that performing the image acquisition process again to collect more single-cell measurements will enhance the quality of the mutants networks. Moreover, it would be interesting to have a deeper biological look at the mutants showing no arc between the cell and the nucleus.

8.5 Bacterial Growth and Size Regulation in *E. coli*

In this last project, we proposed a theoretical approach to explain cell size variability and the growth rate of bacterial population. We devised a population balance equation (PBE) considering the probability distributions of the birth and division lengths of bacterial cells, thus allowing to vary the variance of the distributions. By its impossibility to solve analytically and sometimes numerically, we decided to work on the analytical properties of the equation instead of trying to solve it numerically. We demonstrated that there exist an analytical formulation for the growth rate of the population. Hence, we showed that the growth rate of the population is a time dependent function of the cell size distribution and the division rate of the cells. To measure the growth rate of the population, we simulated the PBE with an agent-based model. As suggested by the definition of the PBE, the growth rate of the population converged to a stationary value when the cell size distribution converged to its steady state. At this point, we introduced a cell cycle regulator in the theoretical model, to see whether regulating the cell sizes would enhance the fitness of the population. The regulator was based on the idea that it can allow us to switch from the timer, sizer and adder models by varying a single parameter. We validated our theoretical model by comparing the simulations to experimental data extracted from the "mother machine" [144]. Unlike a previous study from Schmidt et al. [121] our results showed that the variance of the birth and division lengths affects the growth rate of the population, both in the growth phase and at steady state. We explained this difference of results by noting that in their models, they did not consider the variance of the division length, which probably biased their results as the growth rate (as derived from the PBE) employs the division rate. We believe that the growth rate, as derived from the PBE, is correct as we found similar results of the growth rate at steady state by approaching and solving the problem with a different approach (heuristic). Another main result of our analysis was the presence of a cell cycle regulator, which also proved to exist in the experimental data. By considering that the division length of a cell is a trade-off between its birth length and the average birth length in the bacterial population, we showed that experiments results coincided with the theoretical model, i.e., that the cells needed to employ the adder model. However, we also recall the recent discovery from Tanouchi et al. [134] who found, based on the transient nature of

oscillations for the cell size in *E. coli*, similar results as ours. Their analysis showed that the adder model might in fact be dependent on the growth conditions and the organisms being studied.

Bibliography

- [1] M. Acar, J. T. Mettetal, and A. van Oudenaarden. Stochastic switching as a survival strategy in fluctuating environments. *Nature Genetics*, 40(4):471–475, 2008.
- [2] K. A. Aho, D. Derryberry, and T. Peterson. Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, 95(3):631–636, 2014.
- [3] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell, Fourth Edition*. Garland Science, 4 edition, 2002.
- [4] B. B. Aldridge, M. Fernandez-Suarez, D. Heller, V. Ambravaneswaran, D. Irimia, M. Toner, and S. M. Fortune. Asymmetry and aging of mycobacterial cells lead to variable growth and antibiotic susceptibility. *Science*, 335, 2012.
- [5] A. Amir. Cell size regulation in bacteria. *Phys. Rev. Lett.*, 112:208102, 2014.
- [6] C. A. Athale and H. Chaudhari. Population length variability and nucleoid numbers in escherichia coli. *Bioinformatics*, 27, 2011.
- [7] P. Baldi and S. r. Brunak. *Bioinformatics : the machine learning approach*. Adaptive computation and machine learning. MIT Press, Cambridge (Mass.), 1998.
- [8] J. Baranyi. Comparison of stochastic and deterministic concepts of bacterial lag. *Journal of Theoretical Biology*, 192(3):403 – 408, 1998.
- [9] J. Baranyi. Stochastic modelling of bacterial lag phase. *International Journal of Food Microbiology*, 73(2–3):203 – 206, 2002. 3rd International Conference on Predictive Modelling in Foods.

- [10] N. Barkai and B.-Z. Shilo. Variability and robustness in biomolecular systems. *Molecular Cell*, 28(5):755–760, 2007.
- [11] M. J. Bazin and M. S. d’Abadie. The relationship between cell and population growth. *Antonie van Leeuwenhoek*, 63, 1993.
- [12] R. Bellman and T. E. Harris. On the theory of age-dependent stochastic branching processes. *Proceedings of the National Academy of Sciences of the United States of America*, 34(12):pp. 601–604, 1948.
- [13] K. J. Boor. Bacterial stress responses: What doesn’t kill them can make them stronger. *PloS Biology*, 4, 2006.
- [14] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock. Go::termfinder open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, 2004.
- [15] R. H. Byrd, P. Lu, J. Nocedal, and C. Y. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(6):1190–1208, 1995.
- [16] O. Canela-Xandri, F. Sagués, and J. Buceta. Interplay between intrinsic noise and the stochasticity of the cell cycle in bacterial colonies. *Biophysical Journal*, 98, 2010.
- [17] Y.-H. M. Chan and W. F. Marshall. Scaling properties of cell and organelle size. *Organogenesis*, 6(2):88–96, 2010.
- [18] D. A. Charlebois and M. Kaern. An accelerated method for simulating population dynamics. *Communications in Computational Physics*, 14, 2013.
- [19] J. M. Cherry, E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, E. T. Chan, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, B. C. Hitz, K. Karra, C. J. Krieger, S. R. Miyasato, R. S. Nash,

- J. Park, M. S. Skrzypek, M. Simison, S. Weng, and E. D. Wong. Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Research*, 40(D1):D700–D705, 2012.
- [20] D. M. Chickering. Learning equivalence classes of bayesian-network structures. *J. Mach. Learn. Res.*, 2:445–498, 2002.
- [21] W. Cohen. *A Computer Scientist’s Guide to Cell Biology: A Travelogue from a Stranger in a Strange Land*. Springer, 2007.
- [22] Y. Cohen and J. Cohen. *Statistics and Data with R: An Applied Approach Through Examples*. Wiley, 2008.
- [23] J. F. Collins and M. H. Richmond. Rate of growth of bacillus cereus between divisions. *Journal of General Microbiology*, 28, 1962.
- [24] A. Colman-Lerner, A. Gordon, E. Serra, T. Chin, O. Resnekov, D. Endy, C. Gustavo Pesce, and R. Brent. Regulated cell-to-cell variation in a cell-fate decision system. *Nature*, 437(7059):699–706, 2005.
- [25] S. Cooper. Distinguishing between linear and exponential cell growth during the division cycle: single-cell studies, cell-culture studies, and the object of cell-cycle research. *Theoretical biology & medical modelling*, 3, 2006.
- [26] S. Cooper. Schizosaccharomyces pombe grows exponentially during the division cycle with no rate change points. *FEMS Yeast Research*, 13(7):650–658, 2013.
- [27] T. F. Cooper, E. A. Ostrowski, and M. Travisano. A negative relationship between mutation pleiotropy and fitness effect in yeast. *Evolution*, 61(6):1495–1499, 2007.
- [28] F. H. C. Crick. On protein synthesis. *The Symposia of the Society for Experimental Biology*, 12:138–163, 1958.
- [29] M. E. Csete and J. C. Doyle. Reverse engineering of biological complexity. *Science*, 295(5560):1664–1669, 2002.

- [30] J. Cullum and M. Vincente. Cell growth and length distribution in *Escherichia coli*. *Journal Of Bacteriology*, 134, 1978.
- [31] J. H. Curtiss. On the distribution of the quotient of two chance variables. *The Annals of Mathematical Statistics*, 12(4):pp. 409–421, 1941.
- [32] A. Davison. *Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2003.
- [33] D. Dubnau and R. Losick. Bistability in bacteria. *Molecular Microbiology*, 61, 2006.
- [34] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297, 2002.
- [35] C. Espinosa-Soto, O. C. Martin, and A. Wagner. Phenotypic robustness can increase phenotypic variability after nongenetic perturbations in gene regulatory circuits. *Journal of Evolutionary Biology*, 24(6):1284–1297, 2011.
- [36] B. Everitt and T. Hothorn. *A Handbook of Statistical Analyses Using R*. Chapman & Hall/CRC, 2006.
- [37] W. Feller. On the integral equation of renewal theory. *Ann. Math. Statist.*, 12(3):243–267, 1941.
- [38] L. Fernandez-Ricaud, J. Warringer, E. Ericson, K. Glaab, P. Davidsson, F. Nilsson, G. J. Kemp, O. Nerman, and A. Blomberg. PROPHECY—a yeast phenome database, update 2006. *Nucleic Acids Res*, 35(Database issue), 2007.
- [39] D. Fraser and M. Kaern. A chance at survival: gene expression noise and phenotypic diversification strategies. *Molecular Microbiology*, 71, 2009.
- [40] O. Fridman, A. Goldberg, I. Ronin, N. Shores, and N. Q. Balaban. Optimization of lag time underlies antibiotic tolerance in evolved bacterial populations. *Nature*, 513(7518):418–421, 2014.

- [41] D. Fudenberg and L. A. Imhof. Phenotype switching and mutants in random environments. *Bulletin of Mathematical Biology*, 2001.
- [42] J. M. Guberman, A. Fay, J. Dworkin, N. S. Wingreen, and Z. Gitai. Psicic: noise and asymmetry in bacterial division revealed by computational image analysis at sub-pixel resolution. *PLoS computational biology*, 4, 2008.
- [43] C. A. Hale, H. Meinhardt, and P. A. de Boer. Dynamic localization cycle of the cell division regulator *minE* in *Escherichia coli*. *The EMBO Journal*, 20, 2001.
- [44] T. E. Harris. *Some Mathematical Models for Branching Processes*. University of California Press, Berkeley, Calif., 1951.
- [45] P. D. N. Hebert, A. Cywinska, S. L. Ball, and J. R. Dewaard. Biological identifications through dna barcodes. *Proc. Biol. Sci*, 270:313–321, 2003.
- [46] N. S. Hill, R. Kadoya, D. K. Chatteraj, and P. A. Levin. Cell size and the initiation of DNA replication in bacteria. *PLoS genetics*, 8(3):e1002549, 2012.
- [47] M. Hjortso. *Population Balances in Biomedical Engineering*. McGraw Hill Professional, 2005.
- [48] W.-C. Ho and J. Zhang. The genotype-phenotype map of yeast complex traits: basic parameters and the role of natural selection. *Molecular biology and evolution*, 31(6):1568–1580, 2014.
- [49] S. L. Holland, T. Reader, P. S. Dyer, and S. V. Avery. Phenotypic heterogeneity is a selected trait in natural yeast populations subject to environmental stress. *Environmental Microbiology*, 16(6):1729–1740, 2014.
- [50] M. Howard and A. D. Rutenberg. Pattern formation inside bacteria: Fluctuations due to the low copy number of proteins. *Physical Review Letters*, 90, 2003.
- [51] M. Howard, A. D. Rutenberg, and S. de Vet. Dynamic compartmentalization of bacteria: Accurate division in *E. coli*. *Physical Review Letters*, 87, 2001.

- [52] M. D. D. Huber and L. Gerace. The size-wise nucleus: nuclear volume control in eukaryotes. *J Cell Biol*, 2007.
- [53] F. Jensen and T. Nielsen. *Bayesian Networks and Decision Graphs*. Springer, New York, NY, 2nd edition, 2007.
- [54] P. Jorgensen, N. P. Edgington, B. L. Schneider, I. Rupeš, M. Tyers, and B. Futcher. The size of the nucleus increases as yeast cells grow. *Molecular Biology of the Cell*, 18(9):3523–3532, 2007.
- [55] P. Jorgensen and M. Tyers. How cells coordinate growth and division. *Current Biology*, 14(23):R1014 – R1027, 2004.
- [56] S. Jun and S. Taheri-Araghi. Cell-size maintenance: universal strategy revealed. *Trends in Microbiology*, 23(1):4 – 6, 2015.
- [57] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, 9th edition, 1990.
- [58] H. Kitano. Towards a theory of biological robustness. *Molecular systems biology*, 3, 2007.
- [59] A. L. Koch and M. Schaechter. A model for statistics of the cell division process. *Journal of General Microbiology*, 29(3):435–454, 1962.
- [60] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.
- [61] K. Korb and A. Nicholson. *Bayesian Artificial Intelligence*. Chapman & Hall/CRC Computer Science & Data Analysis. Taylor & Francis, 2003.
- [62] O. Kotte, B. Volkmer, J. L. Radzikowski, and M. Heinemann. Phenotypic bistability in escherichia coli’s central carbon metabolism. *Molecular Systems Biology*, 10(7), 2014.
- [63] J.-U. Kreft, G. Booth, and J. W. Wimpenny. Bacsim, a simulator for individual-based modelling of bacterial colony growth. *Microbiology*, 144, 1998.

