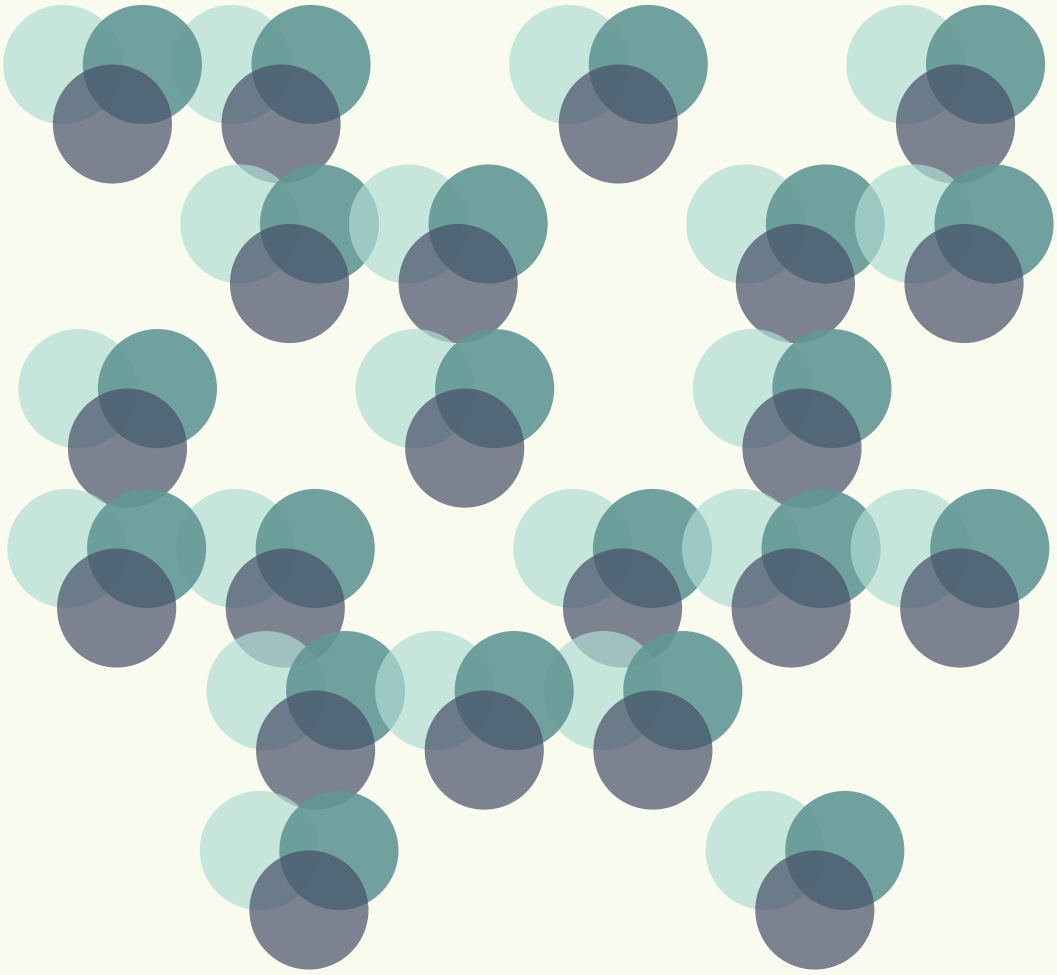


Cell-to-cell variability and its consequences for cellular communities

E.M. Keizer

Propositions

1. Gene expression variability contains a wealth of information. (this thesis)
2. Determining whether observed heterogeneity has functional significance requires both a good understanding of the molecular mechanism and a framework for quantifying heterogeneity. (this thesis)
3. The lack of parity for female research participants in clinical trials and exercise science studies leads to harmful misinformation.
4. Improving scientific literacy across society is a prerequisite for delivering on the promise that science offers for an improved quality of life of current and future generations.
5. Due to its abstract nature, empathy is key for successfully teaching mathematics.
6. Closing the gender gap in the workplace is a non-zero-sum game, and to aid this, paid parental leave serves as an important tool for parents to derive status from childcare.
7. The bicycle represents the efficiencies and benefits of using less of everything.

Propositions belonging to the thesis, entitled

Cell-to-cell variability and its consequences for cellular communities

Emma Keizer
Wageningen, 10 May 2022

Cell-to-cell variability

and its consequences for cellular communities

Emma Mathilde Keizer

Thesis committee

Promotors
Prof. Dr J. Molenaar
Professor of Applied Mathematics
Wageningen University & Research

Prof. Dr Vitor A.P. Martins dos Santos
Personal chair of Biomanufacturing
Wageningen University & Research

Co-promotor
Dr C. Fleck
Spatial Systems Biology Group, University of Freiburg, Germany

Other members
Prof. Dr E. van der Linden, Wageningen University & Research
Dr J.C. Hohlbein, Wageningen University & Research
Prof. Dr F. Bruggeman, Vrije Universiteit Amsterdam
Prof. Dr P.R. ten Wolde, Universiteit van Amsterdam

This research was conducted under the auspices of the Graduate School VLAG (Advanced studies in Food Technology, Agrobiotechnology, Nutrition and Health Sciences).

Cell-to-cell variability
and its consequences for cellular communities

Emma Mathilde Keizer

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 10 May 2022 at 4 p.m. in the Omnia Auditorium

Emma Mathilde Keizer
Cell-to-cell variability and its consequences for cellular communities
261 pages

PhD thesis, Wageningen University, Wageningen, the Netherlands (2022)
With references, with summary in English

ISBN: 978-94-6447-125-0
DOI: https://doi.org/10.18174/565337

Table of contents

Chapter 1

General introduction

The cell, which forms the basic unit of life on earth,[1,2] is a noisy place. Fundamental processes such as growth, replication, metabolism, and signalling depend on the precise coordination of regulatory processes. Yet, nearly all biomolecular interactions ensue from the random collisions of individual molecules, which results in fluctuating molecule concentrations. While cells can exist on their own, often they form part of a cellular community or of the larger structures that make up higher-level organisms, which requires complex organisation and cell-to-cell communication. As of yet, we have a limited understanding of how robustness in development can be guaranteed in spite of this unavoidable stochasticity, and how cellular functioning is maintained in the face of a dynamic environment.

The significant variability that exists between cells in genetically identical populations, even in homogeneous environments, has been observed for a long time. In 1945, Delbrück showed that the burst size of bacteriophages in infected *Escherichia coli* cells varied more than could be accounted for by variations in the size of the bacteria alone, and proposed stochasticity in the viral reproduction process as a possible explanation.[3] However, during this time it was not possible to study this phenomenon at the molecular level due to the lack of appropriate techniques. Despite the inability to directly measure variability in experiments, new mathematical approaches were developed to investigate the degree to which stochasticity in gene expression can lead to phenotypic variations within cell populations.[4] The properties of these gene expression models were analysed using the procedure described by Gillespie, who formulated an approach to simulate the time-evolution of stochastic biochemical systems,[5] and showed that protein concentrations can fluctuate substantially between cells. The theoretical work laid out by McAdams *et al.* was followed up by Arkin *et al.*, who modelled the lysis-lysogeny switch in bacteria infected with the λ phage in order to explain how expression noise can drive genetically identical cells to different fates. Using their stochastic model, they predicted the fraction of the population that selects the lysogenic pathway, a probabilistic outcome that could not be explained using deterministic analysis.[6] The study of gene expression noise really became mainstream with the publication of an article by Elowitz *et al.*, in which they engineered *Escherichia coli* to express yellow (YFP) and cyan fluorescent proteins (CFP) under the control of identical promoters in the same cell.[7] In this by now classic paper, it was shown that expression of a gene inside a single cell could result in significant variation in protein levels, and that this stochasticity could be attributed to different sources.

1.1 Non-genetic origins of cell-to-cell variability

The dual reporter set-up developed by Elowitz *et al.* allows gene expression noise to be decomposed into two different components: *intrinsic* and *extrinsic* noise. Intrinsic noise is the variability in protein levels that arises from the inherent randomness of

biochemical reactions involved in the processes of transcription, translation, and the degradation of mRNA and proteins. Intrinsic noise affects each copy of the gene independently, and thus results in uncorrelated fluctuations in the levels of CFP and YFP. Its magnitude is determined by the structure, reaction rates, and concentrations of molecular species of the underlying reaction network. In contrast, extrinsic noise arises from unobserved variation in cellular processes and components, such as cell size, cell cycle, temperature, concentrations of transcriptional, translational, or metabolic components. These upstream influences affect multiple genes in the cell in the same way.[7–11] Correlations in the fluctuations in the levels of the CFP and YFP reporters then reflect the common environment. A complicating factor is that these sources of variability are not always independent of each other, and the unobserved processes that generate extrinsic noise might have both stochastic and deterministic components.[12,13]

The emergence of techniques such as time-lapse microscopy and single-cell fluorescence tracking has drawn attention to another important source of noise: population dynamics. During balanced growth, on average molecule numbers double between cell birth and division. Across a population, this results in variability in molecular copy number due to heterogeneity in cell age, which is independent of the mean copy number of a molecule.[14] In addition, the interdivision times of *e.g.* bacterial and yeast cells can also exhibit substantial variability,[15,16] causing some cells to have longer generation times than others and thus possess more molecules at the end of their cycle. Upon cell division the cellular contents are randomly partitioned to each daughter cell, which further contributes to intrinsic noise,[17] especially when molecule numbers are low. Next to the variability that originates from the random partitioning of molecules, at cell division the volume of the mother cell is not always symmetrically divided across the two daughter cells. Several attempts at developing stochastic models that incorporate one or more aspects of population dynamics and cell cycle effects into the analysis of gene expression noise have been made, using analytical approaches[14,17–22] or with the use of agent-based simulation models (ABM).[17,20,21,23,24] Yet, despite these efforts, identifying noise sources from experimental data and analysing their respective effects remains challenging.[25]

1.2 Variability across the biological domain

The mechanisms that cause variability permeate biology at all levels. In this thesis, I will consider a number of biological systems to investigate how noise affects functioning. At the sub-cellular level, noise impacts the dynamics of gene expression. According to the central dogma of molecular biology, proteins are produced within cells in two steps: genes are transcribed to synthesise messenger RNA (mRNA), which is then translated to produce proteins. A number of studies have measured

variability in mRNA and protein expression levels for various genes,[26–28] and noise in expression levels has been found to be strongly connected with the mechanism underlying gene regulation.[29] Certain network architectures might have been selected during evolution for their ability to either attenuate or amplify noise. For example, negative feedback loops have been associated with noise reduction,[30, 31] which could be beneficial for the cell when the network needs to be robust to extrinsic noise[32] or is dependent on the reliable processing of molecular signals.[33] In contrast, other gene regulatory networks such as positive feedback loops have been shown to exploit noise. This noise amplification increases population heterogeneity resulting in a possible fitness advantage for the population,[31] and could be required for cellular differentiation.[6, 34] Hence, studying network architectures and characterising their sources of variability can lead to a better understanding of how cells exploit or suppress environmental signals. In addition, this could aid the construction of robust synthetic gene circuits that are able to maintain function in fluctuating environments.

Single-cell variability leads to phenotypic diversity within populations. This might not always benefit individual cells, however it could enable strategies which allow the population to survive in the face of a changing environment. For example, a variable response to stimuli has been shown to increase the adaptability of bacteria to changes in the environment in the context of antibiotic resistance,[35] or to contribute to the development of drug-resistant cancer cells.[36] In this thesis, we will study the cell-to-cell variability of CRISPR-Cas, the adaptive immune system of *Escherichia coli* against invading bacteriophages and other mobile genetic elements (MGEs). The CRISPR-Cas system has two components: Clustered Regulatory Interspaced Short Palindromic Repeats (CRISPR) arrays which represent the immunological memory of past infections, and CRISPR-associated (Cas) proteins which carry out the immune functions of acquiring new memories and destroying invader DNA. In a population of bacteria equipped with CRISPR-Cas, gene expression noise might result in a fraction of cells that have higher levels of Cas proteins and are thus better suited to acquire immunity against MGEs. On the other hand, elevated expression levels are associated with a higher metabolic load, or could increase the probability of acquiring memories that target the cell's own DNA resulting in death. This could result in a bet-hedging scenario where some cell lineages have increased ability to combat environmental conditions, such as subsequent infections, whereas others can invest more energy in reproduction and have a lower chance of self-targeting.

While gene expression noise has been extensively studied in single-celled organisms, its role in the development and functioning of multicellular systems is less clear due to challenges associated with data analysis and mathematical modelling of tissue-bound cells.[37] In a tissue, cells are coupled through processes such as diffusion, active transport, or cell-to-cell signalling, and the population is thus spatially inhomogeneous. Developmental processes such as tissue formation, which

depend on the stochastic interactions between molecules, require a high level of temporal and spatial coordination. Uncontrolled variability at the molecular and cellular level has been shown to disrupt reliable cell differentiation and functioning. For example, non-genetic heterogeneity has been suggested to play a role in tumour progression.[38,39] On the other hand, there is evidence that suggests that the initiation of patterning mechanisms, such as the Notch-Delta signalling pathway, are dependent on this stochasticity.[40]

In this thesis, we study variability in gene expression during development in the context of leaf growth in *Arabidopsis thaliana*. In growing tissues, gene expression is not in stationary state due to cell growth and division. This might result in different contributions of intrinsic and extrinsic noise to gene expression as compared to mature leaves. In addition, adjacent cells in plant leaves are connected by plasmodesmata which leads to spatial coupling. It is unknown to which extent this connectivity affects gene expression in neighbouring cells.

Addressing the many unknowns surrounding the manifestation of gene expression noise in these biological systems requires the interplay between experimental and theoretical approaches. Throughout this thesis, we make use of single-cell measurements from bacteria, mammalian cells, and plants. However, the quantification of gene expression at this detailed level is limited by experimental challenges. In single cells, the number of molecular species that can be measured simultaneously with the use of fluorescent reporters is limited, and as a result often not all processes of interest can be observed. While promising, live-imaging microscopy approaches in bacterial populations and multi-cellular organisms are challenging due to the need for high-resolution images and the development of accurate cell segmentation software.[37] Additionally, in multi-cellular structures it is difficult to ensure homogeneous environmental conditions. Once experimental data are collected, the next task is to try to disentangle the intrinsic and extrinsic contributions to the observed cell-to-cell variability. Mathematical modelling is an indispensable tool to analyse and interpret data, and can be used to test predictions through theoretical analysis or computer simulations. We will now discuss approaches that have been developed for this purpose, and are used throughout this thesis.

1.3 Mathematical and computational approaches to modelling stochasticity

1.3.1 The chemical master equation

When molecule numbers are low, the dynamics of many cellular processes can no longer be accurately described using macroscopic-scale, deterministic models that assume large population numbers. However, to study these processes at the microscopic level requires keeping track of the positions and momenta of all particles

involved, which is usually too computationally expensive. To avoid these problems, we will make use of mesoscopic modelling, which allows us to accurately describe the stochastic nature of biological processes and the discreteness of the population numbers, without the computational burden associated with molecular dynamics. In this approach, the stochastic dynamics of point-particle biochemical systems in well-mixed conditions is described by the chemical master equation (CME),[41] which we will briefly outline next.

We consider a chemical network consisting of N molecular species with particle numbers $X_i(t)$ in a (constant) system volume Ω resulting in species concentrations $x_i(t) = X_i(t)/\Omega$. Depending on its biological origin, extrinsic noise will affect the rate of one or more reactions, which is mathematically represented via the introduction of M extra stochastic variables $\eta_k(t)$. The microscopic rate functions $\tilde{f}_j(\boldsymbol{x}, \boldsymbol{\eta}, \Omega)$ and $[N \times R]$ stoichiometric matrix $\boldsymbol{S}$, where R is equal to the number of possible reactions, together describe the chemical reactions that can take place in the reaction volume. Here, $\boldsymbol{x} = (x_1(t), x_2(t), ..., x_N(t))$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_M)$. The probability that the network is in a state with particle numbers $\boldsymbol{X}$ with stochastic variables $\boldsymbol{\eta}$ at time t is given by $P(\boldsymbol{X}, \boldsymbol{\eta}, t)$, and $\Omega P(\boldsymbol{X}, \boldsymbol{\eta}, t)dt$ is the probability that a reaction of type j occurs in the next infinitesimally small time step dt. The system's state after the reaction is defined by particle numbers $\tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{S}_j$ and the value of $\boldsymbol{\eta}$ at that time. Here, $\boldsymbol{S}_j = (S_{1j}, ..., S_{Nj})$, where S_{ij} is the change in the number of the i^{th} molecular species caused by one occurrence of the j^{th} reaction. The CME describing the mesoscopic system dynamics is obtained by taking the limit $dt \to 0$ and reads as:

$$\frac{\mathrm{d}P(\boldsymbol{X}, \boldsymbol{\eta}, t)}{\mathrm{d}t} = \Omega \sum_{j=1}^{R} \left(\prod_{i=1}^{N} E_i^{-S_{ij}} - 1 \right) \tilde{f}_j(\boldsymbol{x}, \boldsymbol{\eta}, \Omega) P(\boldsymbol{X}, \boldsymbol{\eta}, t) \qquad (1.1)$$

where E_i is a step operator $E_i^q g(X_1, ..., X_i, ..., X_N) = g(X_1, ..., X_{i+q}, ..., X_N)$ on a function g of the particle numbers $\boldsymbol{X}$, the product of which allows to take into account all system states that can evolve to the state given by $\boldsymbol{X}$ and $\boldsymbol{\eta}$. The probability flux away from state $\boldsymbol{X}, \boldsymbol{\eta}$ due to reaction j is given by $-\sum_{j=1}^{R} \tilde{f}_j(\boldsymbol{x}, \boldsymbol{\eta}, \Omega) P(\boldsymbol{X}, \boldsymbol{\eta}, t)$.

Given that the system is well-stirred and in thermal equilibrium, the CME gives an exact description of the system dynamics. Solving Eq. 1.1 in principle yields the probability $P(\boldsymbol{X}, \boldsymbol{\eta}, t)$, which fully characterises the system dynamics. For example, once this probability is known, one may calculate all moments of $\boldsymbol{X}$. However, even in the absence of extrinsic noise, Eq. 1.1 can rarely be solved exactly except for a handful of rather unrealistic toy problems. The analysis is further complicated when extrinsic noise is introduced.[42,43] Since solving the CME directly is seldom an option, there are two possible ways to proceed: stochastic simulation, or analytic

approximation techniques, which we discuss next.

1.3.2 Stochastic simulation

One way to study the dynamics of a biochemical reaction system is through numerical simulation of the system dynamics. Instead of obtaining the probability density function of the system, $P(\boldsymbol{X}, t|\boldsymbol{X}_0, t_0)$, an exact realisation $\boldsymbol{X}(t)$ of the system dynamics is generated. Since the system is noisy, one cannot calculate the system dynamics in a unique way, as would be possible in case of a deterministic system. Instead, every time one simulates the dynamics, a particular realisation of the stochastic noise is obtained. Although each realisation of the model by the SSA will produce a different result, in the limit of a sufficiently large number of samples these together will yield a reliable approximation of the probability $P(\boldsymbol{X}, \boldsymbol{\eta}, t)$.

The most important method for numerical simulation of the CME, termed the stochastic simulation algorithm (SSA), was formulated by Gillespie[5,44] and is used in **Chapter 5**. The principle of the 'direct method' formulation of the SSA is that it explicitly simulates each reaction event by sampling $i)$ the time until the next reaction τ and $ii)$ the type of reaction j, from the reaction probability density function

$$P(\tau, j|\boldsymbol{X}(t), t) = a_j \exp(-a_0 \tau), \tag{1.2}$$

where a_j are the propensity functions of an elementary reaction of type *j*, which depends on the current state of the system and the reaction rate constant. Adding the propensities of all reactions in the system gives us the propensity sum a_0:

$$a_0 = \sum_{j=1}^{R} a_j. \tag{1.3}$$

$P(\tau, j)\,\mathrm{d}\tau$ is then the probability that within the next time interval $[t+\tau, t+\tau+\mathrm{d}\tau]$, a reaction will occur and will be of type j, which depends on the state of the system $\boldsymbol{X}(t)$. As Eq. 1.2 states, it is assumed the system is memoryless and reaction events are thus exponentially distributed. The steps to iteratively simulate the system dynamics can then be summarised as follows:

0. Initialise time $t = 0$ and state $\boldsymbol{X} = \boldsymbol{X}_0$.
1. Compute reaction propensities a_j, and the propensity sum a_0.
2. Draw two random integers r_1 and r_2 from $\mathrm{Uniform}[0, 1]$. Generate $\tau = \frac{1}{a_0}\log(\frac{1}{r_1})$ and select the next reaction index j such that it is the smallest index to satisfy $\sum_{k=1}^{j} a_k(\boldsymbol{X}) \geq r_2 a_0$.
3. Effect the reaction by updating the state of the system according to the stoichiometry of the reaction, and advance time t by τ.

A key requirement for the validity of the algorithm, and the CME in general, is that the system is well mixed, meaning reactant molecules are randomly and uniformly distributed throughout the reaction volume. Though certainly not a universal property of biological systems, in some parts of the cell this required spatial homogeneity is indeed achieved through the natural diffusive motions of molecules.

Extending the SSA to accommodate fluctuating environments

The SSA provides trajectories which are consistent with the underlying CME, provided the rate constants do not fluctuate between reaction occurrences. This means that it can only describe intrinsic noise or static extrinsic noise, but not cases where extrinsic noise manifests as fluctuations in the rate constants, as is the case in **Chapter 2**. Additionally, when dealing with a growing cell populations as is the case in **Chapters 3–4**, the cellular volume will not be constant between reaction events. As the propensity function of a bimolecular reaction depends inversely on the system volume, this results in non-constant propensity functions. Modifications to the SSA have been devised that are able to take into account systems where the reaction propensities become time-varying under the influence of extrinsic processes.[45–47] In **Chapter 2**, **Chapter 3**, and **Chapter 4**, I make use of the *Extrande* extension to the classical stochastic simulation algorithm proposed by Voliotis *et al.*[47] which allows for exact and computationally efficient simulation of a biochemical network subject to extrinsic fluctuations. The principle of the *Extrande* algorithm is to add a virtual reaction to the system that leaves molecule numbers unchanged, but whose propensity is designed to fluctuate over time. This ensures that the extended system fulfils the requirement of the SSA that the sum of propensities remains constant between reaction events. The augmented system is then simulated over time and the reaction events of the virtual reaction channel are discarded, resulting in a trajectory of the original system.

1.3.3 Analytical approximation techniques

Stochastic simulations are computationally expensive. For this reason, various systematic approximation schemes have been derived to obtain analytical results, of which the most well known is van Kampen's system size expansion.[48] This systematic expansion of the CME separates the macroscopic dynamics of the system given by a set of ordinary differential equations (ODEs) from the microscopic fluctuations around it, and has been shown to approximate the CME well for a wide range of systems.[49] The idea behind this approach is as follows.

In the limit of a sufficiently large system volume ($\Omega \to \infty$), the width of the probability distribution of a monostable system is expected to scale with the system volume as $\Omega^{1/2}$, and so the concentrations of the molecular species can be written

as:[48]

$$\boldsymbol{x} = \boldsymbol{\phi} + \Omega^{-1/2}\boldsymbol{\xi}, \tag{1.4}$$

where $\boldsymbol{\phi}$ and $\boldsymbol{f}(\boldsymbol{\phi}, \boldsymbol{\eta})$ are vectors of the macroscopic concentrations and macroscopic reaction rates, respectively, and the new variable $\boldsymbol{\xi}$ denotes the microscopic (Gaussian) fluctuations. The probability distribution can then be rewritten as

$$\lim_{\Omega\to\infty} P(\boldsymbol{X}, \boldsymbol{\eta}, t) = P(\Omega\boldsymbol{\phi} + \Omega^{1/2}\boldsymbol{\xi}, \boldsymbol{\eta}, t) = \Pi(\boldsymbol{\xi}, \boldsymbol{\eta}, t). \tag{1.5}$$

As a next step, the CME is expanded in powers of $\Omega^{-1/2}$, which can be truncated at any order to obtain the desired level of detail. In the limit of large Ω, terms of $\mathcal{O}(\Omega^{-1/2})$ may be neglected, which yields the linear noise approximation (LNA). The LNA forms the basis of the approximation to the CME with extrinsic variables which we derive in **Chapter 2**. For systems consisting of only zero- and first-order reactions, and for a subset of systems with second-order reactions, the mean and variance of the CME and the LNA agree exactly.[50] In cases where the propensities are non-linear, the solution of the LNA should be viewed as an approximation of that of the CME, though the errors in these approximations decrease with increasing molecule number. Another limitation of the LNA is that it can only describe systems characterised by a single steady state. Despite these limitations, the LNA has been used successfully. Some examples include the inference of kinetic parameters and quantification of extrinsic and/or intrinsic noise from single-cell data,[51–53] and the calibration of single-cell fluorescence measurements.[54]

1.4 Scope and thesis outline

The aim of this thesis is to study how fluctuating environments can affect biological systems and cellular communities. We look at how gene expression noise manifests itself over time or across a population, how the system is affected by extrinsic noise, and what possible implications this has for cellular fitness and population survival.

Chapter 2 describes a theoretical approach to this problem, without specifying one type of application, yet. Here, we derive an analytical approximation of the CME to assess the effects of slow environmental noise on gene regulatory networks, which manifests itself as fluctuations in reaction rates. The approximation is based on the linear noise approximation, in combination with separation of time scales between the fast intrinsic fluctuations associated with intracellular interactions and slowly changing environmental variables. We verify its validity using stochastic simulations and show that the mathematical framework provides accurate predictions of system characteristics for a range of biological networks. Lastly, we show that it gives insight into design principles of synthetic networks.

Chapter 3 looks at CRISPR-Cas, the bacterial immune system against bacte-

riophages, in a growing population of *Escherichia coli*. We experimentally monitor invader clearance at a high temporal resolution, and quantify the variation in the various stages of the immune response across the population. By comparing physiological cell features, we determine factors that contribute to this variability. We create a minimal agent-based stochastic simulation framework to replicate the experimental data and provide further insight into the dynamics of the bacterial adaptive immune response at both the molecular and the population scale.

The work started in **Chapter 3** is continued in **Chapter 4**, where we use detailed single-cell lineage data to take a more in-depth look at the adaptation process, the stage of adaptive immunity where cells acquire a new memory of an infection. Using three different targets which have previously been shown to elicit diverse immune responses, we aim to learn more about the largely unknown molecular mechanism behind primed adaptation. With a refined version of the agent-based stochastic simulation model, we contrast competing mechanisms and compare features of the simulated trajectories with experimental data.

In **Chapter 5**, we turn to the realm of plants to study cell-to-cell gene expression noise in *Arabidopsis thaliana*. We aim to quantify intrinsic and extrinsic noise in different tissue types. In young leaves, the system is not in stationary condition, and adjacent cells are connected through plasmodesmata and may be closely related due to recent cell division. We study if this coupling of cells in growing tissues affects gene expression noise in individual cells. We use a combination of experimental approaches, mathematical modelling, and stochastic simulation to show that gene expression in *A. thaliana* cells fluctuates over time, that extrinsic noise is stronger than intrinsic noise, and that there is spatial correlation between young cells due to inheritance of cellular conditions.

In **Chapter 6**, I discuss the main findings and their implications in a broader context. In this chapter, the challenges with defining, measuring, analysing, and computational modelling of noise in biological systems are discussed. I then turn to future research strategies.

Chapter 2

Extending the linear-noise approximation to biochemical systems influenced by intrinsic noise and slow lognormally distributed extrinsic noise

This chapter is published as:
Emma M. Keizer[†], Björn Bastian[†], Robert W. Smith, Ramon Grima, Christian Fleck
Physical Review E 99(5) (2019): 052417

DOI: https://doi.org/10.1103/PhysRevE.99.052417

[†]These authors contributed equally

Abstract

It is well known that the kinetics of an intracellular biochemical network is stochastic. This is due to intrinsic noise arising from the random timing of biochemical reactions in the network as well as due to extrinsic noise stemming from the interaction of unknown molecular components with the network and from the cell's changing environment. While there are many methods to study the effect of intrinsic noise on the system dynamics, few exist to study the influence of both types of noise. Here we show how one can extend the conventional linear-noise approximation to allow for the rapid evaluation of the molecule numbers statistics of a biochemical network influenced by intrinsic noise and by slow lognormally distributed extrinsic noise. The theory is applied to simple models of gene regulatory networks and its validity confirmed by comparison with exact stochastic simulations. In particular we show how extrinsic noise modifies the dependence of the variance of the molecule number fluctuations on the rate constants, the mutual information between input and output signalling molecules and the robustness of feed-forward loop motifs.

2.1 Introduction

Studying the dynamics of biological systems is essential in understanding the design principles underlying biochemical and synthetic networks. However, this task is also challenging given the complexity of interactions between the system's components and its environment. At the molecular level, biological processes possess a certain degree of randomness as chemical reactions are probabilistic events.[55] This stochasticity or *noise* in biological networks has widely varying functional roles, and can be both advantageous and detrimental to cells. Positive effects include phenotypic diversity and the ability to quickly adapt to changing environmental conditions, thus increasing the probability of survival.[56,57] In contrast, noise can also restrict the ability of a cell to resolve input signals of different strengths and hence reduces the information that can be accessed about the external environment .[58] For these reasons, the network's topology either exploits or attenuates noise.

Due to the high complexity of the intracellular machinery, one often can only study a set of reactions between a certain number of observable molecular components. We call this subset of reactions and components *a biochemical system of interest*. The rest of the intracellular and extracellular reactions, species and environmental cues not accounted for, we call *the dynamic environment*. Noise in the biochemical system's dynamics can then stem from either itself or from the dynamic environment. That originating from itself, i.e. from the inherent discreteness of the molecules participating in the biochemical system is called *intrinsic noise*;[41] this type of noise increases with decreasing average molecule numbers and hence is particularly relevant to intracellular dynamics due to the low copy number of genes, messenger RNA (mRNA) and some proteins in a single cell .[59] The noise stemming from the dynamic environment is termed *extrinsic noise* and this affects the biochemical system of interest via modulation of its rate constants. Several studies have shown that a consideration of both types of noise is crucial to understanding gene expression and biochemical dynamics in both prokaryotic and eukaryotic systems.[59–61]

The well-mixed stochastic description of point-particle biochemical systems is given by the chemical master equation (CME).[41] However, this equation rarely can be solved exactly for gene regulatory networks with feedback interactions and when it can, it invariably is for the case of zero extrinsic noise.[42,43] There are two distinct ways to proceed: (i) exact stochastic simulation or (ii) approximate analytic techniques, which we discuss in this order next.

The stochastic simulation algorithm (SSA), as formulated by Gillespie,[55] provides trajectories which are consistent with the CME provided the rate constants are time-independent, i.e, it can only describe intrinsic noise since extrinsic noise manifests as noise in the rate constants. The SSA is noteworthy because it is exact in the limit of an infinite number of samples. Modifications to the SSA that take into account both types of noise have been devised by Shahrezaei *et al.*,[62] Anderson[46]

and by Voliotis *et al.*[47] The latter is the most computationally efficient algorithm of the three and also the only one not suffering from numerical integration error. However, the disadvantage of these methods is that a large number of simulations may be required to obtain statistically significant results.[57]

Alternatively, various approximation schemes have been derived to obtain analytical expressions for statistical moments and the marginal distributions of molecular numbers (for a recent review see[63]). Scott *et al.*[64] and Toni *et al.*[65] develop approximate methods based on the linear-noise approximation[41] which allow the calculation of moments for the case of small intrinsic noise together with extrinsic noise originating from rate constants with a static (time-independent) normally distribution. Roberts *et al.*[66] develop a different type of approximation based on the WKB (Wentzel-Kramers-Brillouin) method and the assumption of extrinsic noise originating from rate constants with a static negative binomial distribution. The advantage of the linear-noise approximation methods is the ease with which they can be calculated for systems with a large number of interacting components since the method amounts to solving a Lyapunov equation that can be computed very efficiently[63] while the major disadvantage is that the fluctuating rate parameters can become negative due to the assumption of a normal distribution. In contrast, the WKB method is difficult to extend to more than one variable however the fluctuating rates are positive.

In this work, we present a novel method based on the linear-noise approximation that is applicable to systems with intrinsic noise together with extrinsic noise originating from rate constants with a static lognormal distribution. It is assumed that the timescale of the extrinsic noise is much longer than that of intrinsic noise, a biologically realistic scenario [56] (correlation times for extrinsic fluctuations in *Escherichia coli* is of the order of 40 minutes which corresponds to the cell cycle period,[67,68] while intrinsic processes typically happen on the order of a minute or shorter timescales). Our method is computationally efficient (due to the use of the linear-noise approximation) while maintaining physical realism by enforcing positive fluctuating rate constants. It hence overcomes the disadvantages of the aforementioned existing frameworks (see previous paragraph). It is also the case that the lognormal distribution appears to be ubiquitous in cell biology[69–71] and hence it is the obvious choice to characterise the generally non-Gaussian distribution of positive fluctuating rates.

The paper is divided as follows. In Section II we develop the theory and derive general expressions for the mean, variances and power spectra of fluctuating molecule numbers in a general biochemical system subjected to lognormal extrinsic noise and intrinsic noise. In Section III, the theory is applied to study how extrinsic noise affects: (1) the second moments of protein numbers in a three-stage model of gene expression and an auto-regulatory genetic feedback loop, (2) the information transfer through a simple biochemical system and (3) the robustness of feed-forward motifs.

2.2 Theory

To correctly model the effects of extrinsic noise, the variables describing extrinsic fluctuations are introduced at the level of the CME. This renders the theory intractable, besides in simplest cases. To overcome this obstacle we propose an asymptotic expansion method relying on three main steps. In Section 2.2.1, we follow van Kampen's system size expansion[72] which is truncated at first order to obtain the well-known linear-noise approximation (LNA)[41] as a function of the time-dependent extrinsic variables. Subsequently, we assume timescale separation between the fast intrinsic fluctuations and slowly changing extrinsic variables (Section 2.2.2). This allows us, as a final step, to employ a small noise expansion of the extrinsic stochastic variables and obtain closed-form expressions for the means, variances and power spectra of the biochemical system components in Section 2.2.3.

2.2.1 LNA with extrinsic variables

We consider a chemical network with system volume Ω and N molecular species with copy numbers $X_i(t)$ and concentrations $x_i(t) = X_i(t)/\Omega$, where $i \in \{1, \ldots, N\}$. The chemical reactions are described by R reaction channels with rate functions $\tilde{f}_j(\mathbf{x}, \boldsymbol{\eta}, \Omega)$ with $j \in \{1, \ldots, R\}$ and stoichiometric matrix $\mathbf{S} \in \mathbb{Z}^{N \times R}$. External fluctuations in the rates, that are not included in the microscopic description, are described by slowly changing stochastic variables $\eta_k(t)$, $k \in \{1, \ldots, M\}$, where M equals the number of fluctuating parameters $\bar{c}_k$,

$$\bar{c}_k(t) = c_k \nu_k(t) = c_k \left(1 + \eta_k(t)\right), \tag{2.1}$$

such that $\langle \bar{c}_k(t) \rangle = c_k$. Lognormal coloured noise $\nu_k(t)$ ensures positive rate constants which avoids spurious production or degradation. Furthermore, lognormal rather than normal distributions have been measured for gene expression rates.[62, 67] The fluctuations around the mean values c_k are then proportional to $\eta_k(t)$. As shown in Appendix 2.A, the lognormal variables $\nu_k(t) = \exp\left(\mu_k(t) - \frac{1}{2}\epsilon_k^2\right)$ with mean 1 may by constructed from an Ornstein-Uhlenbeck process for the normal variables $\mu_k(t)$ with standard deviations ϵ_k.

The probability that the network is in a state with copy numbers $\mathbf{X}$ and stochastic variables $\boldsymbol{\eta}$ at time t is given by $P(\mathbf{X}, \boldsymbol{\eta}, t)$, and $\Omega \tilde{f}_j(\mathbf{x}, \boldsymbol{\eta}, \Omega)\, \mathrm{d}t$ is the probability that a reaction of type j occurs in time $\mathrm{d}t$. The system's state after the reaction is defined by copy numbers $X_i + S_{ij}$ and the value of $\boldsymbol{\eta}$ at that time.