- [64] S. Krishnaswamy, M. H. Spitzer, M. Mingueneau, S. C. Bendall, O. Litvin, E. Stone, D. Pe'er, and G. P. Nolan. Conditional density-based analysis of t cell signaling in single-cell data. *Science*, 346(6213), 2014.
- [65] E. Kussell and S. Leibler. Phenotypic diversity, population growth, and information in fluctuating environments. *Science*, 309(5743):2075–2078, 2005.
- [66] E. Kussell and S. Leibler. Phenotypic diversity, population growth, and information in fluctuating environments. *Science*, 309, 2005.
- [67] M. Langlais. A nonlinear problem in age-dependent population diffusion. *SIAM J. Math. Anal.*, 16:pp. 510–529, 1985.
- [68] B. Lantz. *Machine Learning with R*. Community experience distilled. Packt Publishing, 2013.
- [69] D. Laporte, A. Lebaudy, A. Sahin, B. Pinson, J. Ceschin, B. Daignan-Fornier, and I. Sagot. Metabolic status rather than cell cycle signals control quiescence entry and exit. *The Journal of Cell Biology*, 192(6):949–957, 2011.
- [70] B. Lehner. Modelling genotype-phenotype relationships and human disease with genetic interaction networks. *J Exp Biol*, 210(9):1559–1566, 2007.
- [71] B. Lehner. Conflict between noise and plasticity in yeast. *PLoS genetics*, 6(11):e1001185, 2010.
- [72] B. Lehner. Genes confer similar robustness to environmental, stochastic, and genetic perturbations in yeast. *PloS one*, 5(2):e9035, 2010.
- [73] B. Lehner. Genes confer similar robustness to environmental, stochastic, and genetic perturbations in yeast. *PLoS ONE*, 5(2):e9035, 2010.
- [74] M. Leslie. How does a cell know its size. *Science*, 334, 2011.
- [75] D. L. Levy and R. Heald. Nuclear size is regulated by importin α and ntf2 in xenopus. *Cell*, 143(2):288–298, 2010.

- [76] S. F. Levy and M. L. Siegal. Network hubs buffer environmental variation in *saccharomyces cerevisiae*. *PLoS Biology*, 6:e264 EP –, 2008.
- [77] S. F. Levy, N. Ziv, and M. L. Siegal. Bet hedging in yeast by heterogeneous, age-correlated expression of a stress protectant. *PLoS Biol*, 10(5):e1001325, 2012.
- [78] H. Lodish, A. Berk, C. A. Kaiser, M. Krieger, M. P. Scott, A. Bretscher, H. Ploegh, and P. Matsudaira. *Molecular Cell Biology (Lodish, Molecular Cell Biology)*. W. H. Freeman, 6th edition, 2007.
- [79] A. W. Mahoney and D. Ramkrishna. Efficient solution of population balance equations with discontinuities by finite elements. *Chemical Engineering Science*, 57, 2002.
- [80] T. Malthus. *An Essay on the Principle of Population*. J. Johnson, London, 1798.
- [81] N. V. Mantzaris, P. Daoutidis, and F. Srienc. Numerical solution of multi-variable cell population models: I. finite difference methods. *Computers and Chemical Engineering*, 25, 2001.
- [82] N. V. Mantzaris, P. Daoutidis, and F. Srienc. Numerical solution of multi-variable cell population models: Ii. spectral methods. *Computers and Chemical Engineering*, 25, 2001.
- [83] N. V. Mantzaris, P. Daoutidis, and F. Srienc. Numerical solution of multi-variable cell population models: Iii. finite element methods. *Computers and Chemical Engineering*, 25, 2001.
- [84] D. L. Marchisio and R. O. Fox. Solution of population balance equations using the direct quadrature method of moments. *Journal of Aerosol Science*, 36, 2005.
- [85] F. Markowetz. How to understand the cell by breaking it: network analysis of gene perturbation screens. *PLoS computational biology*, 6(2):e1000655+, 2010.
- [86] W. Marshall, K. Young, M. Swaffer, E. Wood, P. Nurse, A. Kimura, J. Frankel, J. Wallingford, V. Walbot, X. Qu, and A. Roeder. What determines cell size? *BMC Biology*, 10(1):101, 2012.

- [87] W. F. Marshall. Cellular length control systems. *Annual Review of Cell and Developmental Biology*, 20(1):677–693, 2004.
- [88] J. Masel and M. L. Siegal. Robustness: mechanisms and consequences. *Trends in Genetics*, 25(9):395–403, 2009.
- [89] J. H. McDonald. *Handbook of Biological Statistics*. Sparky House Publishing, second edition, 2009.
- [90] J. Monod. The growth of bacterial cultures. *Annual Review of Microbiology*, 3, 1949.
- [91] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.
- [92] I. Nachman, A. Regev, and S. Ramanathan. Dissecting timing variability in yeast meiosis. *Cell*, 131(3):544–556, 2007.
- [93] R. Nagano, H. Akanuma, X.-Y. Qin, S. Imanishi, H. Toyoshiba, J. Yoshinaga, S. Ohsako, and H. Sone. Multi-Parametric profiling network based on gene expression and phenotype data: A novel approach to developmental neurotoxicity testing. *International Journal of Molecular Sciences*, 13(1):187–207, 2011.
- [94] R. Nagarajan, M. Scutari, and S. Lebre. *Bayesian Networks in R with Applications in Systems Biology*. Springer, New York, 2013. ISBN 978-1461464457.
- [95] R. E. Neapolitan. *Learning Bayesian Networks*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2003.
- [96] R. E. Neapolitan. *Learning Bayesian Networks*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2003.
- [97] F. R. Neumann and P. Nurse. Nuclear size control in fission yeast. *Journal of Cell Biology*, 179(4):593–600, 2007.

- [98] J.-B. Nobs and S. J. Maerkl. Long-term single cell analysis of *italic*s. *pombe* on a microfluidic microchemostat array. *PLoS ONE*, 9(4):e93466, 2014.
- [99] Y. Ohya, J. Sese, M. Yukawa, F. Sano, Y. Nakatani, T. L. Saito, A. Saka, T. Fukuda, S. Ishihara, S. Oka, G. Suzuki, M. Watanabe, A. Hirata, M. Ohtani, H. Sawai, N. Frayssé, J.-P. Latgé, J. M. François, M. Aebi, S. Tanaka, S. Muramatsu, H. Araki, K. Sonoike, S. Nogami, and S. Morishita. High-dimensional and large-scale phenotyping of yeast mutants. *Proceedings of the National Academy of Sciences of the United States of America*, 102(52):19015–19020, 2005.
- [100] M. Osella, E. Nugent, and M. Cosentino Lagomarsino. Concerted control of *escherichia coli* cell division. *Proceedings of the National Academy of Sciences*, 111(9):3431–3435, 2014.
- [101] V. Pancaldi. Biological noise to get a sense of direction: an analogy between chemotaxis and stress response. *Frontiers in Genetics*, 5:52, 2014.
- [102] V. Pancaldi, F. Schubert, and J. Bähler. Meta-analysis of genome regulation and expression variability across hundreds of environmental and genetic perturbations in fission yeast. *Molecular BioSystems*, 6(3):543–552, 2010.
- [103] J. Pearl. *Probabilistic Reasoning in Intelligent Systems, Networks of Plausible Inference*. Morgan Kaufmann, Palo Alto, 1988.
- [104] R. Phillips, J. Kondev, and J. Theriot. *Physical biology of the cell*. Garland Science, 2009.
- [105] C. Picioreanu, J.-U. Kreft, and M. C. van Loosdrecht. Particle-based multidimensional multispecies biofilm model. *Applied and Environmental Microbiology*, 70, 2004.
- [106] O. Pierucci. Dimensions of *escherichia coli* at various growth rates: Model for envelope growth. *Journal Of Bacteriology*, 135, 1978.

- [107] E. O. POWELL. A note on koch & schaechter's hypothesis about growth and fission of bacteria. *Journal of General Microbiology*, 37(2):231–249, 1964.
- [108] P. P.R. and M. A. G. Inequality of mean interdivision time and doubling time. *Journal of General Microbiology*, 48:pp. 155–9, 1967.
- [109] W. Qian, D. Ma, C. Xiao, Z. Wang, and J. Zhang. The genomic landscape and evolutionary resolution of antagonistic pleiotropy in yeast. *Cell reports*, 2012.
- [110] A. Raj and A. van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 2008.
- [111] D. Ramkrishnan. *Population balances. Theory and applications to particulate systems in engineering*. Academic Press, 2000.
- [112] Reynolds, A., Richards, G., Iglesia, B., Rayward-Smith, and V. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5(4):475–504, 2006.
- [113] S. Rigopoulos and A. G. Jones. Finite-element scheme for solution of the dynamic population balance equation. *AIChE Journal*, 49, 2003.
- [114] R. Rinott, A. Jaimovich, and N. Friedman. Exploring transcription regulation through cell-to-cell variability. *Proceedings of the National Academy of Sciences of the United States of America*, 108(15):6329–6334, 2011.
- [115] M. Robinson, J. Grigull, N. Mohammad, and T. Hughes. Funspec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, 3(1):35, 2002.
- [116] S. Ross. *A First Course in Probability*. Pearson Prentice Hall, 2010.
- [117] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, 1987.

- [118] K. Sajiki, M. Hatanaka, T. Nakamura, K. Takeda, M. Shimanuki, T. Yoshida, Y. Hanyu, T. Hayashi, Y. Nakaseko, and M. Yanagida. Genetic control of cellular quiescence in *s. pombe*. *Journal of cell science*, 2009.
- [119] A. Sanchez and I. Golding. Genetic Determinants and Cellular Constraints in Noisy Gene Expression. *Science*, 342(6163):1188–1193, 2013.
- [120] M. Schaechter, J. P. Williamson, J. R. Hood, and A. L. Koch. Growth, cell and nuclear divisions in some bacteria. *Journal of General Microbiology*, 29(3):421–434, 1962.
- [121] M. Schmidt, M. Creutziger, and P. Lenz. Influence of molecular noise on the growth of single cells and bacterial populations. *PloS One*, 7, 2012.
- [122] M. Scutari. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, 35(3):1–22, 2010.
- [123] V. Shahrezaei and P. S. Swain. The stochastic nature of biochemical networks. *Current opinion in biotechnology*, 19(4):369–374, 2008.
- [124] A. Solopova, J. van Gestel, F. J. Weissing, H. Bachmann, B. Teusink, J. Kok, and O. P. Kuipers. Bet-hedging during bacterial diauxic shift. *Proceedings of the National Academy of Sciences*, 111(20):7427–7432, 2014.
- [125] S. L. Spencer, S. Gaudet, J. G. Albeck, J. M. Burke, and P. K. Sorger. nature08012-2. *Nature*, 459(7245):428–432, 2009.
- [126] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Adaptive computation and machine learning. MIT Press, 2000.
- [127] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.
- [128] J. L. Spudich and D. E. Koshland. Non-genetic individuality: chance in the single cell. *Nature*, 262(5568):467–471, 1976.

- [129] M. Stumpf, D. Balding, and M. Girolami. *Handbook of Statistical Systems Biology*. Wiley, 2011.
- [130] G. M. Suel, R. P. Kulkarni, J. Dworkin, J. Garcia-Ojalvo, and M. B. Elowitz. Tunability and Noise Dependence in Differentiation Dynamics. *Science*, 315(5819):1716–1719, 2007.
- [131] P. S. Swain, M. B. Elowitz, and E. D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800, 2002.
- [132] P. S. Swain, M. B. Elowitz, and E. D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12795–12800, 2002.
- [133] S. Taheri-Araghi, S. Bradde, J. Sauls, N. Hill, P. Levin, J. Paulsson, M. Vergassola, and S. Jun. Cell-size control and homeostasis in bacteria. *Current Biology*, 25(0):–, 2014.
- [134] Y. Tanouchi, A. Pai, H. Park, S. Huang, R. Stamatov, N. E. Buchler, and L. You. A noisy linear map underlies oscillations in cell size and gene expression in bacteria. *Nature*, advance online publication, 2015.
- [135] T. Toni and B. Tidor. Combined model of intrinsic and extrinsic variability for computational network design with application to synthetic biology. *PLoS Comput Biol*, 9(3):e1002960, 2013.
- [136] I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Mach. Learn.*, 65(1):31–78, 2006.
- [137] P. Turchin. *Complex Population Dynamics: A Theoretical/Empirical Synthesis*. Monographs in population biology. Princeton University Press, 2003.
- [138] J. J. Tyson. Effects of asymmetric division on a stochastic model of the cell division cycle. *Mathematical Biosciences*, 96, 1989.

- [139] J. J. Tyson and O. Diekmann. Sloppy size control of the cell division cycle. *Journal of theoretical biology*, 118, 1986.
- [140] T. S. Verma and J. Pearl. Equivalence and synthesis of causal models. In P. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer, editors, *UAI '90: Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 255–268. Elsevier Science Publishers B.V., 1990.
- [141] M. Vignes, J. Vandel, D. Allouche, N. Ramadan-Alban, C. Cierco-Ayrolles, T. Schiex, B. Mangin, and S. de Givry. Gene regulatory network reconstruction using bayesian networks, the dantzig selector, the lasso and their Meta-Analysis. *PLoS ONE*, 6(12):e29165+, 2011.
- [142] G. Waite and L. Waite. *Applied Cell and Molecular Biology for Engineers*. McGraw-Hill Education, 2007.
- [143] G.-Z. Wang, J. Liu, W. Wang, H.-Y. Zhang, and M. J. Lercher. A gene’s ability to buffer variation is predicted by its fitness contribution and genetic interactions. *PloS one*, 6(3):e17650, 2011.
- [144] P. Wang, L. Robert, J. Pelletier, W. L. Dang, F. Taddei, A. Wright, and S. Jun. Robust growth of escherichia coli. *Current Biology*, 20, 2010.
- [145] P. I. Wang and E. M. Marcotte. It’s the machine that matters: Predicting gene function and phenotype from protein networks. *Journal of Proteomics*, 73(11):2277–2289, 2010.
- [146] D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G. D. Bader, and Q. Morris. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38(suppl 2):W214–W220, 2010.

- [147] M. Webster, K. L. Witkin, and O. Cohen-Fix. Sizing up the nucleus: nuclear shape, size and nuclear-envelope assembly. *Journal of Cell Science*, 122(10):1477–1486, 2009.
- [148] L. Weinberger, Y. Voichek, I. Tirosh, G. Hornung, I. Amit, and N. Barkai. Expression noise and acetylation profiles distinguish {HDAC} functions. *Molecular Cell*, 47(2):193 – 202, 2012.
- [149] E. Wilson. *The Cell in Development and Inheritance*. Biological series. Macmillan, 1896.
- [150] K. L. Witkin, Y. Chong, S. Shao, M. T. Webster, S. Lahiri, A. D. Walters, B. Lee, J. L. Koh, W. A. Prinz, B. J. Andrews, and O. Cohen-Fix. The budding yeast nuclear envelope adjacent to the nucleolus serves as a membrane sink during mitotic delay. *Current Biology*, 22(12):1128 – 1133, 2012.
- [151] V. Wood, M. A. Harris, M. D. McDowall, K. Rutherford, B. W. Vaughan, D. M. Staines, M. Aslett, A. Lock, J. Bähler, P. J. Kersey, and S. G. Oliver. Pombase: a comprehensive online resource for fission yeast. *Nucleic Acids Research*, 40(D1):D695–D699, 2012.
- [152] G. Yvert, S. Ohnuki, S. Nogami, Y. Imanaga, S. Fehrmann, J. Schacherer, and Y. Ohya. Single-cell phenomics reveals intra-species variation of phenotypic noise in yeast. *BMC Systems Biology*, 7(1):54, 2013.

Page intentionally left blank

Appendix A

Modelling the dynamic transition from cellular quiescence to proliferation in *S. pombe*

1. Akaike Information Criterion

1.1 Derivation of the AIC Score when the Residuals Follow a Normal Distribution

Supposing that the n residuals follow a normal distribution $\mathcal{N}(0, \sigma^2)$, the maximum likelihood for $\hat{\sigma}^2$ is given by

$$L = \prod_{i=1}^n \left(\frac{1}{2\pi\hat{\sigma}^2} \right)^{1/2} \exp \left(- \sum_{i=1}^n \frac{(y_i - f(x_i))^2}{2\hat{\sigma}_i^2} \right).$$

Thus, taking the logarithm yields

$$\begin{aligned} \log(L) &= \log \left(\prod_{i=1}^n \left(\frac{1}{2\pi\hat{\sigma}^2} \right)^{1/2} \right) - \frac{1}{2} \sum_{i=1}^n \frac{(y_i - f(x_i))^2}{2\hat{\sigma}_i^2} \\ &= C - \frac{1}{2} \sum_{i=1}^n \frac{(y_i - f(x_i))^2}{2\hat{\sigma}^2}, \end{aligned}$$

where C is constant.

1.2 Distribution of the Residuals

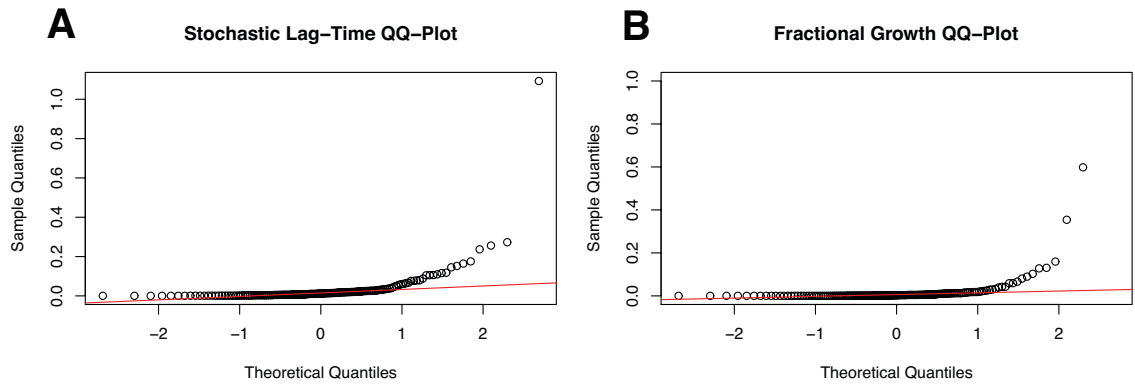


Figure A.1: QQ-Plots of the residuals between the data points and the models. A. QQ-Plot for the stochastic lag-time model. B. QQ-Plot for the fractional growth of the population model.

2. Residuals for the Stochastic Lag Model

We present here the figures for the residuals of the mutants, computed using the wild-type trends, for the two models.

2.1 Residuals for the stochastic lag-time for the mutants at 25 Degrees Celsius

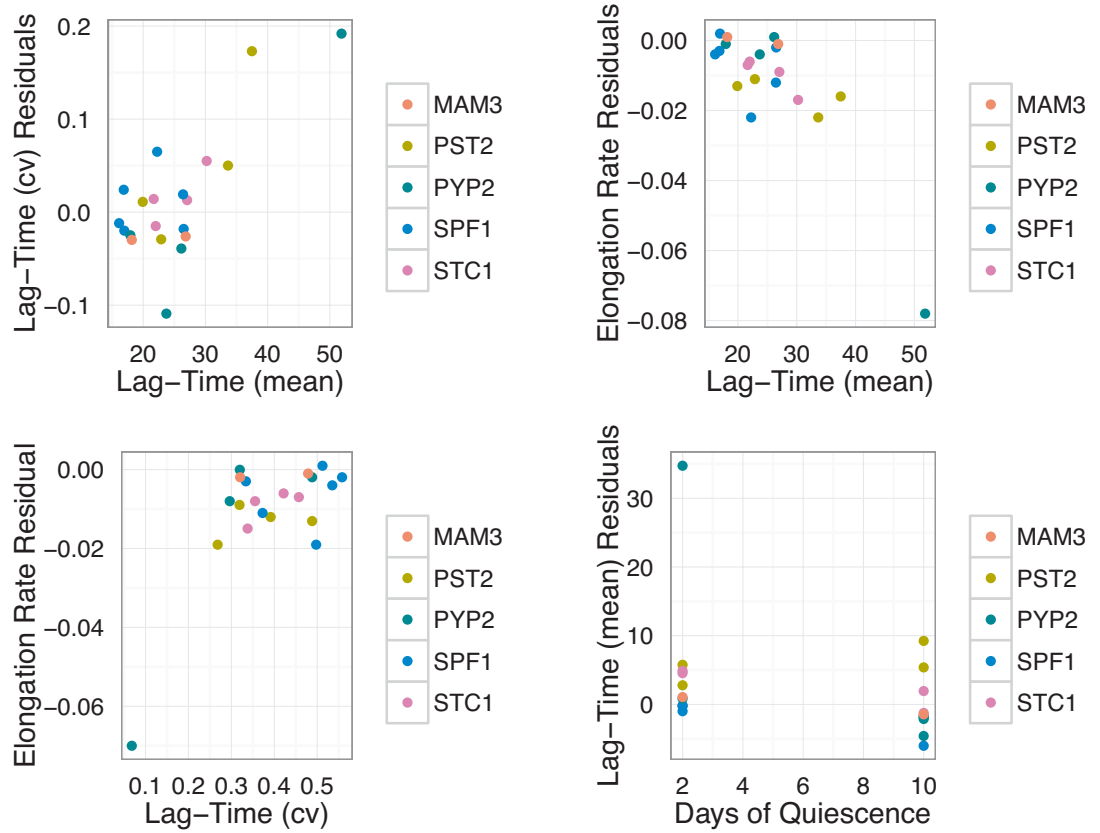
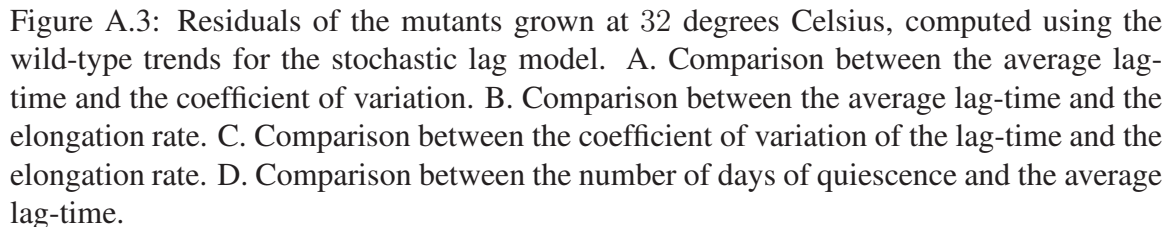


Figure A.2: Residuals of the mutants grown at 25 degrees Celsius, computed using the wild-type trends for the stochastic lag model. A. Comparison between the average lag-time and the coefficient of variation. B. Comparison between the average lag-time and the elongation rate. C. Comparison between the coefficient of variation of the lag-time and the elongation rate. D. Comparison between the number of days of quiescence and the average lag-time.