The chemical master equation describing the microscopic system dynamics is then given by

$$\frac{\mathrm{d}P(\mathbf{X}, \boldsymbol{\eta}, t)}{\mathrm{d}t} = \Omega \sum_{j=1}^{R} \left(\prod_{i=1}^{N} E_i^{-S_{ij}} - 1 \right) \tilde{f}_j(\mathbf{x}, \boldsymbol{\eta}, \Omega) P(\mathbf{X}, \boldsymbol{\eta}, t), \tag{2.2}$$

where E_i is a step operator that is defined by the action $E_i^n g(X_1, ..., X_i, ..., X_N) = g(X_1, ..., X_i + n, ..., X_N)$, the product of which takes into account all system states that can evolve to the state given by $\mathbf{X}$ and $\boldsymbol{\eta}$. The probability flux away from state $(\mathbf{X}, \boldsymbol{\eta})$ due to reaction j is given by $-\Omega \tilde{f}_j(\mathbf{x}, \boldsymbol{\eta}, \Omega) P(\mathbf{X}, \boldsymbol{\eta}, t)$.

If the system volume is sufficiently large, the fluctuations of the concentrations $\mathbf{x}$ due to a couple of reactions on short timescales are relatively small. Thus, the rate functions change more or less continuously on a larger timescale. We can therefore define a macroscopic limit $\Omega \to \infty$ with the concentrations, transition rates and deterministic reaction rate equations are given by:

$$\boldsymbol{\phi} = \lim_{\Omega \to \infty} \mathbf{x}, \tag{2.3a}$$

$$\mathbf{f}(\boldsymbol{\phi}, \boldsymbol{\eta}) = \lim_{\Omega \to \infty} \tilde{\mathbf{f}}(\mathbf{x}, \boldsymbol{\eta}, \Omega), \tag{2.3b}$$

$$\frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}t} = \mathbf{g}(\boldsymbol{\phi}, \boldsymbol{\eta}) = \mathbf{S}\mathbf{f}(\boldsymbol{\phi}, \boldsymbol{\eta}). \tag{2.3c}$$

Following the system size expansion by van Kampen[41] we relate the microscopic and macroscopic vectors via

$$\mathbf{x} = \boldsymbol{\phi} + \Omega^{-\frac{1}{2}} \boldsymbol{\xi}, \tag{2.4}$$

where a new variable $\boldsymbol{\xi}$ denotes the microscopic fluctuations around the macroscopic concentrations $\boldsymbol{\phi}$. It obeys the stochastic differential equation[73]

$$\mathrm{d}\boldsymbol{\xi}(t) = \boldsymbol{A}(\boldsymbol{\eta}(t))\boldsymbol{\xi}(t)\,\mathrm{d}t + \boldsymbol{B}\left(\boldsymbol{\eta}(t)\right)\mathrm{d}\mathbf{W}(t) \tag{2.5}$$

with the Wiener process increments $\mathrm{d}\mathbf{W}(t)$, Jacobian matrix $\boldsymbol{A}(\boldsymbol{\eta})$ and diffusion matrix $\boldsymbol{B}(\boldsymbol{\eta})$.

2.2.2 Timescale separation between dynamics of intrinsic and extrinsic fluctuations

Integrating a stationary solution

As Eq. (2.9) generally cannot be solved analytically, we assume that the supremum of the timescales of intrinsic noise as given by the absolute value of the inverse of the eigenvalues of the Jacobian $\boldsymbol{A}$ is much less than the timescale of extrinsic noise. This allows us to split the time axis into intervals$[t_n, t_n + \triangle t]$, on which the extrinsic variables $\boldsymbol{\eta}$ are treated as constants $\boldsymbol{\eta}^n \equiv \boldsymbol{\eta}(t_n)$. For well defined stationary solutions we require the existence of a unique macroscopic stationary solution $\boldsymbol{\phi}^s(\boldsymbol{\eta}^n)$ of Eq. (2.3c),

$$\mathbf{g}(\boldsymbol{\phi}^s(\boldsymbol{\eta}^n), \boldsymbol{\eta}^n) = 0\,, \tag{2.6}$$

and that the Jacobian matrix has only eigenvalues with negative real part (*i.e.* a stable monotonic system). The stationary Jacobian and diffusion matrices are

$$\boldsymbol{A}(\boldsymbol{\eta}^n) = \frac{\partial \mathbf{g}}{\partial \boldsymbol{\phi}}(\boldsymbol{\phi}^s(\boldsymbol{\eta}^n), \boldsymbol{\eta}^n), \tag{2.7}$$

$$\boldsymbol{B}\left(\boldsymbol{\eta}^n\right)\boldsymbol{B}\left(\boldsymbol{\eta}^n\right)^T = \mathbf{S}\,\mathrm{diag}\left(\mathbf{f}(\phi^s(\boldsymbol{\eta}^n), \boldsymbol{\eta}^n)\right)\mathbf{S}^T. \tag{2.8}$$

Aiming at a stationary solution $\mathbf{x}(t)$ that makes it possible to obtain expressions for the mean, variance and power spectrum, we further follow the well known *linear-noise approximation*:[41] we linearise the rate equations (2.3c) and use Eqs. (2.4) and (2.5) to obtain the linear stochastic differential equation for $t \in [t_n, t_n + \triangle t]$. The stochastic differential equation describing the fluctuations in molecule numbers can be derived by using Eq. (2.4) (with ϕ replaced by $\phi^s(\eta^n)$) together with Eq. (5) (with $\eta(t)$ replaced by η^n) to obtain:

$$\mathrm{d}\mathbf{x}(t) = \boldsymbol{A}(\boldsymbol{\eta}^n)(\mathbf{x}(t) - \boldsymbol{\phi}^s(\boldsymbol{\eta}^n))\,\mathrm{d}t + \frac{1}{\sqrt{\Omega}}\boldsymbol{B}(\boldsymbol{\eta}^n)\,\mathrm{d}\mathbf{W}(t). \tag{2.9}$$

with the stationary solution:[73]

$$\mathbf{x}^s(t) = \boldsymbol{\phi}^s(\boldsymbol{\eta}^n) + \frac{1}{\sqrt{\Omega}}\boldsymbol{\xi}(\boldsymbol{\eta}^n, t), \tag{2.10a}$$

$$\boldsymbol{\xi}(\boldsymbol{\eta}^n, t) = \int_{-\infty}^{t} e^{\boldsymbol{A}(\boldsymbol{\eta}^n)(t-t')}\boldsymbol{B}\left(\boldsymbol{\eta}^n\right)\mathrm{d}\mathbf{W}(t'). \tag{2.10b}$$

Calculating mean concentrations and variances

To evaluate averages and variances of the stationary concentrations $\mathbf{x}^s$ we denote averaging over intrinsic fluctuations $\boldsymbol{\xi}$ by $\langle\ \rangle_i$ and over the extrinsic variables $\boldsymbol{\eta}$ by $\langle\ \rangle_e$. The mean concentrations then simplify to

$$\left\langle \left\langle \mathbf{x}^s \right\rangle_i \right\rangle_e = \left\langle \boldsymbol{\phi}^s(\boldsymbol{\eta}) \right\rangle_e \tag{2.11}$$

where $\langle \boldsymbol{\xi} \rangle_i$ evaluates to zero on each of the time intervals.[41] In the same way, the covariance matrix of $\mathbf{x}^s$ in the timescale separation approximation can be written as

$$\begin{aligned} \mathbf{V}\left(\mathbf{x}^s\right) &= \left\langle \left(\mathbf{x}^s - \left\langle \mathbf{x}^s \right\rangle\right)\left(\mathbf{x}^s - \left\langle \mathbf{x}^s \right\rangle\right)^T \right\rangle \\ &= \mathbf{V}\left(\boldsymbol{\phi}^s\right) + \frac{1}{\Omega}\mathbf{V}\left(\boldsymbol{\xi}\right). \end{aligned} \tag{2.12}$$

where $\langle\ \rangle$ abbreviates $\langle\langle\ \rangle_i\rangle_e$ and we define the covariance matrices of $\boldsymbol{\phi}^s$ and $\boldsymbol{\xi}$ as

$$\mathbf{V}\left(\boldsymbol{\phi}^s\right) = \left\langle \boldsymbol{\phi}^s \boldsymbol{\phi}^{sT} \right\rangle_e - \left\langle \boldsymbol{\phi}^s \right\rangle_e \left\langle \boldsymbol{\phi}^{sT} \right\rangle_e, \tag{2.13}$$

$$\mathbf{V}\left(\boldsymbol{\xi}\right) = \left\langle \left\langle \boldsymbol{\xi}\boldsymbol{\xi}^T \right\rangle_i \right\rangle_e. \tag{2.14}$$

One can obtain $\mathbf{V}(\boldsymbol{\xi})$ algebraically. For the sake of brevity we anticipate the result using Eq. (2.20) and (2.22)

$$\mathbf{V}\left(\boldsymbol{\xi}\right) = \left\langle \left\langle \boldsymbol{\xi}\boldsymbol{\xi}^T \right\rangle_i \right\rangle_e \ = \ \left\langle \mathbf{G}_i(t,t) \right\rangle_e = \left\langle \boldsymbol{C}(\boldsymbol{\eta},\boldsymbol{\eta}) \right\rangle_e \tag{2.15}$$

were the Lyapunov matrix $\boldsymbol{C}$ is evaluated at equal times and we calculate the time correlation function $\mathbf{G}(t_1,t_2)$ as an intermediate step to calculate the power spectrum in what follows.

Calculating power spectra via Fourier transformation

The spectrum matrix of a stationary stochastic process is connected to the time correlation function by the *Wiener-Khinchin theorem* via a Fourier transformation if the time correlation is sufficiently smooth,[41,73]

$$\boldsymbol{P}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-i\omega\tau} \mathbf{G}(\tau,0)\, \mathrm{d}\tau. \tag{2.16}$$

We first introduce the time correlation function of a stationary solution $\mathbf{x}^s$ (Eq. 2.10) as

$$\mathbf{G}(t_1,t_2) = \left\langle \left(\mathbf{x}^s(t_1) - \langle \mathbf{x}^s \rangle\right) \left(\mathbf{x}^s(t_2) - \langle \mathbf{x}^s \rangle\right)^T \right\rangle \tag{2.17}$$

$$= \mathbf{G}_e(t_1,t_2) + \frac{1}{\Omega} \left\langle \mathbf{G}_i(t_1,t_2) \right\rangle_e. \tag{2.18}$$

The second equality holds under timescale separation conditions and represents the sum of the time correlation function of the macroscopic stationary state

$$\mathbf{G}_e(t_1,t_2) = \left\langle \boldsymbol{\phi}^s(\boldsymbol{\eta}^1) \boldsymbol{\phi}^s(\boldsymbol{\eta}^2)^T \right\rangle_e - \left\langle \boldsymbol{\phi}^s \right\rangle_e \left\langle \boldsymbol{\phi}^{sT} \right\rangle_e \tag{2.19}$$

and the time correlation function of the intrinsic noise subject to slow extrinsic fluctuations

$$\mathbf{G}_i(t_1,t_2) = \left\langle \boldsymbol{\xi}\left(\boldsymbol{\eta}^1\right) \boldsymbol{\xi}\left(\boldsymbol{\eta}^2\right)^T \right\rangle_i. \tag{2.20}$$

To evaluate Eq. (2.20) we follow the calculation of the variance in stationary solution by Gardiner[73] (having defined $\boldsymbol{A}$ as the Jacobian it has opposite sign to the matrix in the reference) and generalise it by evaluating the results at different times to

obtain

$$\mathbf{G}_i(t_1,t_2) = e^{\boldsymbol{A}(\boldsymbol{\eta}^1)(t_1-\min(t_1,t_2))}\ \boldsymbol{C}\left(\boldsymbol{\eta}^1,\boldsymbol{\eta}^2\right) e^{\boldsymbol{A}(\boldsymbol{\eta}^2)^T(t_2-\min(t_1,t_2))} \tag{2.21}$$

where the $\boldsymbol{C}$ matrix is defined by the Lyapunov equation

$$\begin{aligned}\boldsymbol{A}(\boldsymbol{\eta}^1)\,\boldsymbol{C}\left(\boldsymbol{\eta}^1,\boldsymbol{\eta}^2\right) + \boldsymbol{C}\left(\boldsymbol{\eta}^1,\boldsymbol{\eta}^2\right)\boldsymbol{A}(\boldsymbol{\eta}^2)^T \\ = -\boldsymbol{B}\left(\boldsymbol{\eta}^1\right)\boldsymbol{B}\left(\boldsymbol{\eta}^2\right)^T.\end{aligned} \tag{2.22}$$

Exploiting timescale separation, we split the spectrum matrix Eq. (2.16) into two terms according to Eq. (2.18),

$$\boldsymbol{P}(\omega) = \boldsymbol{P}_e(\omega) + \frac{1}{\Omega}\left\langle \boldsymbol{P}_i(\omega)\right\rangle_e. \tag{2.23}$$

With Eq. (2.21) and $\boldsymbol{\eta}^n \equiv \boldsymbol{\eta}(t_n)$ with $t_1 = \tau$ and $t_2 = 0$ we can express $\left\langle \boldsymbol{P}_i(\omega)\right\rangle_e$ in terms of the matrices

$$\boldsymbol{R}(\omega) = \left\langle \int_0^\infty e^{-\left(-\boldsymbol{A}(\boldsymbol{\eta}^1)+i\omega\right)\tau}\boldsymbol{C}(\boldsymbol{\eta}^1,\boldsymbol{\eta}^2)\,\mathrm{d}\tau \right\rangle_e, \tag{2.24a}$$

$$\boldsymbol{R}(\omega)^{*T} = \left\langle \int_0^\infty \boldsymbol{C}(\boldsymbol{\eta}^2,\boldsymbol{\eta}^1)e^{\left(\boldsymbol{A}(\boldsymbol{\eta}(\tau))^T+i\omega\right)\tau}\,\mathrm{d}\tau \right\rangle_e. \tag{2.24b}$$

Assuming stationarity of $\boldsymbol{\eta}(t)$ it can be shown that

$$\boldsymbol{R}(\omega) + \boldsymbol{R}(\omega)^{*T} = \int_{-\infty}^{+\infty} e^{-i\omega\tau}\mathbf{G}_i(\tau,0)\,\mathrm{d}\tau. \tag{2.25}$$

Further, we split the Taylor expansion of the Jacobian in two,

$$\boldsymbol{A}(\boldsymbol{\eta}) = \boldsymbol{A}^0 + \left(\boldsymbol{A}(\boldsymbol{\eta}) - \boldsymbol{A}^0\right), \qquad \boldsymbol{A}^0 \equiv \boldsymbol{A}(\mathbf{o}), \tag{2.26}$$

and summarise the quantities of interest for Eq. (2.23):

$$\begin{aligned}\boldsymbol{P}_e(\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-i\omega\tau}\Big(\left\langle \boldsymbol{\phi}^s(\boldsymbol{\eta}^1)\boldsymbol{\phi}^s(\boldsymbol{\eta}^2)^T\right\rangle_e \\ - \left\langle \boldsymbol{\phi}^s\right\rangle_e\left\langle \boldsymbol{\phi}^{sT}\right\rangle_e\Big)\,\mathrm{d}\tau,\end{aligned} \tag{2.27}$$

$$\begin{aligned}\boldsymbol{R}(\omega) = \int_0^\infty e^{-\left(-\boldsymbol{A}^0+i\omega\right)\tau}\times \\ \left\langle e^{\left(\boldsymbol{A}(\boldsymbol{\eta}^1)-\boldsymbol{A}^0\right)\tau}\boldsymbol{C}(\boldsymbol{\eta}^1,\boldsymbol{\eta}^2)\right\rangle_e\,\mathrm{d}\tau,\end{aligned} \tag{2.28}$$

$$\left\langle \boldsymbol{P}_i(\omega)\right\rangle_e = \frac{1}{2\pi}\left(\boldsymbol{R}(\omega) + \boldsymbol{R}(\omega)^{*T}\right). \tag{2.29}$$

2.2.3 Small noise expansion

In the third and last step, we expand equations (2.11), (2.13), (2.15), (2.27) and (2.28) in Taylor series in the M noise variables $\boldsymbol{\eta}$. To calculate expected values, also of the time dependent integrands in (2.27) and (2.28), we consequently need the n-point time correlation functions of the extrinsic noise variables. To this end, we need the following results derived in Appendix 2.B.

For smooth functions $y_k(\mu_k)$ of normal stochastic variables μ_k with m^{th} derivatives $y_k^{(m)}(\mu_k)$, the derived n-point time correlation function reads:

$$\langle y_1 \dots y_n \rangle = \sum_{|\boldsymbol{d}^n|=0}^{\infty} \left(\prod_{k=1}^{n} \left\langle y_k^{(m_k)}(\mu_k) \right\rangle \prod_{\substack{i,j=1 \\ i<j}}^{n} \frac{\Delta_{ij}^{d_{ij}}}{d_{ij}!} \right) \tag{2.30}$$

which generalises a previous result for $n = 2$.[74] Each value of $|\boldsymbol{d}^n| = \sum d_{ij}$ involves summation over all index tuples $\boldsymbol{d}^n = (d_{12}, d_{13}, d_{23}, \dots, d_{(n-1)\,n})$. The 2-point time correlation functions of the normal stochastic variables are denoted by $\Delta_{ij} = \langle \mu_i \mu_j \rangle$. We point out the simplicity of this result for lognormal variables with mean $\langle \nu_k \rangle = 1$,

$$\langle \nu_1 \dots \nu_n \rangle = \exp\left(\sum_{1 \le i < j \le n} \Delta_{ij} \right). \tag{2.31}$$

In the small noise expansion we use the shifted $\eta_k = \nu_k - 1$ with the time correlation functions (see Appendix 2.B)

$$\langle \eta_1 \dots \eta_n \rangle = \sum_{u=\lfloor \frac{n+1}{2} \rfloor}^{\infty} {\sum_{|\boldsymbol{d}^n|=u}}' \prod_{\substack{i,j=1 \\ i<j}}^{n} \frac{\Delta_{ij}^{d_{ij}}}{d_{ij}!} \tag{2.32}$$

where the prime denotes the restriction of the sum to terms where for each $j \in \{1, \dots, n\}$ there is a $i < j$ or $k > j$ with $d_{ij} \neq 0$ or $d_{jk} \neq 0$. The floor function $\lfloor \frac{n+1}{2} \rfloor$ gives the smallest u for which this can be satisfied.