Celsius



3. Residuals for the Fractional Growth of the Population

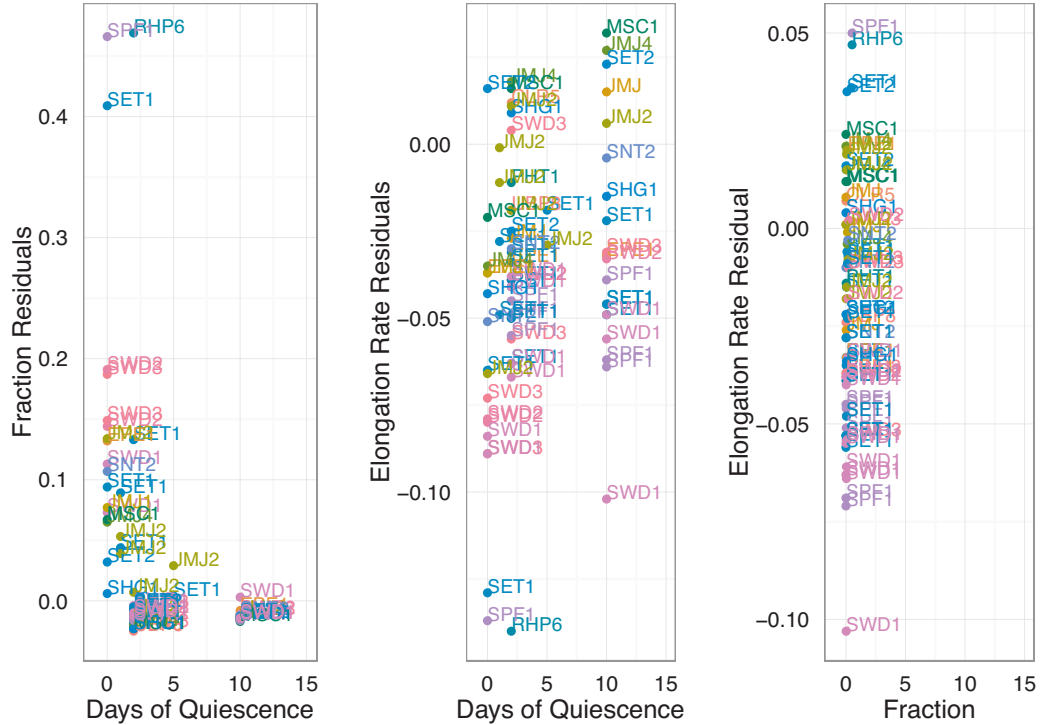


Figure A.4: Residuals of the mutants grown at 32 degrees Celsius, computed using the wild-type trends for the fractional growth of the population. A. Comparison between the number of days of quiescence and the fraction of cells growing. B. Comparison between the number of days of quiescence and the elongation rate. C. Comparison between the fraction of cells growing and the elongation rate.

Page intentionally left blank

Appendix B

Global trends in Phenotypic Variability and Robustness

1. Clustering of the Morphological Phenotypes

Phenotype	Cluster	Description
A7.1_A	1	Actin region size in mother cell on nucleus A
A101_A	1	Actin region ratio on nucleus A
A101_A1B	2	Actin region ratio on nucleus A1B
A7.1_A1B	2	Actin region size in mother cell on nucleus A1B
A7.1_C	3	Actin region size in mother cell on nucleus C
A101_C	3	Actin region ratio on nucleus C
A104_A1B	4	Relative Distance of actin patch center from neck in bud on nucleus A1B
A102_A1B	4	Actin region ratio in bud
A102_C	5	Actin region ratio in bud on nucleus C
A104_C	6	Relative Distance of actin patch center from neck in bud on nucleus C
A121_A	7	Maximum actin patch length on nucleus A
A121_C	7	Maximum actin patch length on nucleus C
A121_A1B	7	Maximum actin patch length on nucleus A1B
A120_C	7	Total length of actin patch link on nucleus C
A120_A	7	Total length of actin patch link on nucleus A
A120_A1B	7	Total length of actin patch link on nucleus A1B
A122_C	8	Number of actin patches on nucleus C
A122_A1B	8	Number of actin patches on nucleus A1B
A122_A	8	Number of actin patches on nucleus A
A123_A1B	8	Ratio of actin patch region to actin region on nucleus A1B
A123_C	8	Ratio of actin patch region to actin region on nucleus C

Continued...

Phenotype	Cluster	Description
A123_A	8	Ratio of actin patch region to actin region on nucleus A
C11.2_A1B	9	Area of daughter cell on nucleus A1B
C12.2_A1B	9	Contour length of daughter cell on nucleus A1B
C107_A1B	9	Long axis length in bud on nucleus A1B
A7.2_A1B	9	Actin region size in bud on nucleus A1B
C108_A1B	9	Short axis length in bud on nucleus A1B
C113_A1B	9	Distance from bud tip to mother cell's long axis along bud direction on nucleus A1B
A7.2_C	10	Actin region size in bud on nucleus C
A8.1_C	11	Actin region brightness in mother cell on nucleus C
A8.1_A1B	11	Actin region brightness in mother cell on nucleus A1B
A8.1_A	11	Actin region brightness in mother cell on nucleus A
A8.2_C	11	Actin region brightness in bud Actin region brightness in bud on nucleus C
A8.2_A1B	11	Actin region brightness in bud on nucleus A1B
A9_A1B	12	Ratio of actin region to bud neck on nucleus A1B
A9_C	13	Ratio of actin region to bud neck on nucleus C
C12.1_A1B	14	Contour length of mother cell on nucleus A1B
C11.1_A1B	14	Mother cell size on nucleus A1B
C101_A1B	14	Cell size on nucleus A1B
C112_A1B	14	Distance from neck to mother cell's center on nucleus A1B
C128_A1B	14	Length from bud neck to the farthest point on mother cell on nucleus A1B
C102_A1B	14	Contour length of cell on nucleus A1B
C103_A1B	14	Long axis length of mother cell on nucleus A1B
C11.1_C	14	Mother cell size on nucleus C
C12.2_C	15	Contour length of daughter cell on nucleus C
C11.2_C	15	Area of daughter cell on nucleus C
C107_C	15	Long axis length in bud on nucleus C
C102_C	15	Contour length of cell on nucleus C
D185_C	15	distance between nuclei through neck
D186_C	15	Relative distance of two nuclear brightest points to middle point of neck on stage C
C108_C	15	Short axis length in bud on nucleus C
C113_C	15	Distance from bud tip to mother cell's long axis along bud direction on nucleus C
C101_C	15	Cell size on nucleus C
C103_C	15	Long axis length of mother cell on nucleus C
C112_C	15	Distance from neck to mother cell's center on nucleus C
C128_C	15	Length from bud neck to the farthest point on mother cell on nucleus C
C12.1_C	15	Contour length of mother cell on nucleus C
C12.1_A	16	Contour length of mother cell on nucleus A
D102_A	16	Distance from nuclear center to mother tip in nucleus A
D127_A	16	Distance between nuclear brightest point and cell tip
C11.1_A	16	Mother cell size on nucleus A
C103_A	16	Long axis length of mother cell on nucleus A
C104_A	16	Short axis length of mother cell on nucleus A

Continued. . .

Phenotype	Cluster	Description
C109_C	17	Neck width on stage C
C109_A1B	17	Neck width on stage A1B
C104_C	17	Short axis length of mother cell on nucleus C
C104_A1B	17	Short axis length of mother cell on nucleus A1B
C106_A1B	18	Bud growth direction on nucleus A1B
C105_A1B	18	Neck position on nucleus A1B
C110_A1B	18	Length from bud tip to mother cell's long axis on nucleus A1B
D158_C	19	Angle between D18.1D1.1 and D18.1C1.2 on stage C
D159_C	19	Angle between D19.1D2.1 and D19.1C1.2 on stage C
C106_C	19	Bud growth direction on nucleus C
C105_C	19	Neck position on nucleus C
C110_C	19	Length from bud tip to mother cell's long axis on nucleus C
C111_C	20	Length from bud tip to mother cell's short axis on nucleus C
C111_A1B	20	Length from bud tip to mother cell's short axis on nucleus A1B
C115_A	21	Roundness of mother cell on nucleus A
C115_A1B	21	Roundness of mother cell on nucleus A1B
C114_C	21	Roundness of bud on nucleus C
C115_C	21	Roundness of mother cell on nucleus C
C114_A1B	21	Roundness of bud on nucleus A1B
C116_A1B	22	Ratio of roundness of mother cell to that of bud in nucleus A1B
C116_C	23	Ratio of roundness of mother cell to that of bud in nucleus C
C117_A1B	24	Ratio of the countour length on nucleus A1B
C118_A1B	24	Ratio of the cell sizes on nucleus A1B
C117_C	25	Ratio of the countour length on nucleus C
C118_C	25	Ratio of the cell sizes on nucleus C
C126_A1B	26	Unevenness of brightness on nucleus A1B
C126_C	26	Unevenness of brightness on nucleus C
C126_A	26	Unevenness of brightness on nucleus A
C127_A	27	Unevenness of cell wall thickness on nucleus A
C127_C	27	Unevenness of cell wall thickness on nucleus C
C127_A1B	27	Unevenness of cell wall thickness on nucleus A1B
C13_C	28	Fitness in nucleus C
C13_A	28	Fitness in nucleus A
C13_A1B	28	Fitness in nucleus A1B
D125_C	29	Distance between nuclear gravity center in mother and mother hip
D141_C	29	Distance between nuclear brightest point in mother and mother hip on stage C
D103_C	29	Distance from nuclear center to mother tip in nucleus C
D128_C	29	Distance between nuclear brightest point in mother and mother tip
D145_C	29	Distance between nuclear outline point C7 and mother hip on stage A1B
D118_A1B	30	Distance from mother cell's center to nucleus center in A1B
D104_A1B	30	Distance from nuclear center to mothre tip in nucleus A1B
D126_A1B	30	Distance between nuclear gravity center and mother hip

Continued...

Phenotype	Cluster	Description
D145_A1B	30	Distance between nuclear outline point C7 and mother hip on stage A1B
D142_A1B	30	Distance between nuclear brightest point and mother hip on stage A1B
D129_A1B	30	Distance between nuclear brightest point and mother tip
D136_A1B	30	Distance between nuclear brightest point and mother center
D105_A	31	Ratio of D102 to C103 in nucleus A
D147_A	31	Relative distance of nuclear gravity center to cell center on stage A
D148_A	31	Relative distance of nuclear brightest point to cell center on stage A
D106_C	32	Ratio of D103 to C103 in nucleus C
D147_A1B	33	Relative distance of nuclear gravity center to cell center on stage A
D148_A1B	33	Relative distance of nuclear brightest point to cell center on stage A
D107_A1B	33	Ratio of D104 to C103 in nucleus A1B
D108_C	34	Distance from neck to mother cell's nucleus in nucleus C
D130_C	34	Distance between nuclear brightest point in mother and middle point of neck
D143_C	34	Nucleus border point close to neck in nucleus A1B
D134_C	34	Distance between two nuclear brightest points
D116_C	34	Distance between two nucleus in nucleus C
D109_C	35	Distance from neck to bud's nucleus in nucleus C
D131_C	35	Distance between nuclear brightest point in bud and middle point of neck
D144_C	35	Nucleus border point close to neck on bud's nucleus
D110_A1B	36	Distance from neck to nucleus center in nucleus A1B
D132_A1B	36	Distance between nuclear brightest point and middle point of neck
D143_A1B	36	Nucleus border point close to neck in nucleus A1B
D112_C	37	Ratio of D108 to C128 on stage C
D152_C	37	Mobility of nucleus in mother cell in nucleus in nucleus C
D113_C	38	Ratio of D109 to C107
D153_C	38	Mobility of nucleus in bud in nucleus C
D114_A1B	39	Ratio of D110 to C128 on stage A1B
D152_A1B	39	Mobility of nucleus in mother cell in nucleus A1B
D135_A	40	Distance between nuclear brightest point and cell center
D117_A	40	Distance from mother cell's center to nucleus in nucleus A
D148_C	41	Relative distance of nuclear brightest point to cell center on stage A
D147_C	41	Relative distance of nuclear gravity center to cell center on stage A
D135_C	41	Distance between nuclear brightest point and cell center
D117_C	41	Distance from mother cell's center to mother cell's nucleus in nucleus C
D150_C	42	Relative distance of nuclear brightest point in bud to bud center on stage C
D149_C	42	Relative distance of nuclear gravity center in bud to bud center on stage C
D119_C	42	Distance from bud center to bud's nucleus in nucleus C
D137_C	42	Distance between nuclear brightest point in bud and bud tip on stage C
D121_C	43	Distance from bud nucleus to bud tip in nucleus C
D139_C	43	Distance between nuclear brightest point in bud and bud tip on stage C
D146_C	43	Distance between nuclear outline point C8 in bud and bud tip on stage C
D123_C	44	Ratio of D121 to C107 on stage C

Continued...