Selected terms of the final results that we present in the following are exemplarily evaluated in Appendix 2.G to further clarify the complex notation.

Mean

We define the Taylor expansion of the mean concentrations in Eq. (2.11) with the unusual but beneficial notation

$$\langle \mathbf{x}^s(\boldsymbol{\eta}) \rangle = \langle \boldsymbol{\phi}^s(\boldsymbol{\eta}) \rangle = \sum_{n=0}^{\infty} \sum_{\#\boldsymbol{r}=n} \boldsymbol{\phi}^s_{(\boldsymbol{r})} \left\langle \eta_{(\boldsymbol{r})} \right\rangle \tag{2.33}$$

where the multi-indices

$$\boldsymbol{r} = (r_1, \ldots, r_n)\,, \quad \#\boldsymbol{r} = n, \quad r_i \in \{1, \ldots, M\} \tag{2.34a}$$

are sorted by their length $\#\boldsymbol{r}$ and we denote the Taylor coefficients and products of the noise variables as

$$\boldsymbol{\phi}^s_{(\boldsymbol{r})} \equiv \frac{1}{n!} \frac{\partial}{\partial \eta_{r_1}} \cdots \frac{\partial}{\partial \eta_{r_n}} \, \boldsymbol{\phi}^s(\boldsymbol{\eta})|_{\boldsymbol{\eta}=\mathbf{o}}\,, \tag{2.34b}$$

$$\eta_{(\boldsymbol{r})} \equiv \eta_{r_1} \ldots \eta_{r_n}. \tag{2.34c}$$

The sum over $\#\boldsymbol{r} = n$ involves all $\boldsymbol{r}$ tuples of length n. By inserting Eq. (2.32) into Eq. (2.33) we obtain the final result for the mean concentrations in terms of the covariances of the independent normal stochastic variables μ_i with standard deviations ϵ_i, that is, their 2-point correlation functions evaluated at equal times

$$\Gamma_{ij} \equiv \Delta_{ij}(0) = \delta_{ij}\epsilon_i^2 \tag{2.35}$$

with the Kronecker delta (see details in Appendix 2.C):

$$\langle \mathbf{x}^s \rangle = \sum_{u=0}^{\infty} \sum_{n=0}^{2u} \sum_{\#\boldsymbol{r}=n} \boldsymbol{\phi}^s_{(\boldsymbol{r})} \sum_{|\boldsymbol{d}^n|=u}{}' \prod_{\substack{i,j=1 \\ i<j}}^{n} \frac{1}{d_{ij}!} \Big(\Gamma_{r_i r_j} \Big)^{d_{ij}}. \tag{2.36}$$

The prime at the sum denotes the same restriction as in Eq. (2.32): the sum is running over all tuples $\boldsymbol{d}^n = (d_{12}, d_{13}, d_{23}, \ldots, d_{(n-1)\,n})$ with $|\boldsymbol{d}^n| = \sum d_{ij} = u$ that obey the condition that for each $j \in \{1, \ldots, n\}$ there is a $i < j$ or $k > j$ such that $d_{ij} \neq 0$ or $d_{jk} \neq 0$.

Variance

With Eq. (2.15) we write the matrix $V(\boldsymbol{\xi}) = \langle \boldsymbol{C}(\boldsymbol{\eta}, \boldsymbol{\eta}) \rangle$ in analogy to $\langle \mathbf{x}^s \rangle = \langle \boldsymbol{\phi}^s \rangle$. Because $\boldsymbol{C}$ is evaluated at equal times here, we may expand it according to (2.34b)

with Taylor coefficients $\boldsymbol{C}_{(\boldsymbol{r})}$. From Eq. (2.36) we then obtain

$$\mathbf{V}(\boldsymbol{\xi}) = \sum_{u=0}^{\infty} \sum_{n=0}^{2u} \sum_{\#\boldsymbol{r}=n} \boldsymbol{C}_{(\boldsymbol{r})} {\sum_{|\boldsymbol{d}^n|=u}}' \prod_{\substack{i,j=1 \\ i<j}}^{n} \frac{1}{d_{ij}!} \Big(\Gamma_{r_i r_j} \Big)^{d_{ij}}. \tag{2.37}$$

The derivation of the series for $\mathbf{V}(\boldsymbol{\phi}^s)$ is conducted similarly. For the expansion of $\boldsymbol{\phi}^s \boldsymbol{\phi}^{s\,T}$ in Eq. (2.13), an ordinary multi-index $\boldsymbol{q} = (q_1, q_2)$, $|\boldsymbol{q}| = q_1 + q_2$ with integer $q_1, q_2 \geq 1$ is used to define the lengths of two multi-indices $\boldsymbol{r}^1$ and $\boldsymbol{r}^2$ (see Eq. 2.34) and finally

$$\mathbf{V}(\boldsymbol{\phi}^s) = \sum_{u=1}^{\infty} \sum_{n=2}^{2u} \sum_{|\boldsymbol{q}|=n} \sum_{\#\boldsymbol{r}^1=q_1} \sum_{\#\boldsymbol{r}^2=q_2} \boldsymbol{\phi}^s_{(\boldsymbol{r}^1)} \boldsymbol{\phi}^{s\,T}_{(\boldsymbol{r}^2)} {\sum_{|\boldsymbol{d}^n|=u}}'' \prod_{\substack{i,j=1 \\ i<j}}^{n} \frac{1}{d_{ij}!} \Big(\Gamma_{f_i^2 f_j^2} \Big)^{d_{ij}}. \tag{2.38}$$

The doubly primed sum (see Appendix 2.D) denotes the restriction of the summation by one additional condition: there is at least one $d_{ij} \neq 0$ for which $i \leq q_1$ and $j > q_1$ since all other terms cancel in the subtraction in Eq. (2.13). For later use, we define the index functions $f_i^k(\boldsymbol{r}^1, \ldots, \boldsymbol{r}^k)$ for a generalised set of multi-indices $\boldsymbol{r}^1, \ldots, \boldsymbol{r}^k$ with $\#\boldsymbol{r}^i = q_i$ and $\boldsymbol{q} = (q_1, \ldots, q_k)$,

$$f_i^k = \begin{cases} r_i^1 & \text{if } 1 \leq i \leq q_1 \\ r^2_{(i-q_1)} & \text{if } q_1 < i \leq q_1 + q_2 \\ \vdots & \vdots \\ r^k_{(i-|\boldsymbol{q}|+q_k)} & \text{if } q_1 + \cdots + q_{k-1} < i \leq |\boldsymbol{q}| \end{cases} \tag{2.39}$$

that we have used in Eq. (2.38) to refer to $\boldsymbol{r}^1$ and $\boldsymbol{r}^2$.

Power spectrum

In Eq. (2.27) for the spectrum matrix $\boldsymbol{P}_e(\omega)$ the integral from $-\infty$ to 0 is the complex conjugate of the integral from 0 to ∞. Thus the small noise expansion proceeds in complete analogy to $\mathbf{V}(\boldsymbol{\phi}^s)$ in Eq. (2.13) with the result in Eq. (2.38) when the product of covariances $\Gamma_{f_i^2 f_j^2}$ is replaced by the Fourier transform plus its complex conjugate

$$\frac{1}{2\pi} \int_0^{\infty} e^{-\mathrm{i}\omega\tau} \prod_{\substack{i,j=1 \\ i<j}}^{n} \frac{1}{d_{ij}!} \Big(\Delta_{f_i^2 f_j^2}(t_i - t_j) \Big)^{d_{ij}} \, \mathrm{d}\tau \tag{2.40}$$

where the 2-point time correlation functions are

$$\Delta_{ij}(t_i - t_j) \equiv \langle \mu_i(t_i)\mu_j(t_j)\rangle = \Gamma_{ij} e^{-K_i|t_i - t_j|} \tag{2.41}$$

with inverse correlation times K_i.

We finally obtain

$$\boldsymbol{P}_e(\omega) = \sum_{u=1}^{\infty}\sum_{n=2}^{2u}\sum_{|\boldsymbol{q}|=n}\sum_{\#\boldsymbol{r}^1=q_1}\sum_{\#\boldsymbol{r}^2=q_2} \boldsymbol{\phi}^s_{(\boldsymbol{r}^1)}{\boldsymbol{\phi}^s_{(\boldsymbol{r}^2)}}^T \times \sum_{|\boldsymbol{d}^n|=u}{}'' \frac{\pi^{-1}\,\Theta(\boldsymbol{d}^n,\boldsymbol{r}^1)}{\omega^2 + \Theta^2(\boldsymbol{d}^n,\boldsymbol{r}^1)} \prod_{\substack{i,j=1\\ i<j}}^{n} \frac{1}{d_{ij}!}\Big(\Gamma_{f_i^2 f_j^2}\Big)^{d_{ij}} \tag{2.42}$$

with $\Theta(\boldsymbol{d}^n,\boldsymbol{r}^1)$ derived from integral (2.40) in Appendix 2.E,

$$\Theta(\boldsymbol{d}^n,\boldsymbol{r}^1) = \sum_{i=1}^{\#\boldsymbol{r}^1}\sum_{j=\#\boldsymbol{r}^1+1}^{n} d_{ij}\, K_{r_i^1}\,. \tag{2.43}$$

To calculate the second spectrum matrix under influence of extrinsic fluctuations, $\langle \boldsymbol{P}_i(\omega)\rangle$ in Eq. (2.29), the $\boldsymbol{R}(\omega)$ matrix in Eq. (2.28) requires expanding. First, we expand the Jacobian matrix $\boldsymbol{A}(\boldsymbol{\eta}^1)$ at time t_1 in the same way as $\boldsymbol{\phi}^s(\boldsymbol{\eta})$ in Eq. (2.33) with Taylor coefficients $\boldsymbol{A}_{(\boldsymbol{r})}$ using the multi-index $\boldsymbol{r}$ from Eq. (2.34). The Lyapunov matrix $\boldsymbol{C}(\boldsymbol{\eta}^1,\boldsymbol{\eta}^2)$ (Eq. 2.22) is evaluated with respect to two arguments corresponding to different times. Thus, the Taylor expansion requires a second multi-index

$$\boldsymbol{\sigma} = (\sigma_1,\ldots,\sigma_a)\,, \qquad \#\boldsymbol{\sigma} = a, \qquad \sigma_i \in \{1,2\}, \tag{2.44}$$

such that index σ_i specifies the argument with respect to which the derivative is taken for the Taylor coefficients and to which the components η_i belong,

$$\boldsymbol{C}_{(\boldsymbol{r},\boldsymbol{\sigma})} \equiv \frac{1}{a!}\frac{\partial}{\partial\eta_{r_1}^{\sigma_1}}\cdots\frac{\partial}{\partial\eta_{r_a}^{\sigma_a}}\,\boldsymbol{C}(\boldsymbol{\eta}^1,\boldsymbol{\eta}^2)\big|_{\boldsymbol{\eta}^1=\boldsymbol{\eta}^2=\mathbf{o}}\,, \tag{2.45}$$

$$\eta_{(\boldsymbol{r},\boldsymbol{\sigma})} \equiv \eta_{r_1}^{\sigma_1}\ldots\eta_{r_a}^{\sigma_a}. \tag{2.46}$$

The Taylor expansion of $\exp\big(\big(\boldsymbol{A}(\boldsymbol{\eta}^1) - \boldsymbol{A}^0\big)\,\tau\big)$ in Eq. (2.28) in c'th order involves c different Taylor coefficients of $\boldsymbol{A}$. To distinguish them, we use sets of multi-indices $\boldsymbol{r}^1,\ldots,\boldsymbol{r}^c$ with an ordinary multi-index

$$\boldsymbol{q} = (q_1,\ldots,q_c), \qquad |\boldsymbol{q}| = q_1 + \cdots + q_c \tag{2.47}$$

and $\#\boldsymbol{r}^i = q_i \geq 1$ for all $i = (1,\ldots,c)$. The result of the expansion and integration

of $\boldsymbol{R}(\omega)$ in Appendix 2.F reads

$$\begin{aligned}\boldsymbol{R}(\omega) = &\sum_{u=0}^{\infty}\sum_{n=0}^{2u}\sum_{a=0}^{n}\sum_{|\boldsymbol{q}|=n-a}\sum_{\#\boldsymbol{r}^1=q_1}\cdots\sum_{\#\boldsymbol{r}^c=q_c}\sum_{\#\boldsymbol{r}^{c+1}=a}\sum_{\#\boldsymbol{\sigma}=a}\\ &\times{\sum_{|\boldsymbol{d}^n|=u}}'\frac{1}{\Big(-\boldsymbol{A}^0+\theta(\boldsymbol{d}^n,|\boldsymbol{q}|,\boldsymbol{r}^{c+1},\boldsymbol{\sigma})+i\omega\Big)^{c+1}}\\ &\times\boldsymbol{A}_{(\boldsymbol{r}^1)}\cdots\boldsymbol{A}_{(\boldsymbol{r}^c)}\boldsymbol{C}_{(\boldsymbol{r}^{c+1},\boldsymbol{\sigma})}\prod_{\substack{i,j=1\\ i<j}}^{n}\frac{1}{d_{ij}!}\Big(\Gamma_{f_i^{c+1}f_j^{c+1}}\Big)^{d_{ij}}\end{aligned}\tag{2.48}$$

where the sum $\sum_{|\boldsymbol{q}|=n-a}$ is carried out over all possible c, $(q_1,\ldots,q_c)$, $q_i\geq 1$ with $q_1+\cdots+q_c+a=n$ and

$$\theta(\boldsymbol{d}^n,|\boldsymbol{q}|,\boldsymbol{r}^{c+1},\boldsymbol{\sigma}) = \sum_{\substack{i,j=1\\ i<j}}^{n} d_{ij}\,K_{r_{j-|\boldsymbol{q}|}^{c+1}}\,\beta_{ij}(|\boldsymbol{q}|),\tag{2.49}$$

$$\beta_{ij}(x)=\begin{cases}1 & \text{if } i\leq x<j \text{ and } \sigma_{(j-x)}\neq 1,\\ 1 & \text{if } x<i<j \text{ and } \sigma_{(j-x)}\neq\sigma_{(i-x)},\\ 0 & \text{else.}\end{cases}\tag{2.50}$$

Finally, we add the complex conjugate to $\boldsymbol{R}(\omega)$ and divide by 2π to obtain the spectrum matrix $\langle\boldsymbol{P}_i(\omega)\rangle$ as described by Eq. (2.29).

Automated sum evaluation

While the derivation of the presented results involves a double expansion and a rather complicated notation, its strength lies in the closed-form expressions for the mean, variance and power spectrum of a biochemical system under time scale separation conditions. Stochastic modelling of the underlying chemical master equation with extrinsic fluctuations requires an extension of the Gillespie algorithm[62,75] and it is usually difficult or inefficient to achieve accuracy at different timescales. A fast automated evaluation of the closed-form expressions allows for a systematic approach to analyse the effect of extrinsic fluctuations as a function of different parameters.

The automated sum evaluation has been implemented with the SymPy library for symbolic mathematics[76] in the `ext_noise_expansion` program that can be obtained from `github.com`.[77] Simple systems can be partially or fully evaluated symbolically while larger systems may require numerical parameter values. The limiting step is the calculation of the stationary state $\boldsymbol{\phi}^s$ in Eq. (2.6) as an analytical function of the extrinsic variables. For this task, an external solver

adapted to the system of interest may be used. All sums are formally evaluated before inserting the Taylor coefficients and subsequent term simplification. Taylor coefficients for $\boldsymbol{\phi}^s$, $\boldsymbol{A}$ and $\boldsymbol{B}$ are directly calculated using memoisation to avoid multiple evaluations. Taylor coefficients of $\boldsymbol{C}\left(\boldsymbol{\eta}^1, \boldsymbol{\eta}^2\right)$ from the Lyapunov equation Eq. (2.22) are obtained by expansion and recursive coefficient comparison (this leads to Lyapunov equations for each coefficient that can be constructed explicitly for any order[78]). More details are presented in Appendix 2.G.

2.3 Results

Here, we will present several applications of the theory to study features of noisy biochemical networks. We first introduce a linear model of gene expression to establish the (limits of) validity of our method and quantify the effects of extrinsic noise in different system parameters on intrinsic and extrinsic cell-to-cell variability. We go on to study the potential of negative feedback control to suppress gene expression noise in the presence of extrinsic noise. In a third example, we apply the notion of mutual information to a simple biochemical system to analyse how extrinsic fluctuations affect a network's ability to relay information. Finally, we present an efficient multi-objective optimisation scheme that combines our analytical framework with deterministic dynamics to obtain optimal network topologies and parameters for feed-forward loops.

2.3.1 Nonlinear effects of extrinsic fluctuations on a linear model of gene expression

We start by verifying the validity of our method on the well-known three-stage model of gene expression[60] by introducing coloured noise in various system parameters and studying its effect on the mean and variance of protein numbers. In this model, the promoter can switch between the active and inactive state through e.g. the binding of transcription factors, changes in the chromatin structure, or through the binding of RNA polymerase.[26,60] Transcription can only be initiated when the promoter is in the active state, and mRNA molecules are synthesised, which can subsequently be translated into proteins. Although each step represents several biochemical reactions, the processes of promoter configuration switching, transcription, translation, as well as degradation of mRNA and protein molecules are assumed to obey first-order kinetics.[31] The gene expression model, shown in Figure 2.1(a), includes the dynamics of active promoter sites D, inactive promoter sites D^*, mRNA molecules M, and protein molecules A (note that this notation is unrelated to any terms of the same name in Section 2.2), and is described by the

following reactions

$$\begin{aligned} D^* \underset{k_1}{\overset{k_0}{\rightleftharpoons}} D \overset{v_0}{\to} D + M, \qquad M \overset{d_0}{\to} \emptyset, \\ M \overset{v_1}{\to} M + A, \qquad A \overset{d_1}{\to} \emptyset. \end{aligned} \tag{2.51}$$

We will now show how to write down the LNA as in Eq. (2.9) for extrinsic noise in one of the reaction rates of this system. To introduce extrinsic fluctuations η into the rate function, we multiply a rate constant, e.g. d_1, by a lognormally distributed stochastic variable $\nu = 1 + \eta$

$$\bar{d_1}(t) = d_1 \nu(t) = d_1(1 + \eta(t)). \tag{2.52}$$

The lognormally distributed variable can be constructed from a normal stochastic variable μ with mean zero, variance ϵ^2, and correlation time $\tau = K^{-1}$ defined by the Ornstein-Uhlenbeck process

$$\mathrm{d}\mu(t) = -K\mu(t) + \sqrt{2K}\epsilon\,\mathrm{d}W(t). \tag{2.53}$$

The shifted lognormal stochastic variable η is then given by

$$\eta(t) = \exp\left(\mu(t) - \frac{1}{2}\epsilon^2\right) - 1, \tag{2.54}$$

and we define the magnitude of the extrinsic noise as the coefficient of variation

$$CV = \sqrt{\exp(\epsilon^2) - 1}, \tag{2.55}$$

such that the average of the lognormal variable $\nu(t) = 1$ and its standard deviation is equal to Eq. (2.55). To be able to write down the LNA in the way of Eq. (2.9), we specify the Jacobian matrix $\boldsymbol{A}$ and the diffusion matrix $\boldsymbol{B}\boldsymbol{B}^T$. The deterministic rate equations for the macroscopic concentrations $\boldsymbol{\phi}$ of the molecular species (where $\phi_1 = D$, $\phi_2 = M$, and $\phi_3 = A$) are

$$\frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}t} = \begin{pmatrix} -k_1\phi_1 + k_0(1-\phi_1) \\ v_0\phi_1 - d_0\phi_2 \\ v_1\phi_2 - d_1(1+\eta)\phi_3 \end{pmatrix}. \tag{2.56}$$

The dynamics of D^* can be eliminated as the total promoter concentration is conserved $D + D^* = 1/\Omega$. One of the conditions for the LNA is that it is valid in the limit of large system size Ω. However, as the current system is linear (i.e. it contains no bimolecular reactions), the LNA will give the exact expressions for the mean and variance independent of the value of Ω.[79] For this reason, we have chosen $\Omega = 1$ here.

The Jacobian as defined in Eq. (2.7) is then given by

$$\boldsymbol{A} = \begin{pmatrix} -(k_0+k_1) & 0 & 0 \\ v_0 & -d_0 & 0 \\ 0 & v_1 & -d_1(1+\eta) \end{pmatrix}, \tag{2.57}$$

and the diffusion matrix as defined in Eq. (2.8) for the system is given by

$$\boldsymbol{B}\boldsymbol{B}^T = \begin{pmatrix} k_1\phi_1^s+k_0(1-\phi_1^s) & 0 & 0 \\ 0 & v_0\phi_1^s+d_0\phi_2^s & 0 \\ 0 & 0 & v_1\phi_2^s+d_1(1+\eta)\phi_3^s \end{pmatrix}, \tag{2.58}$$

where ϕ_i^s are the steady-state concentrations of the active promoter, mRNA, and protein molecules, respectively. The procedure described in Eqs. (2.52)-(2.58) is the same for extrinsic noise in any other system parameter. In summary, for the three-stage model of gene expression the SDE studied in Eq. (2.9) is defined by the steady state concentrations $\boldsymbol{\phi^s}$, the Jacobian as given in Eq. (2.57), the diffusion matrix as given in Eq. (2.58), and the extrinsic noise $\nu(t)$ in d_1 as defined by Eq. (2.52) and Eq. (2.53).

The reaction rates used to obtain the results in Figure 2.1 are shown in Table 2.1 and are representative for gene expression in mammalian cells as determined by Schwanhäusser *et al.* and Suter *et al.*[27,28] Schwanhäusser *et al.* experimentally determined transcription and translation rates for over 5 000 genes in mammalian cells.[27] We selected the mode of these parameter distributions as parameter values for our linear model of gene expression. To ensure relatively fast intrinsic timescales, we chose a protein degradation and mRNA degradation rate associated with a gene that was classified as having both unstable protein and mRNA (see Figure 5 in [27]), while also enforcing that the protein lifetime is much longer than the mRNA lifetime. Promoter activation and deactivation rates were chosen as the average over various mammalian genes as measured by Suter *et al.*[28]

In the three-stage gene expression model, the three intrinsic timescales are $1/(k_0+k_1)$, $1/d_0$, and $1/d_1$, corresponding to the lifetimes of the molecular species D, M, and A respectively. The timescale of the extrinsic noise process is $\tau = K^{-1}$ as defined in Eq. (2.53). The ratio of the extrinsic and intrinsic timescales is then

$$\lambda = \tau / \max(\tau_{\text{int}}), \tag{2.59}$$

where $\max(\tau_{\text{int}})$ is the longest intrinsic timescale in the model, here $1/d_1 = 15\,625$ s. To see for which values of λ our method gives good results, we simulate the model in Figure 2.1 with the mRNA degradation rate d_0 subject to extrinsic noise with various values of τ using the Extrande algorithm. Figure 2.1(b) shows that for longer extrinsic correlation times τ (corresponding to larger values of λ), the stochastic simulations (denoted "SSA" in Figure 2.1) approach the analytically calculated mean number of protein A, given by $\langle\phi_3^s\rangle$ (see Eq. (2.36)). For both extrinsic noise magnitudes $CV = 0.1$ and $CV = 0.25$, timescale separation conditions are

Table 2.1: Parameter values used for the gene expression model.

parameter	value (s^{-1})
k_0	0.00085
k_1	0.0017
v_0	0.00028
v_1	0.028
d_0	0.00019
d_1	0.000064

satisfied for $\lambda \approx 10$ and we are able to accurately predict the mean number of proteins. This corresponds to an extrinsic timescale of $\tau = 10^5$ s, which is of the same order of magnitude as the period of the mammalian cell cycle (approximately 27.5 h[27]).

Noise in different parameters can affect the system in different ways. Figure 2.1(c) shows that the mean protein number can decrease (k_0), increase (d_0, d_1, k_1), or remain constant (v_0, v_1) in response to increasing extrinsic noise magnitude (CV). The total variance in protein A, denoted by $V(x_3^s)$ (see Eq. (2.12)), always increases when a system is affected by extrinsic noise (Figure 2.1(d)). Extrinsic noise in a parameter that controls the lifetime of molecular components, such as the mRNA or protein degradation rate, might cause intrinsic and extrinsic timescales to mix and thus have non-trivial effects. This can be seen, for example, when extrinsic fluctuations are applied to the promoter activation rate k_0, causing a decrease in the intrinsic variance, denoted by $V(\xi_3)$ (Figure 2.1(e)). The potential of extrinsic fluctuations to decrease the intrinsic noise in gene expression was already noted by Shahrezaei *et al.*[62]

2.3.2 Regulated gene expression

Feedback control is often proposed as a possible strategy to reduce gene expression noise, and therefore it is considered to have great potential for the design of robust synthetic gene regulatory networks.[65] Several studies have confirmed that such strategies are indeed capable of reducing the variability in protein concentrations as well as influencing the number of modes of the protein number distribution.[32,65,81–85] To study the potential of negative feedback to reduce protein noise in a system subject to extrinsic noise, we adapt the genetic feedback loop model proposed by Grima *et al.*[42] In this model, the mRNA dynamics are omitted, which is a valid assumption in the absence of translational bursting.[45,86] The genetic feedback model contains a negative feedback loop where the gene product can bind to the promoter area, thus preventing protein production (Figure 2.2(a)).

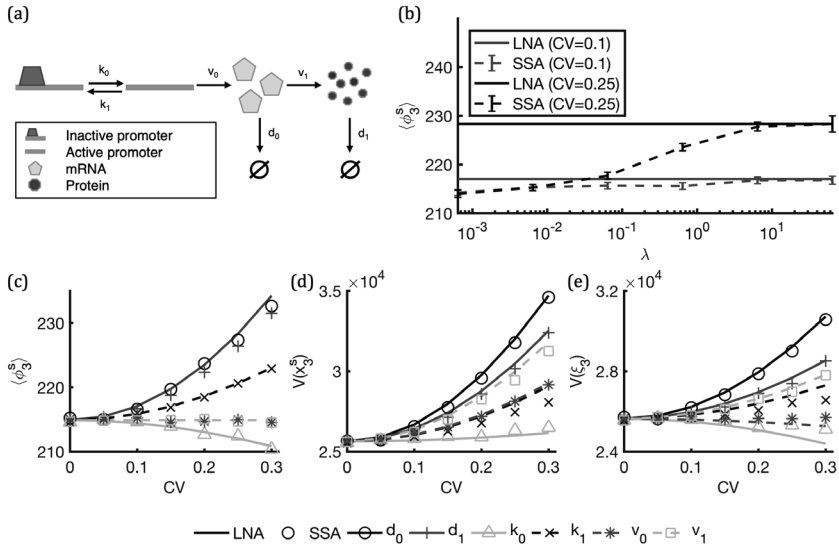


Figure 2.1: (a) Model of gene expression where the promoter can switch between the active and inactive state. (b) Validity of the timescale separation condition depends on the ratio of extrinsic and intrinsic timescales λ and the magnitude of the extrinsic noise denoted by the CV (here in parameter d_0). The mean number of protein A at steady state, denoted by $\langle \phi_3^s \rangle$, with grey lines corresponding to extrinsic noise with $CV = 0.1$, and black lines corresponding to extrinsic noise with $CV = 0.25$. Solid lines are analytical predictions (denoted by "LNA"), dashed lines are stochastic simulation results (denoted by "SSA") from the Extrande algorithm.[47] 95% confidence intervals for the mean were calculated following Brančík *et al.* in.[80] (c) Effect of extrinsic noise in different parameters on the mean number of protein A. Analytical predictions are denoted by lines (d_0 (black line), d_1 (dark grey line), k_0 (light grey line), k_1 (dashed black line), v_0 (dashed dark grey line), and v_1 (dashed light grey line)), stochastic simulation results by markers (d_0 (circles), d_1 (plus sings), k_0 (triangles), k_1 (crosses), v_0 (asterisks), and v_1 (squares)). We calculate terms up to sixth order in (substitute u = 3 in Eq. (36)). Note that the analytical prediction for extrinsic noise in parameters d_0 and d_1 overlap, as well for v_0 and v_1. We calculate terms up to sixth order in ϵ (substitute $u = 3$ in Eq. (2.36)). (d) Effect of extrinsic noise in different parameters on the total variance of protein A, given by $V(x_3^s)$ (see Eq. 2.12). We calculate terms up to second order in ϵ (substitute $u = 1$ in Eqs. (2.38), (2.37)). Line colours and marker styles correspond to extrinsic noise in the same parameters as in (c). (e) Effect of extrinsic noise in different parameters on the intrinsic variance of protein A, given by $V(\xi_3)$ (see Eq. 2.12). We calculate terms up to second order in ϵ (substitute $u = 1$ in Eq. (2.37)). We use the dual reporter technique [59] to compute estimates of the intrinsic variance from stochastic simulations. Line colours and marker styles correspond to extrinsic noise in the same parameters as in (c).

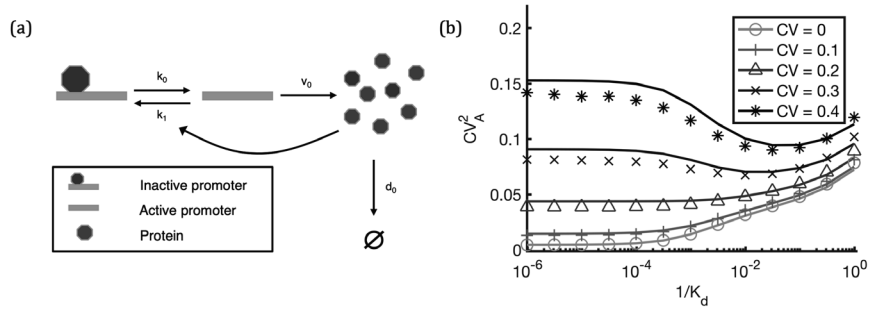


Figure 2.2: (a) Model of autoregulatory gene expression. The feedback strength $1/K_d = k_1/(k_0\Omega)$ is determined by the binding affinity of the protein to the promoter. (b) Analytical predictions (solid lines) and stochastic simulation results (markers) for the CV_A^2 of the number of molecules of protein A as a function of feedback strength $1/K_d$ for extrinsic noise in protein production rate v_0 ranging in magnitude from no extrinsic noise ($CV = 0$, light grey lines and circles) to strong extrinsic noise ($CV = 0.4$, black lines and asterisks). The analytical solution for the mean was calculated up to sixth order in ϵ (substitute $u = 3$ in Eq. (2.36)), variances were calculated up to second order in ϵ (substitute $u = 1$ in Eqs. (2.38), (2.37)).

This system consists of the following reactions

$$D \xrightarrow{v_0} D + A \underset{k_0}{\overset{k_1}{\rightleftharpoons}} D^*, \qquad A \xrightarrow{d_0} \emptyset. \tag{2.60}$$

Here, D denotes the unbound promoter, D^* the bound, inactive promoter, and A the protein. The values for the parameters k_0, v_0, and d_0 (Table 2.2) were chosen so that they are consistent with those in the linear gene expression model, as follows. Since the mRNA dynamics are omitted, d_0 now refers to the protein degradation rate. To ensure the same steady-state species concentrations as in the unregulated gene expression model, we choose the protein production rate as

$$v_0 = \phi_2^s v_1, \tag{2.61}$$

where ϕ_2^s refers to the macroscopic steady-state concentration of mRNA in the linear gene expression model, respectively, and v_1 is equal to the parameter value of the translation rate of the linear gene expression model as stated in Table 2.1. The probability that the binding reaction occurs in a small time interval is proportional to k_1/Ω, where k_1 is the protein-DNA binding rate constant and Ω is the cell volume (approximately 2 picolitres[87]). We vary the binding rate k_1 of protein A to the promoter over a range of biologically relevant specificities[88] and define the

Table 2.2: Parameter values used for the autoregulatory gene expression model.

parameter	value (s^{-1})
k_0	0.00085
k_1/Ω	$8.3 \times 10^{-10} - 8.3 \times 10^{-4}$
v_0	0.014
d_0	0.000064

feedback strength as the inverse of the non-dimensional dissociation constant K_d

$$\frac{1}{K_d} = \frac{k_1}{k_0 \Omega}. \tag{2.62}$$

We note that for very small values of $1/K_d$, the system is weakly non-linear since protein binding becomes a rare event compared to promoter activation.

In a negative feedback system with no extrinsic noise, it is well known that the noise increases with $1/K_d$.[89–91] This is confirmed in Figure 2.2(b), where the blue lines and circles ($CV = 0$) show that the noise in the protein population, quantified by its coefficient of variation squared CV_A^2, increases with increasing negative feedback strength. We consider the case of extrinsic noise affecting the protein production rate v_0 and show in Figure 2.2(b) how this changes the noise in the number of proteins A. Slow fluctuations in the protein production rate could for example be the result of changes in the cell's mitochondrial or ribosomal content over time.[92] While adding extrinsic noise increases protein noise, as the extrinsic noise magnitude increases beyond a threshold ($CV \sim 0.3$) the magnitude of protein noise does not increase monotonically with $1/K_d$. Rather, it has a minimum at $1/K_d \approx 10^{-1}$. This finding is supported by experimental data that shows that negative autoregulation mechanisms are able to negate the effects of slow extrinsic noise.[68] In the absence of extrinsic noise, the variance of protein numbers only has an intrinsic component. The magnitude of intrinsic noise is influenced the average expression level of molecular species and the response time of the system, which is the time it takes for any initial perturbation to decay and the system to return to its equilibrium. In one instance, protein expression levels decrease monotonically with increasing negative feedback strength, and smaller molecular numbers are associated with higher intrinsic noise. In the second, negative feedback is known to speed up the response time of a system, leading to the attenuation of protein noise .[93] These effects can be seen from the blue line and circles in Figure 2.2(b) indicate that smaller values of $1/K_d$ result in a less noisy system when the system is not subject to extrinsic noise, which implies that the decrease of intrinsic noise due to a faster response time cannot compensate for the increase in noise resulting from smaller protein levels. The extrinsic contribution to the protein noise is also a function of the response time of the system,[91] which causes the

extrinsic component of the protein noise to decrease monotonically with increasing feedback strength. As the CV of the extrinsic noise source increases above 0.3 (Figure 2.2(b), purple and green line and circles), the reduction in extrinsic noise is larger than the increase in intrinsic noise up until $1/K_d \approx 10^{-1}$. As the negative feedback strength increases further, the reduction of extrinsic noise is negated by an increase in intrinsic noise due to decreasing protein levels, resulting in an increase of CV_A^2.

2.3.3 Signal transduction and extrinsic noise

Cells are embedded in highly fluctuating environments. It is vital for biological systems that they can sense external stimuli and process this information in order to adapt to their environment accordingly. Information theoretic approaches have, for example, been applied to biological systems to address the question of how well a network subject to biochemical noise is able to transmit information that arrives at cell receptors into the intracellular environment.

In order to analyse the effects of extrinsic noise on the signal transduction process, we consider the simple two-stage gene expression model shown in Figure 2.3(a), which contains mRNA molecules M and protein molecules A, the concentrations of which fluctuate over time

$$\emptyset \xrightarrow{v_0} M \xrightarrow{d_0} \emptyset, \qquad M \xrightarrow{v_1} M + A, \qquad A \xrightarrow{d_1} \emptyset. \tag{2.63}$$

We again choose parameters for this motif such that they are consistent with those in the linear gene expression model. The values for v_1, d_0 and d_1 are the same as in Table 2.1, while the mRNA production rate v_0 is calculated as

$$v_0 = \phi_1^s v_0^*, \tag{2.64}$$

where ϕ_1^s corresponds to the macroscopic steady-state concentration of the active promoter in the linear gene expression model and the asterisk refers to parameters as stated in Table 2.1.