Phenotype	Cluster	Description
D176_A	45	nucleus diameter in mother cell
D173_A	45	nucleus maximum radius in mother cell
D14.1_A	45	Area of nucleus region in mother cell in nucleus A
D179_A	45	nucleus minimum radius in mother cell
D176_C	46	nucleus diameter in mother cell
D173_C	46	nucleus maximum radius in mother cell
D14.1_C	46	Area of nucleus region in mother cell in nucleus C
D179_C	46	nucleus minimum radius in mother cell
D14.3_C	46	Area of nucleus region in nucleus C
D181_A1B	46	nucleus minimum radius
D14.2_C	47	Area of nucleus region in bud in nucleus C
D177_C	47	nucleus diameter in bud
D174_C	47	nucleus maximum radius in bud
D180_C	47	nucleus minimum radius in bud
D178_A1B	48	nucleus diameter
D175_A1B	48	nucleus maximum radius
D14.3_A1B	48	Area of nucleus region in nucleus A1B
D15.3_C	49	Sum of brightness in nucleus region in nucleus C
D15.1_C	49	Sum of brightness in nucleus region in mother cell in nucleus C
D15.1_A	49	Sum of brightness in nucleus region in mother cell in nucleus A
D15.2_C	49	Sum of brightness in nucleus region in bud in nucleus C
D15.3_A1B	49	Sum of brightness in nucleus region in nucleus A1B
D151_C	50	Distance ratio of two nuclei from neck in nucleus C
D155_A	51	Angle between C1D2.1 and C1C1.2 on stage A
D154_A	51	Angle between C1D1.1 and C1C1.2 on stage A
D155_A1B	52	Angle between C1D2.1 and C1C1.2 on stage A
D154_A1B	52	Angle between C1D1.1 and C1C1.2 on stage A
D154_C	53	Angle between C1D1.1 and C1C1.2 on stage A
D155_C	53	Angle between C1D2.1 and C1C1.2 on stage A
D156_C	54	Angle between C2D1.2 and C2C4.2 on stage C
D157_C	54	Angle between C2D2.2 and C2C4.2 on stage C
D193_C	55	Average of nuclear brightness in whole cell
D191_C	55	Average of nuclear brightness in mother cell
D16.1_C	55	Maximum brightness of a nucleus in the mother cell
D16.3_C	55	Maximum brightness of nuclei
D192_C	55	Average of nuclear brightness in bud
D191_A	55	Average of nuclear brightness in mother cell
D16.2_C	55	Maximum brightness of nuclei
D16.1_A	55	Maximum brightness of a nucleus in the mother cell
D16.3_A1B	56	Maximum brightness of nuclei
D193_A1B	56	Average of nuclear brightness in whole cell
D162_C	57	slope between two nuclei to neck position

Continued...

Phenotype	Cluster	Description
D163_C	57	Angle between D23D2.1 and D23C1 on stage C
D167_C	58	Angle between D19.2D2.2 and D19.2C4.2 on stage C
D166_C	58	Angle between D18.2D1.2 and D18.2C4.2 on stage C
D169_A1B	59	Angle between M1D1.1 and M1C1 on stage A1B
D170_A1B	59	Angle between M1D2.1 and M1C1 on stage A1B
D170_C	60	Angle between M1D2.1 and M1C1 on stage A1B
D169_C	60	Angle between M1D1.1 and M1C1 on stage A1B
D182_A	61	nucleus roundness in mother cell
D17.1_A	61	Fitness to ellipse of the nucleus in the mother cell
D17.2_C	62	Fitness to ellipse of the nucleus in the daughter cell
D183_C	62	nucleus roundness in bud
D17.1_C	62	Fitness to ellipse of the nucleus in the mother cell
D184_A1B	63	nucleus roundness
D17.3_A1B	63	Fitness to ellipse of the nucleus
D182_C	64	nucleus roundness in mother cell
D189_C	65	Distance between nuclear gravity center and brightest point in bud
D188_A	65	Distance between nuclear gravity center and brightest point on stage A
D188_C	65	Distance between nuclear gravity center and brightest point on stage A
D196_C	66	Maximal intensity of nuclear brightness divided by average
D194_A	66	Maximal intensity of nuclear brightness divided by average
D194_C	66	Maximal intensity of nuclear brightness divided by average
D196_A1B	66	Maximal intensity of nuclear brightness divided by average
D195_C	66	Maximal intensity of nuclear brightness divided by average in mother
D198_C	67	Ratio of nuclear brightness
D197_C	67	nuclei size ratio

2. Analysis of Fitness for the Residuals Calculated with the Coefficient of Variation instead of the $L1$ -regression

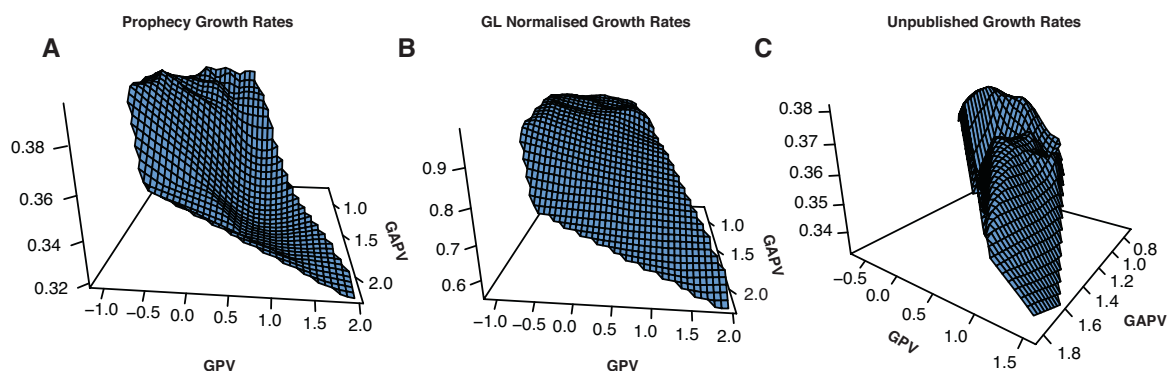


Figure B.1: Similarly to our first analysis using the $L1$ -regression to calculate the residuals, we plot the growth rates of the three main categories of genes, based on the new measure of variability.

3. Gene Ontology Annotations Obtained with the Residuals Calculated with the Coefficient of Variation instead of the $L1$ -regression

Inducers

GO Molecular Function	p-value	k	f
RNA helicase activity [GO:0003724]	0.004331	2	10
GO Biological Process			
proline transport [GO:0015824]	0.0001014	2	2
response to virus [GO:0009615]	0.0006007	2	4
establishment of cell polarity [GO:0030010]	0.0007394	4	41
protein targeting to vacuole [GO:0006623]	0.008084	3	41
regulation of cell shape [GO:0008360]	0.008532	2	14
GO Cellular Component			
polysomal ribosome [GO:0042788]	0.0006007	2	4
septin complex [GO:0031105]	0.002061	2	7
cellular bud neck septin ring [GO:0000144]	0.004331	2	10
nuclear nucleosome [GO:0000788]	0.005259	2	11
nucleosome [GO:0000786]	0.005259	2	11

Maintainers

GO Molecular Function	p-value	k	f
structural constituent of ribosome [GO:0003735]	2.62E-06	111	218
catalytic activity [GO:0003824]	0.001146	194	455
transferase activity [GO:0016740]	0.002159	252	611
sequence-specific DNA binding [GO:0043565]	0.004	76	165
transferase activity, transferring pentosyl groups [GO:0016763]	0.005903	5	5
acyltransferase activity [GO:0008415]	0.007711	35	69
GO Biological Process			
carbohydrate metabolic process [GO:0005975]	2.06E-06	56	94
glycerol metabolic process [GO:0006071]	3.46E-05	10	10
pseudohyphal growth [GO:0007124]	3.55E-05	39	64
metabolic process [GO:0008152]	0.0003051	186	425
late endosome to vacuole transport [GO:0045324]	0.0004114	15	20
response to osmotic stress [GO:0006970]	0.001078	19	29
phosphatidylcholine biosynthetic process [GO:0006656]	0.001661	8	9
fatty acid metabolic process [GO:0006631]	0.002013	18	28
MAPKKK cascade involved in cell wall biogenesis [GO:0000196]	0.002113	6	6
protein retention in Golgi apparatus [GO:0045053]	0.002454	9	11
GO Cellular Component			
cytoplasm [GO:0005737]	2.61E-10	839	2026
mitochondrial membrane [GO:0031966]	2.82E-07	41	60
ribosome [GO:0005840]	7.53E-07	152	310
mitochondrion [GO:0005739]	2.16E-05	444	1072
mitochondrial large ribosomal subunit [GO:0005762]	8.57E-05	28	43
intracellular [GO:0005622]	0.0003857	168	381
peroxisome [GO:0005777]	0.001393	36	66
endosome [GO:0005768]	0.00165	54	108
nuclear telomeric heterochromatin [GO:0005724]	0.002113	6	6
endosome membrane [GO:0010008]	0.003112	31	57

Capacitors

GO Molecular Function	p-value	k	f
proton-transporting ATPase activity, rotational mechanism [GO:0046961]	0.0001282	9	29
hydrogen ion transmembrane transporter activity [GO:0015078]	0.0003792	6	15
telomeric DNA binding [GO:0042162]	0.0005703	6	16
tubulin binding [GO:0015631]	0.0007706	4	7
protein-cysteine S-palmitoyltransferase activity [GO:0019706]	0.001454	4	8
protein heterodimerization activity [GO:0046982]	0.001454	4	8
microtubule motor activity [GO:0003777]	0.003885	4	10
general transcriptional repressor activity [GO:0016565]	0.005143	2	2
G-protein alpha-subunit binding [GO:0001965]	0.005143	2	2
H3 histone acetyltransferase activity [GO:0010484]	0.005143	2	2
nicotinamide riboside transporter activity [GO:0034257]	0.005143	2	2
ubiquitin-protein ligase activity [GO:0004842]	0.008235	12	77
GO Biological Process			
response to DNA damage stimulus [GO:0006974]	9.85E-10	40	197
double-strand break repair via single-strand annealing [GO:0045002]	8.74E-07	6	7
mitotic sister chromatid cohesion [GO:0007064]	2.59E-06	10	24
DNA repair [GO:0006281]	5.06E-06	31	183
chromatin remodeling [GO:0006338]	5.19E-06	14	49
negative regulation of transcription, RNA-mediated [GO:0010526]	9.25E-06	6	9
mRNA 3'-end processing [GO:0031124]	1.10E-05	7	13
mRNA export from nucleus [GO:0006406]	3.20E-05	12	43
histone exchange [GO:0043486]	4.49E-05	6	11
vacuolar acidification [GO:0007035]	4.86E-05	9	26
GO Cellular Component			
prefoldin complex [GO:0016272]	3.47E-05	5	7
proton-transporting V-type ATPase, V0 domain [GO:0033179]	0.0001237	4	5
carboxy-terminal domain protein kinase complex [GO:0032806]	0.0003678	3	3
telomerase holoenzyme complex [GO:0005697]	0.0003678	3	3
vacuolar proton-transporting V-type ATPase, V0 domain [GO:0000220]	0.0007706	4	7
proton-transporting two-sector ATPase complex, proton-transporting domain [GO:0033177]	0.0007706	4	7
THO complex part of transcription export complex [GO:0000445]	0.001392	3	4
Cul8-RING ubiquitin ligase complex [GO:0035361]	0.001392	3	4
cytoplasmic dynein complex [GO:0005868]	0.001392	3	4
nucleoplasmic THO complex [GO:0000446]	0.001392	3	4

Figure B.2: Using the average coefficient of variation instead of the the $L1$ -regression to analyse the GO annotations, we also find only a few numbers of annotations for the inducer genes and annotations related to DNA damage for the capacitor genes.