We are interested in how the rate of information transfer from mRNA to protein is affected by an extrinsic noise source that perturbs the translation process (extrinsic fluctuations in v_1). Fluctuations in the rate of translation arise because this process is dependent on the number of free ribosomes in the cytoplasm, which changes stochastically over time. To quantify how well networks can transmit information in noisy environments, we can make use of the mutual information rate (MIR) as a metric. Calculation of the mutual information of trajectories is a challenging task,[94,95] but for linear Gaussian processes an expression involving the power spectra of continuous-time input signal $s(t)$ and output signal $x(t)$ can be

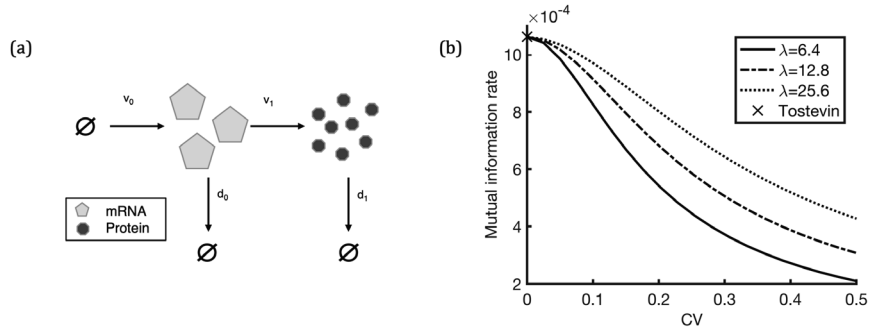


Figure 2.3: (a) Model of the two-stage model of gene expression, where information about the concentration of mRNA (input) is transduced to the concentration of protein (output). (b) Mutual information rate between input and output variables as a function of extrinsic noise magnitude CV for extrinsic noise in parameter v_1 with different correlation times, corresponding to a ratio of timescale separation λ between intrinsic and extrinsic processes of 6.4 (solid line), 12.8 (dash-dot line), and 25.6 (dotted line). The black cross denotes the result by Tostevin *et al.*[94] in the absence of extrinsic noise. The analytical solution of the power spectrum was calculated up to second order in ϵ (substitute $u = 1$ in Eqs. (2.42), (2.48)).

derived:

$$R(s(t), x(t)) = -\frac{1}{4\pi}\int_{-\infty}^{\infty} \ln\left[1 - \frac{|P_{sx}(\omega)|^2}{P_{ss}(\omega)P_{xx}(\omega)}\right] \mathrm{d}\omega. \tag{2.65}$$

Here, $P_{sx}(\omega)$ is the cross-power spectrum of $s(t)$ and $x(t)$ (the Fourier transform of the cross-correlation function of $s(t)$ and $x(t)$), and $P_{ss}(\omega)$ and $P_{xx}(\omega)$ are the power spectra of $s(t)$ and $x(t)$, respectively. Note that for non-Gaussian and/or non-linear systems this expression provides a lower bound for the channel capacity.[96] We consider fluctuations in the mRNA concentration as the input signal, and fluctuations in the protein concentration as the output signal. In,[94] Tostevin *et al.* have derived the analytical expression for the MIR of the input and output trajectories of the motif in Figure 2.3(a) in the absence of extrinsic noise. We extend this result by considering an external process that affects the translation process, causing the parameter v_1 to fluctuate over time. By substituting the analytical expressions for the (cross-)power spectra resulting from Eq. (2.23) into Eq. (2.65), we are able to quantify how the accuracy of information transmission from mRNA to protein concentration is affected by this disrupting process.

Figure 2.3(b) shows that in the absence of extrinsic noise, our approximate solution agrees with the solution by Tostevin *et al.* (marked by the black cross).[94] The presence of extrinsic noise decreases the fidelity of information processing, and the extent of this effect depends on both the extrinsic noise magnitude CV and the ratio of timescale separation λ, as defined in Eq. (2.59), between the intrinsic

Table 2.3: Parameter values used for the information processing motif.

parameter	value (s^{-1})
v_0	0.000093
v_1	0.028
d_0	0.00019
d_1	0.000064

and extrinsic processes. Extrinsic noise sources with long extrinsic correlation times (large λ) are less disruptive than those with shorter correlation times (smaller λ). This implies that slowly fluctuating environmental variables have a smaller negative effect on the MIR than external variables that fluctuate quickly. This result is intuitive, as in the extreme case where faster fluctuating extrinsic processes happen on roughly the same timescale as the intrinsic processes, it may prove harder to distinguish between signal and noise and information might be lost. In the limit of infinitely slowly fluctuating external variables, the value of the external variable is constant with respect to the intrinsic timescale, and will have no effect on the signal transduction process regardless of the noise magnitude.

2.3.4 Robustness of feed-forward loop motifs

Feed-forward loops (FFLs) are capable of responding in a precise, robust manner to external signals.[97] These motifs are defined by a gene X that regulates a second gene Y. Both X and Y then regulate a target gene Z and are very common networks, appearing in many organisms, including *E. coli*, *B. subtilis*, *S. cerevisiae*, *D. melanogaster*, *C. elegans*, and humans.[98] There are multiple types of FFLs (see Figure 2.4(a)) since regulation can take place either through activation or repression, and much effort has been devoted to extract the general features of each one. However, this is not straightforward because both the transient and equilibrium behaviour is characteristic of a particular system.[99] For these reasons, constructing an optimal system that fulfils certain design requirements can be a considerable computational task. Here, we aim to present an efficient optimisation scheme to generate optimal parameters for a FFL to ensure it responds in a precise manner to input signals but remains robust to noise. To do this, we devise two objective functions that quantify both the dynamic as well as the stochastic behaviour of the system. We aim to generate parameter sets for the network such that the system responds to a switching-on of the input signal X by a negative pulse in the concentration of Z, before going back to its original steady state (black line, Figure 2.4(b)). Moreover, the variation around this steady state concentration of Z for a noisy input signal X should be minimal. Taking these requirements into

account, the form of the objective function is

$$S = w_1 c_{ODE} + w_2 c_{LNA}, \tag{2.66}$$

where w_1 and w_2 are the weights of the respective objective functions c_{ODE} and c_{LNA}, with $w_1 + w_2 = 1$ and

$$c_{ODE} = \sum_{i=1}^{5} q_i s_i, \tag{2.67}$$

$$c_{LNA} = CV_Z^2 = \frac{V(x_3^s)}{\langle x_3^s \rangle^2}, \tag{2.68}$$

where x_3^s is the steady-state concentration of protein Z and

$$s_1 = (\phi_1(t_f) - \phi_1(t_f - \Delta t))^2, \tag{2.69}$$

$$s_2 = (\phi_2(t_f) - \phi_2(t_f - \Delta t))^2, \tag{2.70}$$

$$s_3 = (\phi_3(t_f) - \phi_3(t_f - \Delta t))^2, \tag{2.71}$$

$$s_4 = (\phi_3(t_f) - \phi_3(t_0))^2, \tag{2.72}$$

$$s_5 = \frac{\min(\phi_3)}{\langle \phi_3 \rangle}, \tag{2.73}$$

with ϕ_i the macroscopic concentrations of the proteins X, Y, and Z respectively, $q_i = \frac{1}{5}$, $i = 1, \ldots, 5$ the subweights of each ODE (ordinary differential equation) objective s_i, $t_0 = 0$ the initial time point, $t_f = 1000$ the final time point of the simulation, and $\Delta t = 5$. To obtain c_{ODE} we simulate the FFLs using an ODE solver, where Eqs. (2.69)–(2.71) ensure each of the three system components reaches steady state, Eq. (2.72) ensures that Z reaches pre-input concentration, and Eq. (2.73) aims to produce a significant drop in the concentration of Z upon a change in input X (Figure 2.4(b)). The score c_{LNA} is obtained from our analytical solution (Eq. (2.11) and Eq. (2.12)).

We used the most general model of the FFL from Macía *et al.* in,[99] that is able to describe all eight FFL topologies. The macroscopic rate equations for this model are given by

$$\frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}t} = \begin{pmatrix} \alpha_0(1+\eta) - d_0\phi_1 \\ \alpha_1\left(\frac{1+\beta_0 K_1\phi_1}{1+K_1\phi_1}\right) - d_1\phi_2 \\ \alpha_2\left(\frac{1+\beta_1 K_2\phi_1 + \beta_2 K_3\phi_2 + \beta_3 K_2 K_3\phi_1\phi_2}{1+K_2\phi_1+K_3\phi_2+K_2K_3\phi_1\phi_2}\right) - d_2\phi_3 \end{pmatrix}. \tag{2.74}$$

In this model, α_i describes the basal production of the proteins X, Y, and Z, and d_i denotes the degradation rate of a species. The type of regulatory interaction

between the regulator gene and the gene it targets is described by parameters β_i, where values $\beta_i < 1$ correspond to an inhibitory interaction as the production rate decreases proportional to the basal level, whereas $\beta_i > 1$ describes activation.[99] K_i describes the binding equilibrium of the regulator with the gene it targets. Extrinsic noise enters the model in the production rate of X, α_0, with a magnitude of $CV = 0.5$. We fix the ratio of basal production/degradation ($\alpha_i/d_i = 100$) of all three molecular species to ensure that all species are sufficiently abundant for the LNA to hold. Initial conditions are $X(t_0) = 0$, $Y(t_0) = Z(t_0) = 100$. We generate 2×10^6 random parameter sets for the 10 remaining system parameters and optimise for the objective score S for the cases $\{w_1 = 1, w_2 = 0\}$ and $\{w_1 = 0.5, w_2 = 0.5\}$ (for details on the optimisation procedure, see Section 2.5.2). The 0.01% top scoring parameter sets are then selected for either set of objective scores, and their corresponding topology is determined. The results in Table 2.4 show that only three of eight possible topologies, shown in the dashed boxes in Figure 2.4(a), are present in the results when only the ODE objective function is considered ($w_1 = 1$), and that cFFL4 is the most prevalent motif. If both the ODE and LNA objectives are given equal importance ($w_1 = w_2 = 0.5$), then the cFFL3 and iFFL1 topologies are also present among the top-scoring parameter sets (solid line boxes, Figure 2.4(a)). However, since these do not occur when only considering the ODE criteria, this implies that they are not very suitable to give the desired dynamic behaviour. In addition, Macía *et al.* also find that the iFFL1 motif is not capable of producing the desired negative pulse.[99]

Since the cFFL4 topology is the most prevalent, this motif has the highest probability to produce the desired system behaviour. For this reason, we perform an optimisation within the local parameter space of the cFFL4 motif for 2×10^4 randomly generated parameter sets and select the 1 500 best-scoring sets. Figure 2.4(c) shows how the parameter space changes with the addition of the c_{LNA} objective. If only the robustness of the network to noise is considered ($w_1 = 0$, blue circles), no specific parameter values for the parameters K_1 and d_2 are preferred. When only the dynamic behaviour of the FFL is prioritised ($w_1 = 1$, yellow asterisks), high values of K_1 and d_2 more likely result in the desired dynamics. Interestingly, when both the c_{ODE} and c_{LNA} objectives are given equal weight ($w_1 = 0.5$, red downward-pointing triangles) the parameter space is further constrained compared to optimising for a single objective. Thus, for robust FFLs the degradation rate of Z and the binding equilibrium of protein X to the gene that produces Y needs to be tuned.

2.4 Discussion

In this work, we have derived an analytical framework to quantify the contribution of coloured extrinsic noise to fluctuations in gene expression. We have shown that

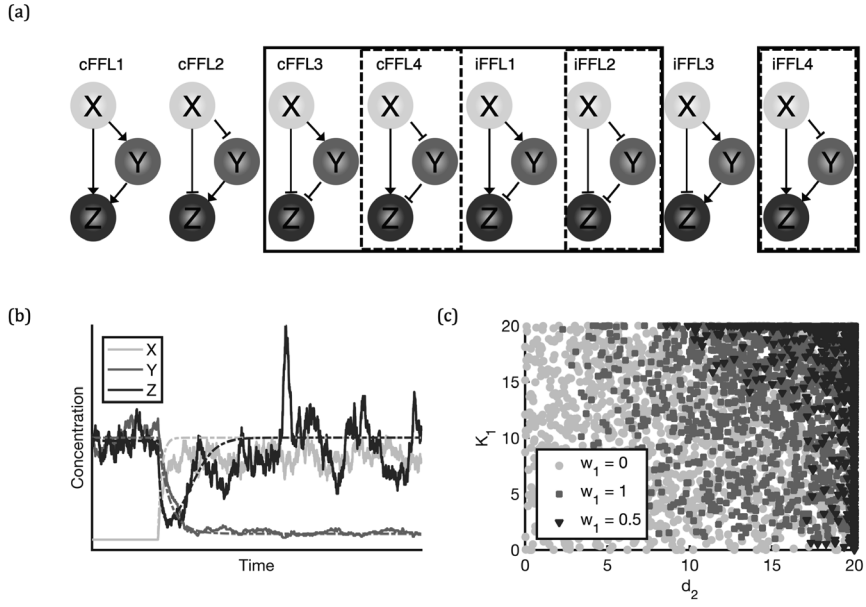


Figure 2.4: (a) All eight feed-forward loop topologies, where the arrow indicates the type of regulation (activation or repression). The motifs enclosed by a dashed line denote the subset of optimal FFL topologies for the case of $\{w_1 = 1, w_2 = 0\}$ in Eq. (2.66), whereas the motifs enclosed by a line denote the subset of optimal topologies when $\{w_1 = 0.5, w_2 = 0.5\}$ (See Table 2.4). (b) Sample result of the FFL dynamic behaviour of our optimisation scheme. After the production of protein X is induced, this produced a drop in the concentration of protein Z. The concentration of Z then recovers to its original steady-state level. Solid lines denote example stochastic simulations of the concentration of proteins X (light grey), Y (medium grey), and Z (black) over time, while dashed lines correspond to the respective deterministic dynamics. (c) How the parameter space of parameters K_1 and d_2 is constrained by considering an additional objective function for the system's stochastic behaviour. $w_1 = 0$ (light grey circles) refers to the case where only the dynamic behaviour is prioritised, $w_1 = 1$ (medium grey squares) refers to the case where only the stochastic behaviour is prioritised, and $w_1 = 0.5$ (black downward-pointing triangles) refers to the case where both objectives are given equal importance.

Table 2.4: Occurrence of FFL topologies of top (0.01%) scoring parameter sets out of 2×10^6 parameter sets.

topology	$w_1 = 1$	$w_1 = 0.5$
iFFL2	1%	2%
cFFL4	75%	39.5%
iFFL4	24%	1%
cFFL3	0%	14.5%
iFFL1	0%	43%

when the conditions underlying the theory are satisfied, we are able to accurately describe the mean, variance and power spectrum of molecule number fluctuations subject to both intrinsic and extrinsic noise sources. Using several examples, we show that the theory is relevant in a wide range of applications, and can be used to distil the principles underlying fundamental system behaviour, noise sources, and information processing in biochemical networks.

Our framework relies on three main approximations: the linear-noise approximation, the separation between timescales of the intrinsic and extrinsic fluctuations, and the small extrinsic noise expansion. First, the LNA will give an accurate approximation of the CME when the molecular species populations are sufficiently large, when the nonlinearity in the reaction rates is sufficiently weak or else for special classes of biochemical systems (;[79] see discussion later). Second, our theory requires the extrinsic fluctuations to be slow with respect to the system's intrinsic dynamics. As most intrinsic processes happen on the timescale of seconds or sub-seconds and extrinsic fluctuations have a typical correlation time corresponding to the cell cycle period (many minutes), we expect that our timescale separation assumption is reasonable. The assumption of small extrinsic noise might appear at the first sight very limiting, however, in practice we find that the approximation yields sufficiently accurate results for noise magnitudes at least as large as $CV = 0.25$ for the mean protein number (Figure 2.1(b)) and variances (Figure 2.1(d)) of the linear model and $CV = 0.4$ for the means and variances of the regulated gene expression model (Figure 2.2(b)).

A more general issue with studying stochastic systems is computational speed. To obtain statistics of biochemical systems subject to both intrinsic and extrinsic noise with reasonable confidence levels, one needs to simulate many trajectories of the system for a considerable time. The advantage of obtaining closed-form expressions for these statistics is that parameter values simply need to be substituted, and there is no need to re-evaluate the system. For example, it takes approximately 3 hours to generate one data point (100 trajectories of 10^8 seconds each) for the model in Figure 2.1(a), whereas evaluating the analytical expressions for the mean and variance of each molecular component takes less than a second on a

typical desktop computer. As pointed out in Section 2.2.3, the limiting step in the automated sum evaluation is the calculation of the stationary state ϕ^s. This task becomes increasingly computationally intensive for more complex systems. For this reason, the limit of complexity that we can study is determined by the computational power available. For specific cases, a more specialised external solver could be employed to accelerate this task.

Due to this speed-up, we have been able to use the analytical framework to perform a computationally efficient multi-objective optimisation of FFLs. With this optimisation routine, we are able to explore both network topology and parameter space to generate systems with optimal dynamic and stochastic features, which is generally infeasible for non-trivial systems using simulation-based approaches. Our analysis shows that even in simple networks such as FFLs, there exists a complex relationship between system structure and function. With this optimisation scheme we are able to quickly generate recommendations for an optimal network topology and parameter ranges. Compared to optimisation using stochastic simulation algorithms, this optimisation scheme gives an improvement in computational time of several orders of magnitude. The results of the optimisation scheme suggest that not every FFL motif is capable of producing a specific dynamic response, and that not all FFL types have the same extrinsic noise tolerance. Although there does not appear to be a trade-off between these two objectives, choosing optimal networks and their parameters such that they fulfil both requirements can be a substantial task given the high dimensionality of the problem. In this case, combining deterministic dynamics with stochastic analysis of equilibrium behaviour is an efficient and effective approach.

Since analytical expressions for both intrinsic and extrinsic contributions to variability can be obtained, the proposed method allows a systematic analysis of how changing the properties of extrinsic fluctuations affects intrinsic variability and total noise in gene expression models. Similarly, these expressions can provide predictions on which network parameters are susceptible to perturbations and contribute to high variability and can be used as a tool for stochastic sensitivity analysis. Such a method can be of interest for synthetic biology applications as it could provide universal design principles for network construction that exploits (suppresses) the positive (negative) effects of cellular stochasticity.

The framework developed here rests on the validity of the linear-noise approximation first and foremostly. This limits the current approach to analysis of nonlinear biochemical systems with large numbers of molecules in all species or else to those systems with arbitrary number of molecules but weakly nonlinear reaction rates. However, we note that the linear-noise approximation has been, over the past decade, extended to estimate the first and second moments of the molecule number distributions of nonlinear biochemical systems in which one or more molecular species is present in low copy numbers.[100–103] The corrections to the LNA power spectrum of fluctuations due to low molecule numbers have also been systematically

studied.[104,105] Hence by starting from these frameworks and repeating the same analysis as we performed here, i.e., applying the assumption of timescale separation between intrinsic and extrinsic noise and subsequently assuming small extrinsic noise, would likely result in a new theory which overcomes the major limitations of the present approach.

In conclusion, we have proposed a fast, systematic analytical framework to assess the effects of coloured environmental noise on biochemical systems. We have shown that the mathematical framework provides accurate predictions of system characteristics for a wide range of biological networks. Given the speed and flexibility of our approach, the research community can now further access the sources of variability in gene expression data. This will lead to a better understanding of how biological systems exploit or suppress environmental signals. There is, thus, the potential to uncover new design principles to aid the construction of new, robust *in vivo* synthetic circuits.

2.5 Methods

2.5.1 Stochastic simulations

Stochastic simulations were performed using the Extrande algorithm by Voliotis *et al.*[47] implemented in `C++11`. Each stochastic simulation data point in Figures 2.1 and 2.2 was obtained from 100 trajectories of 10^8 seconds each.

2.5.2 Optimisation routine

All optimisations were performed using the L-BFGS-B algorithm (`fmin_l_bfgs_b()`) provided in the SciPy library.[106] Parameters were generated from a uniform distribution with bounds $[0.0001, 20]$.

Acknowledgements

E.K. and C.F. were supported by HFSP Research grant RGP0025/2013. R.G. was supported by BBSRC grant BB/M025551/1. C.F. acknowledges fruitful discussions with Benjamin Lindner.

Appendices

2.A Construction of the noise variables

To introduce extrinsic fluctuations to a rate constant c_k, we multiply it with a lognormal variable $\bar{\nu}_k(t)$,

$$\bar{c}_k(t) = c_k \bar{\nu}_k(t) = c_k \left(1 + \bar{\eta}_k(t)\right). \tag{2.75}$$

We require $\langle \bar{c}_k(t) \rangle = c_k$ and also define a shifted stochastic variable $\bar{\eta}_k(t)$ that will be needed in the small noise expansion. The lognormal variables $\bar{\nu}_k(t)$ can be constructed from normally distributed variables $\bar{\mu}_k(t)$ with variances $\langle \bar{\mu}_k \, \bar{\mu}_k \rangle = \epsilon_k^2$ and inverse correlation times K_k. The latter may be described by the Ornstein-Uhlenbeck process[73]

$$\mathrm{d}\bar{\mu}_k(t) = -K_k \bar{\mu}_k(t) \, \mathrm{d}t + \sqrt{2K_k}\epsilon_k \, \mathrm{d}W(t). \tag{2.76}$$

We define the lognormally distributed stochastic variable by

$$\bar{\nu}_k(t) = \exp\left(\bar{\mu}_k(t) - \frac{1}{2}\epsilon_k^2\right) \tag{2.77}$$

and use Wick's theorem[107] to calculate its mean (only even powers in $\bar{\mu}_k^c$ do not vanish):

$$\langle \bar{\nu}_k \rangle = \exp\left(-\frac{1}{2}\epsilon_k^2\right) \sum_{c=0}^{\infty} \frac{1}{(2c)!} \left\langle \bar{\mu}_k^{2c} \right\rangle = \exp\left(-\frac{1}{2}\epsilon_k^2\right) \sum_{c=0}^{\infty} \frac{1}{2^c c!} \underbrace{\langle \bar{\mu}_k \, \bar{\mu}_k \rangle}_{\epsilon_k^2}^{c} = 1 \tag{2.78}$$

in agreement with the requirement $\langle \bar{c}_k(t) \rangle = c_k$. The shifted lognormal stochastic variable is then

$$\bar{\eta}_k(t) = \exp\left(\bar{\mu}_k(t) - \frac{1}{2}\epsilon_k^2\right) - 1. \tag{2.79}$$

2.B Arbitrary time correlation functions of lognormal stochastic variables

Given a set of independent normally distributed stochastic variables $\{\bar{\mu}_k\}$, a tuple of indices $(r_1, \ldots, r_n)$ and a tuple of times $(t_1, \ldots, t_n)$ we define $\mu_k \equiv \bar{\mu}_{r_k}(t_k)$, and the two-point time correlation functions Δ_{ij} for $1 \leq i < j \leq n$,

$$\Delta_{ij} \equiv \langle \mu_i \mu_j \rangle \tag{2.80}$$

for all $i < j$. We consider a smooth function and its derivatives

$$y_k \equiv V(\mu_k) = \sum_{c=0}^{\infty} \frac{a_c}{c!} \mu_k^c \,, \quad y_k^{(m)}(\mu_k) \equiv \frac{\mathrm{d}^m V(\mu_k)}{\mathrm{d}\mu_k^m} = \sum_{c=m}^{\infty} \frac{1}{(c-m)!} a_c \, \mu_k^{c-m} \tag{2.81}$$

to derive the generalisation of the $n = 2$ result by Malakhov[74] for the n-point time correlation function

$$\langle y_1 \ldots y_n \rangle = \sum_{c_1=0}^{\infty} \cdots \sum_{c_n=0}^{\infty} \frac{a_{c_1} \ldots a_{c_n}}{c_1! \ldots c_n!} \langle \mu_1^{c_1} \ldots \mu_n^{c_n} \rangle \,. \tag{2.82}$$

According to Wick's theorem,[107] the correlation functions $\langle \mu_1^{c_1} \ldots \mu_n^{c_n} \rangle$ decompose into sums of partitions into two-point correlation functions and they are zero for odd n. We apply the theorem partially to isolate the two-point correlation functions $\Delta_{ij} = \langle \mu_i \mu_j \rangle$ with $i < j$,

$$\langle y_1 \ldots y_n \rangle = \sum_{l_1=0}^{\infty} \cdots \sum_{l_n=0}^{\infty} \sum_{u=0}^{\infty} \sum_{|\boldsymbol{d}^n|=u} \left(\frac{c_1! \ldots c_n!}{l_1! \ldots l_n! \prod\limits_{i<j} d_{ij}!} \frac{a_{c_1} \ldots a_{c_n}}{c_1! \ldots c_n!} \langle \mu_1^{l_1} \rangle \ldots \langle \mu_n^{l_n} \rangle \prod_{\substack{i,j=1 \\ i<j}}^{n} \Delta_{ij}^{d_{ij}} \right) \tag{2.83}$$

where the inner sum is taken over all tuples $\boldsymbol{d}^n = (d_{12}, d_{13}, d_{23}, \ldots, d_{(n-1)\,n})$ with $|\boldsymbol{d}^n| = \sum d_{ij} = u$. There are $c_k!/l_k!$ possibilities to assign $m_k = c_k - l_k$ from a total of c_k factors μ_k to the u Δ_{ij} pairs. However, to obtain the number of *different* partitions into the pairs, the product of the former must be divided by the product of $d_{ij}!$ permutations of d_{ij} identical factors Δ_{ij}. We notice that $l_k = c_k - m_k$ to recognise the derivatives $y_k^{(m_k)}$ from Eq. (2.81) so finally

$$\langle y_1 \ldots y_n \rangle = \sum_{u=0}^{\infty} \sum_{|\boldsymbol{d}^n|=u} \left(\prod_{k=1}^{n} \left\langle y_k^{(m_k)}(\mu_k) \right\rangle \prod_{\substack{i,j=1 \\ i<j}}^{n} \frac{\Delta_{ij}^{d_{ij}}}{d_{ij}!} \right). \tag{2.84}$$

For $n = 2$ with $k = m_1 = m_2 = d_{12}$ and $B_\mu[\tau] = \Delta_{12}(\tau) = \langle \mu(0)\mu(\tau)\rangle$ we recover the result by Malakhov,[74]

$$B_y[\tau] = \langle y(0)y(\tau)\rangle - \langle y\rangle^2 = \sum_{k=1}^{\infty} \frac{1}{k!} \langle y^{(k)}(\mu)\rangle^2 B_\mu^k[\tau]. \tag{2.85}$$

Evaluation for normal stochastic variables With normally distributed $y_k = \mu_k$ (mean 0) the term $\langle y_k^{(m_k)}(\mu_k)\rangle$ is 1 for $m_k = 1$ and 0 else. Consequently, $m_k = 1$ for $k \in \{1, \ldots, n\}$, that is each index must occur exactly once in the product of two-point correlation functions in Eq. (2.84) so also all d_{ij} are 1 and Wick's theorem is recovered.

Evaluation for lognormal stochastic variables Lognormally distributed $y_k = \nu_k$ with mean 1 is invariant under differentiation with respect to μ_k (see Eq. 2.77) so the term $\langle y_k^{(m_k)}(\mu_k)\rangle$ becomes identical 1 for all m_k. This leads to significant simplification of Eq. (2.84) and we use the multinomial theorem to obtain

$$\langle \nu_1 \ldots \nu_n\rangle = \sum_{u=0}^{\infty} \sum_{|\boldsymbol{d}^n|=u} \left(\prod_{\substack{i,j=1 \\ i<j}}^{n} \frac{\Delta_{ij}^{d_{ij}}}{d_{ij}!} \right) = \sum_{k=0}^{\infty} \frac{1}{k!} \left(\sum_{\substack{i,j=1 \\ i<j}}^{n} \Delta_{ij} \right)^k = \exp\left(\sum_{\substack{i,j=1 \\ i<j}}^{n} \Delta_{ij} \right). \tag{2.86}$$

Evaluation for shifted lognormal stochastic variables The mean of the stochastic variables η_k in Eq. (2.79) is 0 and all derivatives with respect to μ_k are identical to ν_k in Eq. (2.77) with mean 1. Therefore, the product of $\langle y_k^{(m_k)}(\mu_k)\rangle$ terms in Eq. (2.84) vanishes if $m_k = 0$ for any $k \in \{1, \ldots, n\}$ and is 1 else. The final result is

$$\langle \eta_1 \ldots \eta_n\rangle = \sum_{u=0}^{\infty} {\sum_{|\boldsymbol{d}^n|=u}}' \left(\prod_{\substack{i,j=1 \\ i<j}}^{n} \frac{\Delta_{ij}^{d_{ij}}}{d_{ij}!} \right) \tag{2.87}$$

where the prime denotes the condition that for each $j \in \{1, \ldots, n\}$ there is a $i < j$ or $k > j$ such that $d_{ij} \neq 0$ or $d_{jk} \neq 0$ (consequently $m_j \neq 0$). For example, with $n = 2$ we obtain $\langle \eta_1\eta_2\rangle = \exp(\Delta_{12}) - 1$. Evaluation for $r_1 = r_2 = k$ at equal times gives the variance $\langle \bar{\eta}_k^2\rangle = \exp(\epsilon_k^2) - 1$ for $\bar{\eta}_k$ in Eq. (2.79) where the variance of μ_k is $\epsilon_k^2 = \langle \mu_k\, \mu_k\rangle$.

2.C Small noise expansion for the mean

The calculation of the mean concentrations $\langle \mathbf{x}^s \rangle$ in Eq. (2.33) is mediated by the multi-index notation defined by Eq. (2.34) so the correlation functions in the small noise expansion read $\langle \eta_{r_1} \dots \eta_{r_n} \rangle$ and can be evaluated by means of Eq. (2.32):

$$\langle \mathbf{x}^s \rangle = \sum_{n=0}^{\infty} \sum_{\#\boldsymbol{r}=n} \boldsymbol{\phi}^s_{(\boldsymbol{r})} \sum_{u=\lfloor \frac{n+1}{2} \rfloor}^{\infty} {\sum_{|\boldsymbol{d}^n|=u}}' \prod_{\substack{i,j=1 \\ i<j}}^{n} \frac{1}{d_{ij}!} \Big(\Gamma_{r_i r_j} \Big)^{d_{ij}}. \tag{2.88}$$

Γ_{ij} denotes the covariances $\langle \mu_i(t)\, \mu_j(t) \rangle$ of the normal stochastic variables. We rearrange the order of summation with u as the principal summation index, then n runs from 0 to $2u$, and obtain the final result in Eq. (2.36).

2.D Small noise expansion for the covariance matrix

The small noise expansion of $\mathbf{V}\,(\boldsymbol{\xi})$ in analogy to $\langle \mathbf{x}^s \rangle$ is detailed in the main text, see Eq. (2.37). For $\mathbf{V}(\boldsymbol{\phi}^s)$ we use the multi-index $\boldsymbol{q} = (q_1, q_2)$, $|\boldsymbol{q}| = q_1 + q_2$, and the Taylor series of $\boldsymbol{\phi}^s$ in Eq. (2.33) to expand Eq. (2.13):

$$\mathbf{V}(\boldsymbol{\phi}^s) = \sum_{n=0}^{\infty} \sum_{|\boldsymbol{q}|=n} \sum_{\#\boldsymbol{r}^1=q_1} \sum_{\#\boldsymbol{r}^2=q_2} \boldsymbol{\phi}^s_{(\boldsymbol{r}^1)} \boldsymbol{\phi}^{s\,T}_{(\boldsymbol{r}^2)} \left(\left\langle \eta_{(\boldsymbol{r}^1)} \eta_{(\boldsymbol{r}^2)} \right\rangle - \left\langle \eta_{(\boldsymbol{r}^1)} \right\rangle \left\langle \eta_{(\boldsymbol{r}^2)} \right\rangle \right). \tag{2.89}$$

To evaluate the correlation functions we use Eq. (2.32) and the index functions f^2_{ij} from Eq. (2.39) to obtain

$$\left\langle \eta_{(\boldsymbol{r}^1)} \eta_{(\boldsymbol{r}^2)} \right\rangle = \sum_{u=\lfloor \frac{n+1}{2} \rfloor}^{\infty} {\sum_{|\boldsymbol{d}^n|=u}}' \prod_{\substack{i,j=1 \\ i<j}}^{n} \frac{1}{d_{ij}!} \Big(\Gamma_{f^2_i f^2_j} \Big)^{d_{ij}}, \tag{2.90}$$

$$\begin{aligned} \left\langle \eta_{(\boldsymbol{r}^1)} \right\rangle \left\langle \eta_{(\boldsymbol{r}^2)} \right\rangle = & \sum_{v=\lfloor \frac{n_1+1}{2} \rfloor}^{\infty} {\sum_{|\boldsymbol{d}^{n_1}|=v}}' \prod_{\substack{i,j=1 \\ i<j}}^{n_1} \frac{1}{d_{ij}!} \Big(\Gamma_{r^1_i r^1_j} \Big)^{d_{ij}} \quad \times \\ & \sum_{w=\lfloor \frac{n_2+1}{2} \rfloor}^{\infty} {\sum_{|\boldsymbol{d}^{n_2}|=w}}' \prod_{\substack{i,j=1 \\ i<j}}^{n_2} \frac{1}{d_{ij}!} \Big(\Gamma_{r^2_i r^2_j} \Big)^{d_{ij}}. \end{aligned} \tag{2.91}$$

Each term in Eq. (2.91) appears as well in Eq. (2.90) when $u = v + w$. The other way round, every term in (2.90) that contains only two-point correlation functions

that occur in one of the two factors in Eq. (2.91) cancels in the difference

$$\left\langle \eta_{(\boldsymbol{r}^1)}\eta_{(\boldsymbol{r}^2)}\right\rangle - \left\langle \eta_{(\boldsymbol{r}^1)}\right\rangle \left\langle \eta_{(\boldsymbol{r}^2)}\right\rangle = \sum_{u=\lfloor\frac{n+1}{2}\rfloor}^{\infty} {\sum_{|\boldsymbol{d}^n|=u}}'' \prod_{\substack{i,j=1\\ i<j}}^{n} \frac{1}{d_{ij}!}\Big(\Gamma_{f_i^2 f_j^2}\Big)^{d_{ij}}. \quad (2.92)$$

The new restriction indicated by the second prime in the sum, there is at least one $d_{ij} \neq 0$ for which $i \leq n_1$ and $j > n_1$, asserts that only those terms from Eq. (2.90) are kept that do not cancel with the corresponding term of the sum in Eq. (2.91). We substitute this result into Eq. (2.89) and change the order of summation. For $n < 2$ this restriction cannot be fulfilled so the sum in the result Eq. (2.38) in the main text starts with $u = 1$ and $n = 2$.

2.E First integral for the spectrum matrix

To compute Eq. (2.40) for the spectrum matrix $\boldsymbol{P}_e(\omega)$ we first evaluate the two-point time correlation functions $\Delta_{ij}(t_i - t_j)$ of the normal stochastic variables μ_i according to Eq. (2.41),

$$\frac{1}{2\pi}\int_0^\infty e^{-\mathrm{i}\omega\tau} \prod_{\substack{i,j=1\\ i<j}}^{n} \frac{1}{d_{ij}!}\Big(\Delta_{f_i^2 f_j^2}(t_i - t_j)\Big)^{d_{ij}}\,\mathrm{d}\tau \quad (2.93)$$

$$= \frac{1}{2\pi}\int_0^\infty e^{-\mathrm{i}\omega\tau} \prod_{\substack{i,j=1\\ i<j}}^{n} \frac{1}{d_{ij}!}\left(\Gamma_{f_i^2 f_j^2} e^{-K_{f_i^2}|t_i - t_j|}\right)^{d_{ij}}\,\mathrm{d}\tau \quad (2.94)$$

$$= \frac{1}{2\pi}\int_0^\infty e^{-\left(\mathrm{i}\omega + \sum_{i<j=1}^{n} d_{ij} K_{f_i^2}|t_i - t_j|\tau^{-1}\right)\tau}\,\mathrm{d}\tau \prod_{\substack{i,j=1\\ i<j}}^{n} \frac{1}{d_{ij}!}\left(\Gamma_{f_i^2 f_j^2}\right)^{d_{ij}}. \quad (2.95)$$

So far we have not treated the times t_i, t_j precisely and need to follow them back to equation Eq. (2.27) where $t_1 = \tau$ corresponds to $\boldsymbol{\eta}^1$ and $t_2 = 0$ to $\boldsymbol{\eta}^2$. The information needed for their correct evaluation is traceable in the definition of the index function f_i^2 in Eq. (2.39) that maps to components of $\boldsymbol{\eta}^1$ if $i \leq \#\boldsymbol{r}^1$ $(t_i = \tau)$ and to $\boldsymbol{\eta}^2$ else $(t_i = 0)$. The difference $|t_i - t_j|\tau^{-1}$ becomes either 0 or 1 and the sum in the exponent reduces to pairs i, j that obey $1 \leq i \leq \#\boldsymbol{r}^1 < j \leq n$. With finite and positive inverse correlation times K_i, the exponent has a negative real part so the integral with the solution

$$\frac{1}{2\pi}\left(\mathrm{i}\omega + \sum_{i=1}^{\#\boldsymbol{r}^1}\sum_{j=\#\boldsymbol{r}^1+1}^{n} d_{ij}\, K_{f_i^2}\right)^{-1} \quad (2.96)$$

exists. Finally, we evaluate the index function $f_i^2 = r_i^1$ for $i \leq \#\boldsymbol{r}^1$, denote the double sum as $\Theta(\boldsymbol{d}^n, \boldsymbol{r}^1)$ in Eq. (2.43) and add the complex conjugate of the whole expression to obtain $\boldsymbol{P}_e(\omega)$ in Eq. (2.42).