Appendix C

Extraction and Analysis of Haploid Genes Modifying Phenotypic Variability and Cell Cycle Regulation in *S. cerevisiae*

1. Accepted Gene Ontology Annotations

GO.ID	GO.Description
GO:0004693	cyclin-dependent protein kinase activity
GO:0000221	vacuolar proton-transporting V-type ATPase
GO:0006283	transcription-coupled nucleotide-excision repair
GO:0006335	DNA replication-dependent nucleosome assembly
GO:0006336	DNA replication-independent nucleosome assembly
GO:0010526	negative regulation of transposition
GO:0015616	DNA translocase activity
GO:0016272	prefoldin complex
GO:0030491	heteroduplex formation
GO:0045002	double-strand break repair via single-strand annealing
GO:0000032	cell wall mannoprotein biosynthetic process
GO:0000372	Group I intron splicing
GO:0006407	rRNA export from nucleus
GO:0032958	inositol phosphate biosynthetic process

Continued...

GO.ID	GO.Description
GO:0000289	nuclear-transcribed mRNA poly(A) tail shortening
GO:0015631	tubulin binding
GO:0015991	ATP hydrolysis coupled proton transport
GO:0032447	protein urmylation
GO:0000220	vacuolar proton-transporting V-type ATPase
GO:0000290	deadenylation-dependent decapping of nuclear-transcribed ...
GO:0000743	nuclear migration involved in conjugation with cellular fusion
GO:0005680	anaphase-promoting complex
GO:0005819	spindle
GO:0006857	oligopeptide transport
GO:0007004	telomere maintenance via telomerase
GO:0016586	RSC complex
GO:0016925	protein sumoylation
GO:0033177	proton-transporting two-sector ATPase complex
GO:0033178	proton-transporting two-sector ATPase complex
GO:0033962	cytoplasmic mRNA processing body assembly
GO:0043140	ATP-dependent 3'-5' DNA helicase activity
GO:0043486	histone exchange
GO:0045143	homologous chromosome segregation
GO:0046982	protein heterodimerization activity
GO:0070478	nuclear-transcribed mRNA catabolic process
GO:0000009	alpha-1
GO:0000023	maltose metabolic process
GO:0000033	alpha-1
GO:0000086	G2/M transition of mitotic cell cycle
GO:0000092	mitotic anaphase B
GO:0000162	tryptophan biosynthetic process
GO:0000727	double-strand break repair via break-induced replication
GO:0000812	Swr1 complex
GO:0000956	nuclear-transcribed mRNA catabolic process
GO:0001080	nitrogen catabolite activation of transcription from RNA ...
GO:0001308	negative regulation of chromatin silencing involved in ...
GO:0005761	mitochondrial ribosome
GO:0006259	DNA metabolic process
GO:0006333	chromatin assembly or disassembly
GO:0006337	nucleosome disassembly
GO:0006356	regulation of transcription from RNA polymerase I promoter
GO:0006401	RNA catabolic process
GO:0006409	tRNA export from nucleus

Continued...

GO.ID	GO.Description
GO:0006474	N-terminal protein amino acid acetylation
GO:0006986	response to unfolded protein
GO:0007021	tubulin complex assembly
GO:0008138	protein tyrosine/serine/threonine phosphatase activity
GO:0019563	glycerol catabolic process
GO:0030490	maturation of SSU-rRNA
GO:0032040	small-subunit processome
GO:0033588	Elongator holoenzyme complex
GO:0034476	U5 snRNA 3'-end processing
GO:0034967	Set3 complex
GO:0042277	peptide binding
GO:0043935	sexual sporulation resulting in formation of a cellular spore
GO:0044267	cellular protein metabolic process
GO:0046034	ATP metabolic process
GO:0046540	U4/U6 x U5 tri-snRNP complex
GO:0048478	replication fork protection
GO:0000096	sulfur amino acid metabolic process
GO:0000133	polarisome
GO:0000151	ubiquitin ligase complex
GO:0000165	MAPK cascade
GO:0000175	3'-5'-exoribonuclease activity
GO:0000196	MAPK cascade involved in cell wall biogenesis
GO:0000447	endonucleolytic cleavage in ITS1 to separate SSU-rRNA ...
GO:0000463	maturation of LSU-rRNA from tricistronic rRNA transcript ...
GO:0000706	meiotic DNA double-strand break processing
GO:0000722	telomere maintenance via recombination
GO:0000742	karyogamy involved in conjugation with cellular fusion
GO:0000755	cytogamy
GO:0003899	DNA-directed RNA polymerase activity
GO:0004003	ATP-dependent DNA helicase activity
GO:0004090	carbonyl reductase (NADPH) activity
GO:0004129	cytochrome-c oxidase activity
GO:0004169	dolichyl-phosphate-mannose-protein mannosyltransferase ...
GO:0004540	ribonuclease activity
GO:0004713	protein tyrosine kinase activity
GO:0004792	thiosulfate sulfurtransferase activity
GO:0005048	signal sequence binding
GO:0005724	nuclear telomeric heterochromatin
GO:0005751	mitochondrial respiratory chain complex IV

Continued...

GO.ID	GO.Description
GO:0005871	kinesin complex
GO:0006000	fructose metabolic process
GO:0006166	purine ribonucleoside salvage
GO:0006260	DNA replication
GO:0006261	DNA-dependent DNA replication
GO:0006270	DNA-dependent DNA replication initiation
GO:0006362	transcription elongation from RNA polymerase I promoter
GO:0006493	protein O-linked glycosylation
GO:0006564	L-serine biosynthetic process
GO:0006796	phosphate-containing compound metabolic process
GO:0006892	post-Golgi vesicle-mediated transport
GO:0007035	vacuolar acidification
GO:0007231	osmosensory signaling pathway
GO:0007266	Rho protein signal transduction
GO:0008104	protein localization
GO:0008134	transcription factor binding
GO:0008156	negative regulation of DNA replication
GO:0008299	isoprenoid biosynthetic process
GO:0009051	pentose-phosphate shunt
GO:0009082	branched chain family amino acid biosynthetic process
GO:0009097	isoleucine biosynthetic process
GO:0009263	deoxyribonucleotide biosynthetic process
GO:0010038	response to metal ion
GO:0010255	glucose mediated signaling pathway
GO:0010696	positive regulation of spindle pole body separation
GO:0015078	hydrogen ion transmembrane transporter activity
GO:0015174	basic amino acid transmembrane transporter activity
GO:0015205	nucleobase transmembrane transporter activity
GO:0015677	copper ion import
GO:0015851	nucleobase transport
GO:0015986	ATP synthesis coupled proton transport
GO:0016818	hydrolase activity
GO:0019904	protein domain specific binding
GO:0030286	dynein complex
GO:0030472	mitotic spindle organization in nucleus
GO:0031011	Ino80 complex
GO:0031298	replication fork protection complex
GO:0031578	mitotic cell cycle spindle orientation checkpoint
GO:0031990	mRNA export from nucleus in response to heat stress

Continued...

GO.ID	GO.Description
GO:0032120	ascospore-type prospore membrane assembly
GO:0032312	regulation of ARF GTPase activity
GO:0032592	integral to mitochondrial membrane
GO:0032865	ERMES complex
GO:0033100	NuA3 histone acetyltransferase complex
GO:0033108	mitochondrial respiratory chain complex assembly
GO:0033263	CORVET complex
GO:0034087	establishment of mitotic sister chromatid cohesion
GO:0034398	telomere tethering at nuclear periphery
GO:0042162	telomeric DNA binding
GO:0042274	ribosomal small subunit biogenesis
GO:0042800	histone methyltransferase activity (H3-K4 specific)
GO:0043130	ubiquitin binding
GO:0043137	DNA replication
GO:0043328	protein targeting to vacuole involved in ubiquitin-dependent ...
GO:0045048	protein insertion into ER membrane
GO:0045835	negative regulation of meiosis
GO:0048188	Set1C/COMPASS complex
GO:0048278	vesicle docking
GO:0051017	actin filament bundle assembly
GO:0051123	RNA polymerase II transcriptional preinitiation complex ...
GO:0060277	negative regulation of transcription involved in G1 phase ...
GO:0070086	ubiquitin-dependent endocytosis
GO:0070772	PAS complex
GO:0072380	TRC complex
GO:0080130	L-phenylalanine:2-oxoglutarate aminotransferase activity
GO:0097034	mitochondrial respiratory chain complex IV biogenesis

Page intentionally left blank

Appendix D

A Primer on Cell Size and Nuclear Size Regulation in *S. cerevisiae* using Bayesian Network Analysis

1. Bootstrapped Networks for the Wild Types

1.1 Strengths and Directions for Stage A

From	To	Strength	Direction
Mother Cell Size	Axis Ratio in Mother	1	0.54
Mother Cell Size	Actin Region Ratio	1	0.61
Mother Cell Size	Nuclear Size in Mother	1	0.89
Axis Ratio in Mother	Actin Region Ratio	0.04	0.5
Axis Ratio in Mother	Nuclear Size in Mother	0.99	0.82
Actin Region Ratio	Axis Ratio in Mother	0.04	0.5
Nuclear Size in Mother	Actin Region Ratio	0.55	0.64
Hip Nuclear	Mother Cell Size	1	0.52
Hip Nuclear	Axis Ratio in Mother	1	0.59
Hip Nuclear	Actin Region Ratio	1	0.66
Mother Cell Center Nuclear	Mother Cell Size	0.11	1
Mother Cell Center Nuclear	Actin Region Ratio	0.26	0.56
Mother Cell Center Nuclear	Nuclear Size in Mother	1	1
Mother Cell Center Nuclear	Hip Nuclear	1	0.53

1.2 Strengths and Directions for Stage A1B

From	To	Strength	Direction
Mother Cell Size	Bud Cell Size	0.78	0.99
Mother Cell Size	Bud Ratio	1	1
Mother Cell Size	Axis Ratio in Bud	0.01	1
Mother Cell Size	Neck Width	1	0.73
Mother Cell Size	Actin Region Ratio	0.22	1
Mother Cell Size	Nuclear Size in Mother	1	1
Mother Cell Size	Nuclear Size in Cell	1	1
Mother Cell Size	Hip Nuclear	0.4	0.95
Bud Cell Size	Bud Ratio	1	0.62
Bud Cell Size	Axis Ratio in Bud	0.72	0.79
Bud Cell Size	Nuclear Size in Mother	0.6	0.96
Bud Cell Size	Nuclear Size in Cell	0.6	1
Bud Cell Size	Nuclear Axis Ratio in Mother	0.83	0.95
Bud Cell Size	Nuclear Axis Ratio in Cell	0.83	1
Bud Cell Size	Hip Nuclear	0.25	0.88
Bud Cell Size	Mother Cell Center Nuclear	0.08	0.88
Bud Cell Size	Neck Nuclear in Mother	0.17	0.65
Bud Ratio	Axis Ratio in Bud	0.98	0.69
Bud Ratio	Actin Region Ratio	0.99	0.96
Bud Ratio	Bud Actin Ratio	0.02	0.5
Bud Ratio	Nuclear Size in Mother	0.09	1
Bud Ratio	Nuclear Size in Cell	0.09	1
Bud Ratio	Nuclear Axis Ratio in Mother	0.21	0.95
Bud Ratio	Nuclear Axis Ratio in Cell	0.21	1
Bud Ratio	Hip Nuclear	0.31	1
Bud Ratio	Mother Cell Center Nuclear	0.8	0.78
Bud Ratio	Neck Nuclear in Mother	0.33	0.7
Axis Ratio in Mother	Mother Cell Size	1	0.6
Axis Ratio in Mother	Bud Ratio	0.02	1
Axis Ratio in Mother	Axis Ratio in Bud	0.08	0.5
Axis Ratio in Mother	Bud Direction	1	0.68
Axis Ratio in Mother	Neck Position	0.56	0.52
Axis Ratio in Mother	Neck Width	0.29	0.67
Axis Ratio in Mother	Actin Region Ratio	0.08	0.69
Axis Ratio in Mother	Bud Actin Ratio	0.18	0.86
Axis Ratio in Mother	Nuclear Size in Mother	0.14	1
Axis Ratio in Mother	Nuclear Size in Cell	0.13	1

Continued...

From	To	Strength	Direction
Axis Ratio in Mother	Neck Nuclear in Mother	0.55	0.87
Axis Ratio in Bud	Axis Ratio in Mother	0.08	0.5
Axis Ratio in Bud	Actin Region Ratio	0.6	0.63
Axis Ratio in Bud	Nuclear Size in Mother	0.3	0.5
Axis Ratio in Bud	Nuclear Size in Cell	0.29	1
Axis Ratio in Bud	Nuclear Axis Ratio in Cell	0.01	1
Axis Ratio in Bud	Hip Nuclear	0.68	0.87
Axis Ratio in Bud	Neck Nuclear in Mother	0.59	0.63
Bud Direction	Mother Cell Size	0.96	0.57
Bud Direction	Bud Cell Size	0.64	0.98
Bud Direction	Axis Ratio in Bud	0.08	1
Bud Direction	Actin Region Ratio	0.5	0.98
Bud Direction	Bud Actin Gravity Point	0.04	1
Bud Direction	Nuclear Size in Mother	0.84	1
Bud Direction	Nuclear Size in Cell	0.84	1
Bud Direction	Hip Nuclear	0.39	0.92
Bud Direction	Mother Cell Center Nuclear	0.09	0.89
Bud Direction	Neck Nuclear in Mother	0.84	0.98
Neck Position	Mother Cell Size	1	0.66
Neck Position	Bud Cell Size	0.01	1
Neck Position	Bud Ratio	0.01	1
Neck Position	Axis Ratio in Bud	0.24	1
Neck Position	Bud Direction	1	0.7
Neck Position	Neck Width	0.01	1
Neck Position	Actin Region Ratio	0.01	1
Neck Position	Bud Actin Gravity Point	0.16	1
Neck Position	Nuclear Size in Mother	0.02	1
Neck Position	Nuclear Size in Cell	0.01	1
Neck Position	Hip Nuclear	1	0.92
Neck Position	Mother Cell Center Nuclear	0.18	0.89
Neck Position	Neck Nuclear in Mother	0.23	0.89
Neck Width	Bud Cell Size	1	0.86
Neck Width	Bud Ratio	0.72	0.91
Neck Width	Axis Ratio in Bud	1	0.9
Neck Width	Bud Direction	0.01	1
Neck Width	Actin Region Ratio	0.46	0.96
Neck Width	Bud Actin Ratio	0.85	0.72
Neck Width	Nuclear Axis Ratio in Mother	0.11	0.73
Neck Width	Nuclear Axis Ratio in Cell	0.24	1

Continued...

Appendix D. A Primer on Cell Size and Nuclear Size Regulation in *S. cerevisiae* using Bayesian Network Analysis

236

From	To	Strength	Direction
Neck Width	Neck Nuclear in Mother	0.55	0.96
Actin Region Ratio	Bud Cell Size	0.07	0.57
Actin Region Ratio	Nuclear Size in Cell	0.89	1
Actin Region Ratio	Nuclear Axis Ratio in Mother	0.02	1
Actin Region Ratio	Nuclear Axis Ratio in Cell	0.03	1
Actin Region Ratio	Mother Cell Center Nuclear	0.02	0.5
Bud Actin Ratio	Bud Cell Size	0.02	1
Bud Actin Ratio	Bud Ratio	0.02	0.5
Bud Actin Ratio	Axis Ratio in Bud	0.56	0.86
Bud Actin Ratio	Actin Region Ratio	1	0.95
Bud Actin Ratio	Bud Actin Gravity Point	1	0.67
Bud Actin Ratio	Nuclear Size in Mother	0.01	1
Bud Actin Ratio	Nuclear Size in Cell	0.04	1
Bud Actin Ratio	Hip Nuclear	0.42	0.74
Bud Actin Ratio	Mother Cell Center Nuclear	0.25	0.68
Bud Actin Ratio	Neck Nuclear in Mother	0.86	0.82
Bud Actin Gravity Point	Bud Cell Size	0.9	0.99
Bud Actin Gravity Point	Bud Ratio	0.96	0.92
Bud Actin Gravity Point	Axis Ratio in Bud	1	0.96
Bud Actin Gravity Point	Neck Width	0.19	0.84
Bud Actin Gravity Point	Nuclear Size in Mother	0.08	0.88
Bud Actin Gravity Point	Nuclear Size in Cell	0.07	1
Bud Actin Gravity Point	Nuclear Axis Ratio in Mother	0.7	0.94
Bud Actin Gravity Point	Nuclear Axis Ratio in Cell	0.64	1
Bud Actin Gravity Point	Hip Nuclear	0.19	0.89
Bud Actin Gravity Point	Mother Cell Center Nuclear	0.02	1
Bud Actin Gravity Point	Neck Nuclear in Mother	0.27	0.61
Nuclear Size in Mother	Axis Ratio in Bud	0.3	0.5
Nuclear Size in Mother	Actin Region Ratio	0.89	0.6
Nuclear Size in Mother	Bud Actin Gravity Point	0.08	0.13
Nuclear Size in Mother	Nuclear Axis Ratio in Cell	1	0.98
Nuclear Size in Mother	Hip Nuclear	0.02	1
Nuclear Axis Ratio in Mother	Nuclear Size in Mother	1	0.94
Nuclear Axis Ratio in Mother	Nuclear Size in Cell	1	1
Hip Nuclear	Axis Ratio in Mother	0.02	1
Hip Nuclear	Mother Cell Center Nuclear	1	0.67
Hip Nuclear	Neck Nuclear in Mother	1	0.81
Mother Cell Center Nuclear	Actin Region Ratio	0.02	0.5
Mother Cell Center Nuclear	Nuclear Size in Mother	1	0.95

Continued...

Appendix D. A Primer on Cell Size and Nuclear Size Regulation in *S. cerevisiae* using Bayesian Network Analysis

237

From	To	Strength	Direction
Mother Cell Center Nuclear	Nuclear Size in Cell	1	1
Mother Cell Center Nuclear	Nuclear Axis Ratio in Mother	0.98	0.73
Mother Cell Center Nuclear	Nuclear Axis Ratio in Cell	0.98	0.99
Neck Nuclear in Mother	Actin Region Ratio	0.99	0.74
Neck Nuclear in Mother	Nuclear Size in Mother	0.32	0.88
Neck Nuclear in Mother	Nuclear Size in Cell	0.33	1
Neck Nuclear in Mother	Nuclear Axis Ratio in Mother	1	0.9
Neck Nuclear in Mother	Nuclear Axis Ratio in Cell	1	0.99
Neck Nuclear in Mother	Mother Cell Center Nuclear	1	0.67

1.2 Strengths and Directions for Stage C

From	To	Strength	Direction
Mother Cell Size	Bud Cell Size	1	0.89
Mother Cell Size	Bud Ratio	1	0.84
Mother Cell Size	Axis Ratio in Mother	0.93	0.63
Mother Cell Size	Bud Direction	0.31	0.66
Mother Cell Size	Neck Width	0.21	1
Mother Cell Size	Bud Actin Gravity Point	0.01	1
Mother Cell Size	Nuclear size in Mother	0.93	0.52
Mother Cell Size	Nuclear size in Bud	0.37	0.99
Mother Cell Size	Nuclear size in Cell	0.5	0.52
Mother Cell Size	Nuclear Axis Ratio in Mother	0.04	1
Mother Cell Size	Hip Nuclear	0.07	1
Mother Cell Size	Mother Cell Center Nuclear	0.03	1
Mother Cell Size	Bud Cell Center Nuclear	0.02	1
Mother Cell Size	Length between Nucleus	0.92	1
Bud Cell Size	Bud Ratio	1	0.7
Bud Cell Size	Axis Ratio in Mother	0.01	1
Bud Cell Size	Neck Width	0.6	0.93
Bud Cell Size	Actin Region Ratio	0.19	0.74
Bud Cell Size	Nuclear size in Mother	0.2	0.5
Bud Cell Size	Nuclear size in Bud	0.32	0.83
Bud Cell Size	Nuclear size in Cell	0.52	0.98
Bud Cell Size	Mother Cell Center Nuclear	0.01	1
Bud Cell Size	Bud Top Nuclear	0.14	0.86
Bud Cell Size	Neck Nuclear in Bud	0.06	1
Bud Cell Size	Length between Nucleus	0.73	1
Bud Ratio	Axis Ratio in Bud	0.07	0.86
Bud Ratio	Neck Width	0.17	0.88
Bud Ratio	Actin Region Ratio	0.05	0.6
Bud Ratio	Bud Actin Ratio	0.06	1
Bud Ratio	Bud Actin Gravity Point	0.38	0.74
Bud Ratio	Nuclear size in Bud	0.7	0.94
Bud Ratio	Nuclear Axis Ratio in Mother	1	0.75
Bud Ratio	Nuclear Axis Ratio in Bud	0.22	0.86
Bud Ratio	Mother Cell Center Nuclear	0.44	0.89
Bud Ratio	Neck Nuclear in Mother	0.32	0.73
Bud Ratio	Length between Nucleus	0.01	1
Axis Ratio in Mother	Axis Ratio in Bud	0.53	0.5

Continued...

From	To	Strength	Direction
Axis Ratio in Mother	Bud Direction	1	0.96
Axis Ratio in Mother	Neck Position	0.08	0.81
Axis Ratio in Mother	Neck Width	0.07	0.86
Axis Ratio in Mother	Actin Region Ratio	0.53	0.94
Axis Ratio in Mother	Bud Actin Ratio	0.03	1
Axis Ratio in Mother	Nuclear Axis Ratio in Mother	0.1	1
Axis Ratio in Mother	Neck Nuclear in Mother	0.45	0.77
Axis Ratio in Bud	Bud Cell Size	0.05	0.8
Axis Ratio in Bud	Axis Ratio in Mother	0.53	0.5
Axis Ratio in Bud	Bud Direction	0.8	0.96
Axis Ratio in Bud	Neck Position	0.1	0.7
Axis Ratio in Bud	Neck Width	1	0.93
Axis Ratio in Bud	Actin Region Ratio	0.8	0.81
Axis Ratio in Bud	Bud Actin Ratio	1	0.8
Axis Ratio in Bud	Bud Actin Gravity Point	0.09	1
Axis Ratio in Bud	Nuclear size in Mother	0.28	0.82
Axis Ratio in Bud	Nuclear size in Bud	0.02	1
Axis Ratio in Bud	Nuclear size in Cell	0.02	1
Axis Ratio in Bud	Nuclear Axis Ratio in Mother	0.03	1
Axis Ratio in Bud	Hip Nuclear	0.01	1
Axis Ratio in Bud	Neck Nuclear in Mother	0.07	0.71
Axis Ratio in Bud	Bud Top Nuclear	0.11	0.77
Axis Ratio in Bud	Bud Cell Center Nuclear	0.39	0.94
Axis Ratio in Bud	Length between Nucleus	1	1
Bud Direction	Bud Cell Size	0.63	0.76
Bud Direction	Bud Ratio	0.32	0.81
Bud Direction	Neck Width	0.86	0.71
Bud Direction	Bud Actin Ratio	0.01	1
Bud Direction	Bud Actin Gravity Point	0.02	1
Bud Direction	Nuclear size in Mother	0.02	1
Bud Direction	Nuclear size in Bud	0.13	0.85
Bud Direction	Nuclear size in Cell	0.06	0.5
Bud Direction	Nuclear Axis Ratio in Mother	0.01	1
Bud Direction	Nuclear Axis Ratio in Bud	0.26	0.88
Bud Direction	Hip Nuclear	0.36	0.61
Bud Direction	Neck Nuclear in Mother	0.01	1
Bud Direction	Bud Top Nuclear	0.07	0.71
Bud Direction	Bud Cell Center Nuclear	0.31	0.81
Neck Position	Mother Cell Size	1	0.75

Continued...

Appendix D. A Primer on Cell Size and Nuclear Size Regulation in *S. cerevisiae* using Bayesian Network Analysis

240

From	To	Strength	Direction
Neck Position	Bud Cell Size	0.67	0.92
Neck Position	Bud Ratio	0.8	0.93
Neck Position	Bud Direction	1	0.92
Neck Position	Neck Width	0.21	0.79
Neck Position	Bud Actin Gravity Point	0.68	1
Neck Position	Nuclear size in Bud	0.12	1
Neck Position	Nuclear size in Cell	0.03	0.67
Neck Position	Hip Nuclear	0.8	0.99
Neck Position	Neck Nuclear in Mother	0.72	0.88
Neck Position	Length between Nucleus	0.52	1
Neck Width	Bud Actin Gravity Point	0.04	0.75
Neck Width	Nuclear size in Mother	0.04	0.75
Neck Width	Nuclear size in Bud	0.33	0.85
Neck Width	Length between Nucleus	0.23	0.78
Actin Region Ratio	Neck Width	0.88	0.81
Actin Region Ratio	Bud Actin Ratio	0.11	0.55
Actin Region Ratio	Bud Actin Gravity Point	1	0.93
Actin Region Ratio	Nuclear size in Bud	0.03	0.67
Actin Region Ratio	Nuclear Axis Ratio in Mother	0.84	0.77
Bud Actin Ratio	Bud Cell Size	0.31	0.97
Bud Actin Ratio	Neck Width	0.93	0.76
Bud Actin Ratio	Bud Actin Gravity Point	1	0.94
Bud Actin Ratio	Nuclear Axis Ratio in Bud	0.96	0.83
Bud Actin Ratio	Hip Nuclear	0.96	0.95
Bud Actin Ratio	Bud Top Nuclear	0.21	0.81
Bud Actin Gravity Point	Bud Cell Size	0.01	1
Bud Actin Gravity Point	Actin Region Ratio	1	0.07
Bud Actin Gravity Point	Nuclear Axis Ratio in Mother	0.73	0.89
Bud Actin Gravity Point	Nuclear Axis Ratio in Bud	0.03	1
Bud Actin Gravity Point	Hip Nuclear	0.01	1
Bud Actin Gravity Point	Length between Nucleus	0.18	0.72
Nuclear size in Mother	Bud Cell Size	0.2	0.5
Nuclear size in Mother	Bud Ratio	0.13	0.73
Nuclear size in Mother	Actin Region Ratio	0.21	0.88
Nuclear size in Mother	Nuclear size in Cell	0.88	0.95
Nuclear size in Mother	Nuclear Axis Ratio in Mother	1	0.9
Nuclear size in Mother	Nuclear Axis Ratio in Bud	0.31	1
Nuclear size in Mother	Hip Nuclear	0.56	0.63
Nuclear size in Mother	Neck Nuclear in Mother	0.87	0.57

Continued...

From	To	Strength	Direction
Nuclear size in Mother	Bud Top Nuclear	0.47	0.81
Nuclear size in Mother	Neck Nuclear in Bud	0.36	0.63
Nuclear size in Mother	Length between Nucleus	0.68	0.63
Nuclear size in Bud	Axis Ratio in Mother	0.19	0.55
Nuclear size in Bud	Nuclear size in Mother	0.12	0.63
Nuclear size in Bud	Nuclear Axis Ratio in Bud	0.09	0.89
Nuclear size in Cell	Bud Ratio	0.17	0.59
Nuclear size in Cell	Axis Ratio in Mother	0.03	0.67
Nuclear size in Cell	Bud Direction	0.06	0.5
Nuclear size in Cell	Neck Width	0.09	0.56
Nuclear size in Cell	Nuclear size in Mother	0.88	0.05
Nuclear size in Cell	Nuclear size in Bud	1	0.88
Nuclear size in Cell	Nuclear Axis Ratio in Mother	0.84	0.58
Nuclear size in Cell	Nuclear Axis Ratio in Bud	0.85	0.76
Nuclear Axis Ratio in Mother	Bud Cell Size	0.25	0.68
Nuclear Axis Ratio in Mother	Bud Actin Ratio	0.35	0.69
Nuclear Axis Ratio in Mother	Nuclear size in Bud	0.05	1
Nuclear Axis Ratio in Mother	Nuclear Axis Ratio in Bud	1	0.97
Nuclear Axis Ratio in Bud	Actin Region Ratio	0.01	1
Hip Nuclear	Neck Width	0.24	0.63
Hip Nuclear	Nuclear Axis Ratio in Mother	0.01	1
Hip Nuclear	Length between Nucleus	0.22	0.82
Mother Cell Center Nuclear	Bud Direction	0.25	0.52
Mother Cell Center Nuclear	Neck Width	0.17	0.85
Mother Cell Center Nuclear	Actin Region Ratio	0.74	0.89
Mother Cell Center Nuclear	Bud Actin Ratio	0.36	0.78
Mother Cell Center Nuclear	Bud Actin Gravity Point	0.99	0.97
Mother Cell Center Nuclear	Nuclear size in Mother	0.93	0.58
Mother Cell Center Nuclear	Nuclear size in Bud	0.25	0.88
Mother Cell Center Nuclear	Nuclear size in Cell	0.83	0.83
Mother Cell Center Nuclear	Hip Nuclear	1	0.97
Mother Cell Center Nuclear	Neck Nuclear in Mother	1	0.67
Mother Cell Center Nuclear	Bud Top Nuclear	0.88	0.82
Mother Cell Center Nuclear	Bud Cell Center Nuclear	0.01	1
Mother Cell Center Nuclear	Length between Nucleus	0.26	0.92
Neck Nuclear in Mother	Mother Cell Size	0.09	0.67
Neck Nuclear in Mother	Bud Cell Size	0.05	0.8
Neck Nuclear in Mother	Neck Width	0.61	0.95
Neck Nuclear in Mother	Actin Region Ratio	0.35	0.74

Continued...

From	To	Strength	Direction
Neck Nuclear in Mother	Bud Actin Gravity Point	0.78	0.94
Neck Nuclear in Mother	Nuclear size in Bud	0.01	1
Neck Nuclear in Mother	Nuclear size in Cell	0.83	0.88
Neck Nuclear in Mother	Nuclear Axis Ratio in Mother	0.3	1
Neck Nuclear in Mother	Nuclear Axis Ratio in Bud	0.16	1
Neck Nuclear in Mother	Hip Nuclear	1	0.91
Neck Nuclear in Mother	Bud Top Nuclear	0.98	0.68
Neck Nuclear in Mother	Bud Cell Center Nuclear	0.67	0.93
Neck Nuclear in Mother	Neck Nuclear in Bud	0.18	0.83
Neck Nuclear in Mother	Length between Nucleus	1	1
Bud Top Nuclear	Neck Width	0.44	0.8
Bud Top Nuclear	Bud Actin Gravity Point	0.08	0.63
Bud Top Nuclear	Nuclear size in Bud	0.02	1
Bud Top Nuclear	Nuclear size in Cell	0.17	0.56
Bud Top Nuclear	Nuclear Axis Ratio in Mother	0.27	0.94
Bud Top Nuclear	Nuclear Axis Ratio in Bud	0.32	1
Bud Top Nuclear	Hip Nuclear	0.18	0.67
Bud Top Nuclear	Bud Cell Center Nuclear	0.45	0.67
Bud Top Nuclear	Neck Nuclear in Bud	1	0.88
Bud Top Nuclear	Length between Nucleus	0.47	1
Bud Cell Center Nuclear	Actin Region Ratio	1	0.53
Bud Cell Center Nuclear	Bud Actin Gravity Point	1	0.92
Bud Cell Center Nuclear	Nuclear size in Bud	0.3	0.93
Bud Cell Center Nuclear	Nuclear size in Cell	0.82	0.83
Bud Cell Center Nuclear	Nuclear Axis Ratio in Mother	0.99	0.94
Bud Cell Center Nuclear	Nuclear Axis Ratio in Bud	0.06	1
Bud Cell Center Nuclear	Hip Nuclear	0.68	0.66
Bud Cell Center Nuclear	Neck Nuclear in Bud	1	0.75
Bud Cell Center Nuclear	Length between Nucleus	0.5	0.76
Neck Nuclear in Bud	Bud Direction	0.15	0.8
Neck Nuclear in Bud	Neck Width	0.13	0.85
Neck Nuclear in Bud	Bud Actin Ratio	0.54	0.9
Neck Nuclear in Bud	Nuclear size in Bud	0.04	1
Neck Nuclear in Bud	Nuclear size in Cell	0.08	0.88
Neck Nuclear in Bud	Nuclear Axis Ratio in Mother	1	0.99
Neck Nuclear in Bud	Nuclear Axis Ratio in Bud	0.6	1
Neck Nuclear in Bud	Hip Nuclear	0.81	0.98
Neck Nuclear in Bud	Mother Cell Center Nuclear	0.16	0.63
Neck Nuclear in Bud	Length between Nucleus	1	1

Continued...

Appendix D. A Primer on Cell Size and Nuclear Size Regulation in *S. cerevisiae* using Bayesian Network Analysis

243

From	To	Strength	Direction
Length between Nucleus	Bud Actin Ratio	0.01	1
Length between Nucleus	Nuclear size in Bud	0.72	0.94
Length between Nucleus	Nuclear size in Cell	0.52	0.63
Length between Nucleus	Nuclear Axis Ratio in Mother	0.9	0.81
Length between Nucleus	Nuclear Axis Ratio in Bud	0.68	0.98

Page intentionally left blank

Appendix E

Bacterial Population Growth and Variability in Cell Size and Division Process

Fit of the Mother Machine Data

The mother machine data collected on the MG1655 strain can be fitted with log-normal distributions for the mother and daughter lengths. For an overview of the other strains, we encourage the reader to have a look at the supporting material available in [144].

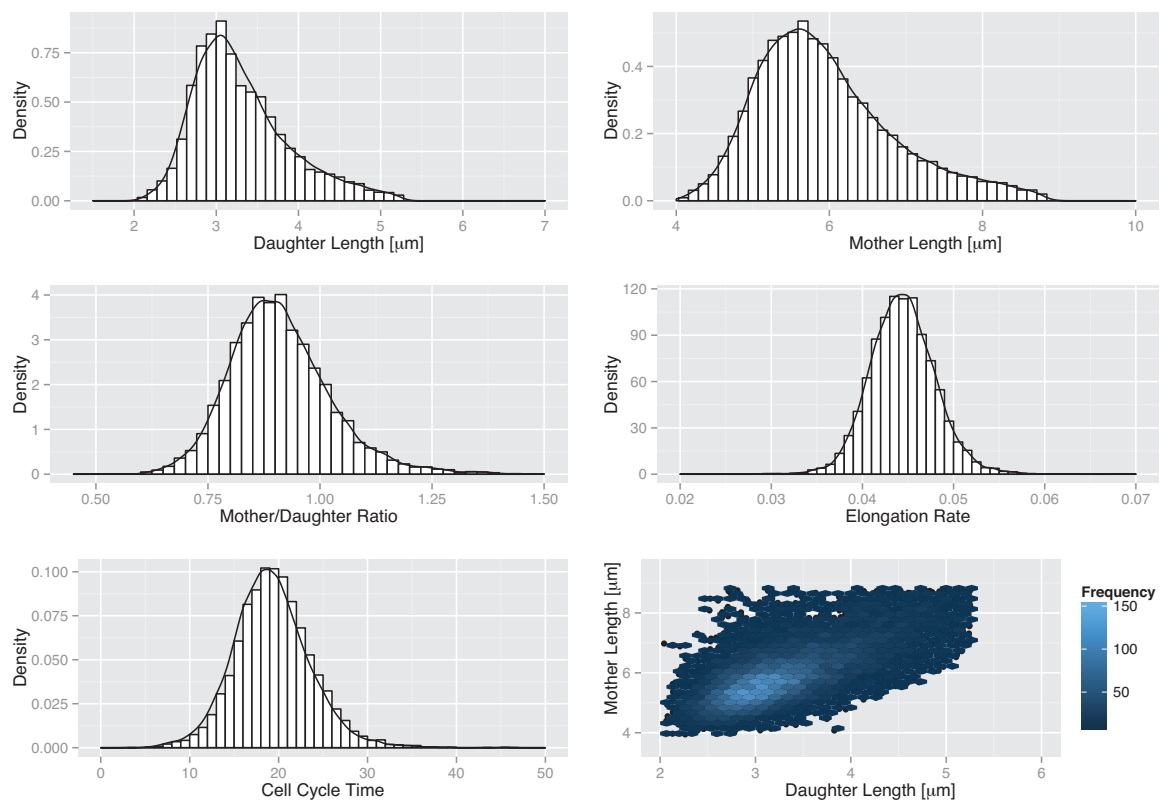


Figure E.1: Fit of the mother machine data for the MG1655 strain. For the mother and daughter lengths, a log-normal distribution can be used to fit the data.