2.F Second integral for the spectrum matrix

To calculate the spectrum matrix $\langle \boldsymbol{P}_i(\omega) \rangle$ in Eq. (2.29), we expand Eq. (2.28) with the Taylor coefficients for $\boldsymbol{A}(\boldsymbol{\eta}_1)$, using Eq. (2.34), and $\boldsymbol{C}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$, using Eq. (2.44) and (2.45),

$$\boldsymbol{R}(\omega) = \int_0^\infty e^{-(-\boldsymbol{A}^0 + i\omega)\tau} \left\langle \sum_{c=0}^{\infty} \frac{\tau^c}{c!} \left(\sum_{q=1}^{\infty} \sum_{\#\boldsymbol{r}=q} \boldsymbol{A}_{(\boldsymbol{r})} \eta_{(\boldsymbol{r})} \right)^c \left(\sum_{a=0}^{\infty} \sum_{\#\boldsymbol{r}=a} \sum_{\#\boldsymbol{\sigma}=a} \boldsymbol{C}_{(\boldsymbol{r},\boldsymbol{\sigma})} \eta_{(\boldsymbol{r},\boldsymbol{\sigma})} \right) \right\rangle \mathrm{d}\tau. \quad (2.97)$$

The term in the average is a sum of terms of the form

$$\frac{\tau^c}{c!} \left(\sum_{\#\boldsymbol{r}^1=q_1} \boldsymbol{A}_{(\boldsymbol{r}^1)} \eta_{(\boldsymbol{r}^1)} \right) \ldots \left(\sum_{\#\boldsymbol{r}^c=q_c} \boldsymbol{A}_{(\boldsymbol{r}^c)} \eta_{(\boldsymbol{r}^c)} \right) \left(\sum_{\#\boldsymbol{r}^{c+1}=a} \sum_{\#\boldsymbol{\sigma}=a} \boldsymbol{C}_{(\boldsymbol{r}^{c+1},\boldsymbol{\sigma})} \eta_{(\boldsymbol{r}^{c+1},\boldsymbol{\sigma})} \right). \quad (2.98)$$

With the multi-index $\boldsymbol{q} = (q_1, \ldots, q_c)$ from Eq. (2.47) we change to $n = |\boldsymbol{q}| + a$ (the order in $\boldsymbol{\eta}$) as principal sum index,

$$\boldsymbol{R}(\omega) = \sum_{n=0}^{\infty} \sum_{a=0}^{n} \sum_{|\boldsymbol{q}|=n-a} \sum_{\#\boldsymbol{r}^1=q_1} \ldots \sum_{\#\boldsymbol{r}^c=q_c} \sum_{\#\boldsymbol{\sigma}=a} \times \int_0^\infty e^{-(-\boldsymbol{A}^0 + i\omega)\tau} \frac{\tau^c}{c!} \Big(\boldsymbol{A}_{(\boldsymbol{r}^1)} \ldots \boldsymbol{A}_{(\boldsymbol{r}^c)} \boldsymbol{C}_{(\boldsymbol{r}^{c+1},\boldsymbol{\sigma})} \left\langle \eta_{(\boldsymbol{r}^1)} \ldots \eta_{(\boldsymbol{r}^c)} \eta_{(\boldsymbol{r}^{c+1},\boldsymbol{\sigma})} \right\rangle \Big) \mathrm{d}\tau.$$

The sum $\sum_{|\boldsymbol{q}|=n-a}$ is carried out over all possible c, $(q_1, \ldots, q_c)$ with $q_1 + \cdots + q_c + a = n$. The correlation functions are calculated according to Eq. (2.32). After

changing the order of summation, a comparison to $\boldsymbol{R}(\omega)$ in Eq. (2.48) gives

$$\left(-\boldsymbol{A}^0 + \theta(\boldsymbol{d}^c, |\boldsymbol{q}|, \boldsymbol{r}^{c+1}, \boldsymbol{\sigma}) + \mathrm{i}\omega\right)^{-(c+1)} \tag{2.99}$$

$$= \frac{1}{c!}\int_0^\infty e^{-\left(-\boldsymbol{A}^0 + \mathrm{i}\omega + \sum_{i<j=1}^{n} d_{ij} K_{f_i^{c+1}} |t_i - t_j| \tau^{-1}\right)\tau} \tau^c \,\mathrm{d}\tau \tag{2.100}$$

$$= \left(-\boldsymbol{A}^0 + \mathrm{i}\omega + \sum_{\substack{i,j=1 \\ i<j}}^{n} d_{ij} K_{f_i^{c+1}} |t_i - t_j| \tau^{-1}\right)^{-(c+1)} \tag{2.101}$$

where the sum in the exponent stems from $\Delta_{ij}(t_i - t_j)$ in Eq. (2.41) and $\int_0^\infty e^{-a\tau}\,\tau^c\,\mathrm{d}\tau = \left(-\frac{\mathrm{d}}{\mathrm{d}a}\right)^c \int_0^\infty e^{-a\tau}\,\mathrm{d}\tau = \frac{c!}{a^{c+1}}$ was used in the last equality. We identify the θ function as the sum in the last term that we evaluate following the arguments in section 2.E. The difference $|t_i - t_j|\tau^{-1}$ is 0 or 1 and non zero if and only if f_j^{c+1} in Eq. (2.39) maps to a component of r^{c+1}, that is $j > |\boldsymbol{q}|$, and the index function f_i^{c+1} either maps to a component of $\boldsymbol{r}^i$, $i \in \{1, \ldots, c\}$ (corresponding to t_1 in Eq. 2.28) nd $\sigma_{i-|\boldsymbol{q}|} = 2$ (corresponding to t_2) or also $i > |\boldsymbol{q}|$ and $\sigma_{i-|\boldsymbol{q}|} \neq \sigma_{j-|\boldsymbol{q}|}$ (so $t_1 \neq t_2$). This result is formalised in Eq. (2.49) and (2.50) in the main text. For simplification of the notation in Eq. (2.49), $K_{f_i^{c+1}}$ is evaluated to $K_{r_{j-|\boldsymbol{q}|}^{c+1}}$ which is allowed due to the $\Gamma_{f_i^{c+1} f_j^{c+1}}$ in Eq. (2.48) that is proportional to $\delta_{f_i^{c+1} f_j^{c+1}}$.

2.G Exemplary evaluation of the small noise expansion

While the closed-form expressions obtained from the small noise expansion are well suited for automated evaluation, the notation is rather complicated and will be unfamiliar to most readers. To facilitate reading of the sums to the reader we evaluate the first terms in more detail here. The simplest of the sums is Eq. (2.36) for the mean concentrations

$$\langle \mathbf{x}^s \rangle = \sum_{u=0}^{\infty} \sum_{n=0}^{2u} \sum_{\#\boldsymbol{r}=n} \boldsymbol{\phi}^s_{(\boldsymbol{r})} \sum_{|\boldsymbol{d}^n|=u}{}' \prod_{\substack{i,j=1 \\ i<j}}^{n} \frac{1}{d_{ij}!} \left(\Gamma_{r_i r_j}\right)^{d_{ij}}$$

for which we evaluate all terms for $u = 0$ and $u = 1$:

$$
\begin{array}{llll}
u=0 & n=0 & \boldsymbol{r}=() & \boldsymbol{\phi}^s_{(\boldsymbol{r})} = \boldsymbol{\phi}^s_{()} = \boldsymbol{\phi}^s(\mathbf{o}) \\
u=1 & n=0 & \boldsymbol{r}=() & 0 \\
 & n=1 & \boldsymbol{r}=(i) & 0 \\
 & n=2 & \boldsymbol{r}=(i,j) & \boldsymbol{\phi}^s_{(i,j)} \Gamma_{ij} = \frac{1}{2}\frac{\partial}{\partial\eta_i}\frac{\partial}{\partial\eta_j}\,\boldsymbol{\phi}^s(\boldsymbol{\eta})|_{\boldsymbol{\eta}=\mathbf{o}}\,\delta_{ij}\epsilon_i^2 \\
 & & & \qquad = \frac{1}{2}\frac{\partial^2}{\partial\eta_i^2}\,\boldsymbol{\phi}^s(\boldsymbol{\eta})|_{\boldsymbol{\eta}=\mathbf{o}}\,\epsilon_i^2
\end{array}
$$

where the last term is a sum over the index i that enumerates the extrinsic noise variables η_i (Einstein notation). The $\Gamma_{r_i r_j}$ symbol has been evaluated according to Eq. (2.35). We obtain the first order result in the variances ϵ_i^2 of the stochastic variables $\mu_i(t)$ from which we constructed the lognormal variables (Eq. 2.79)

$$\langle \mathbf{x}^s \rangle = \boldsymbol{\phi}^s(\mathbf{o}) + \frac{1}{2}\frac{\partial^2}{\partial\eta_i^2}\,\boldsymbol{\phi}^s(\boldsymbol{\eta})|_{\boldsymbol{\eta}=\mathbf{o}}\,\epsilon_i^2 + \mathcal{O}(\epsilon_i^4). \tag{2.102}$$

The contribution $\mathbf{V}(\boldsymbol{\xi})$ in Eq. (2.37) to the total variance is formally equivalent and we obtain

$$\mathbf{V}(\boldsymbol{\xi}) = \boldsymbol{C}(\mathbf{o}) + \frac{1}{2}\frac{\partial^2}{\partial\eta_i^2}\,\boldsymbol{C}(\boldsymbol{\eta})|_{\boldsymbol{\eta}=\mathbf{o}}\,\epsilon_i^2 + \mathcal{O}(\epsilon_i^4). \tag{2.103}$$

As opposed to Eq. (2.45), $\frac{\partial}{\partial\eta_i}\,\boldsymbol{C}(\boldsymbol{\eta})|_{\boldsymbol{\eta}=\mathbf{o}}$ is a Taylor coefficient of $\boldsymbol{C}(\boldsymbol{\eta}) \equiv \boldsymbol{C}(\boldsymbol{\eta},\boldsymbol{\eta})$ from Eq. (2.22) at equal times.

For the purely extrinsic contribution to the variance in Eq. (2.38)

$$\mathbf{V}(\boldsymbol{\phi}^s) = \sum_{u=1}^{\infty}\sum_{n=2}^{2u}\sum_{|\boldsymbol{q}|=n}\sum_{\#\boldsymbol{r}^1=q_1}\sum_{\#\boldsymbol{r}^2=q_2}\boldsymbol{\phi}^s_{(\boldsymbol{r}^1)}\boldsymbol{\phi}^{s\,T}_{(\boldsymbol{r}^2)}\sum_{|\boldsymbol{d}^n|=u}{}''\prod_{\substack{i,j=1\\ i<j}}^{n}\frac{1}{d_{ij}!}\Big(\Gamma_{f_i^2 f_j^2}\Big)^{d_{ij}}$$

the sum starts with $u=1$ and with the double prime only terms with derivatives of both $\boldsymbol{\phi}^s$ and $\boldsymbol{\phi}^{sT}$ contribute,

$$
\begin{array}{llll}
u=1 & \boldsymbol{q}=(2,0) & \boldsymbol{r}^1=(i,j) & \boldsymbol{r}^2=() \qquad 0 \\
 & \boldsymbol{q}=(1,1) & \boldsymbol{r}^1=(i) & \boldsymbol{r}^2=(j) \qquad \boldsymbol{\phi}^s_{(i)}{\boldsymbol{\phi}^s_{(j)}}^T\Gamma_{ij} = \\
 & & & \qquad \Big(\frac{\partial}{\partial\eta_i}\,\boldsymbol{\phi}^s(\boldsymbol{\eta})|_{\boldsymbol{\eta}=\mathbf{o}}\Big)\Big(\frac{\partial}{\partial\eta_j}\,\boldsymbol{\phi}^s(\boldsymbol{\eta})|_{\boldsymbol{\eta}=\mathbf{o}}\Big)^T\delta_{ij}\epsilon_i^2 \\
 & \boldsymbol{q}=(0,2) & \boldsymbol{r}^1=() & \boldsymbol{r}^2=(i,j) \quad 0\ .
\end{array}
$$

The $\Gamma_{f_i^2 f_j^2}$ symbol has been evaluated with Eq. (2.35) and (2.39). The first order

result in the variances ϵ_i^2 is

$$\mathbf{V}(\boldsymbol{\phi}^s) = \left(\frac{\partial}{\partial\eta_i}\,\boldsymbol{\phi}^s(\boldsymbol{\eta})|_{\boldsymbol{\eta}=\mathbf{o}}\right)\left(\frac{\partial}{\partial\eta_i}\,\boldsymbol{\phi}^s(\boldsymbol{\eta})|_{\boldsymbol{\eta}=\mathbf{o}}\right)^T \epsilon_i^2 + \mathcal{O}(\epsilon_i^4). \tag{2.104}$$

The total variance is $\mathbf{V}\left(\mathbf{x}^s\right) = \mathbf{V}\left(\boldsymbol{\phi}^s\right) + \frac{1}{\Omega}\mathbf{V}\left(\boldsymbol{\xi}\right)$ according to Eq. (2.12).

In comparison to $\mathbf{V}(\boldsymbol{\phi}^s)$, the purely extrinsic contribution to the spectrum matrix $\boldsymbol{P}_e(\omega)$ in Eq. (2.42) needs evaluation of the additional factor $\frac{\pi^{-1}\Theta}{\omega^2+\Theta^2}$ with $\boldsymbol{d}^2 = (d_{12}) = (1)$, $\boldsymbol{r}^1 = (i)$ and $n = 2$. Eq. (2.43) then gives $\Theta(\boldsymbol{d}^2, \boldsymbol{r}^1) = K_i$ and

$$\boldsymbol{P}_e(\omega) = \left(\frac{\partial}{\partial\eta_i}\,\boldsymbol{\phi}^s(\boldsymbol{\eta})|_{\boldsymbol{\eta}=\mathbf{o}}\right)\left(\frac{\partial}{\partial\eta_i}\,\boldsymbol{\phi}^s(\boldsymbol{\eta})|_{\boldsymbol{\eta}=\mathbf{o}}\right)^T \frac{K_i\epsilon_i^2}{\pi\left(\omega^2+K_i^2\right)} + \mathcal{O}(\epsilon_i^4). \tag{2.105}$$

For the second spectrum matrix $\boldsymbol{P}_e(\omega)$ we need to evaluate Eq. (2.48),

$$\boldsymbol{R}(\omega) = \sum_{u=0}^{\infty}\sum_{n=0}^{2u}\sum_{a=0}^{n}\sum_{|\boldsymbol{q}|=n-a}\sum_{\#\boldsymbol{r}^1=q_1}\cdots\sum_{\#\boldsymbol{r}^c=q_c}\sum_{\#\boldsymbol{r}^{c+1}=a}\sum_{\#\boldsymbol{\sigma}=a}\sum_{|\boldsymbol{d}^n|=u}{}'$$
$$\times\frac{1}{\left(-\boldsymbol{A}^0+\theta(\boldsymbol{d}^n,|\boldsymbol{q}|,\boldsymbol{r}^{c+1},\boldsymbol{\sigma})+i\omega\right)^{c+1}}\,\boldsymbol{A}_{(\boldsymbol{r}^1)}\ldots\boldsymbol{A}_{(\boldsymbol{r}^c)}\boldsymbol{C}_{(\boldsymbol{r}^{c+1},\boldsymbol{\sigma})}\prod_{\substack{i,j=1\\ i<j}}^{n}\frac{1}{d_{ij}!}\left(\Gamma_{f_i^{c+1}f_j^{c+1}}\right)^{d_{ij}}.$$

To elucidate the full complexity of the sum, we here evaluate some terms for $u = 2$. We use an abbreviated notation for derivatives as exemplarily defined by

$$\boldsymbol{A}''_{ij} \equiv \frac{1}{2!}\frac{\partial}{\partial\eta_i}\frac{\partial}{\partial\eta_j}\,\boldsymbol{A}(\boldsymbol{\eta})|_{\boldsymbol{\eta}=\mathbf{o}} \qquad \text{and} \qquad \boldsymbol{C}_i^{\sigma} \equiv \frac{\partial}{\partial\eta_i^{\sigma}}\,\boldsymbol{C}(\boldsymbol{\eta}^1,\boldsymbol{\eta}^2)\big|_{\boldsymbol{\eta}^1=\boldsymbol{\eta}^2=\mathbf{o}}\,. \tag{2.106}$$

For $u = 2$, $n = 4$, $a = 1$ and $c = 2$ we obtain the terms

$$\sum_{\substack{|\boldsymbol{q}|=3\\ \#\boldsymbol{q}=2}} \sum_{\#\boldsymbol{r}^1=q_1} \sum_{\#\boldsymbol{r}^2=q_2} \sum_{r_1^3} \sum_{\sigma_1} \sum_{|\boldsymbol{d}^4|=2}{}' \frac{1}{\left(-\boldsymbol{A}^0 + \theta(\boldsymbol{d}^4, 3, \boldsymbol{r}^3, \boldsymbol{\sigma}) + i\omega\right)^3} \boldsymbol{A}_{(\boldsymbol{r}^1)} \boldsymbol{A}_{(\boldsymbol{r}^2)} \boldsymbol{C}_{(\boldsymbol{r}^3,\boldsymbol{\sigma})} \prod_{\substack{i,j=1\\ i<j}}^{4} \frac{1}{d_{ij}!} \left(\Gamma_{f_i^3 f_j^3}\right)^{d_{ij}}$$

$$= \left(\frac{\delta_{ij}\delta_{kl}\epsilon_i^2\epsilon_k^2}{\left(-\boldsymbol{A}^0 + K_{kl}\delta_{2\sigma} + i\omega\right)^3} + \frac{\delta_{ik}\delta_{jl}\epsilon_i^2\epsilon_j^2}{\left(-\boldsymbol{A}^0 + K_{jl}\delta_{2\sigma} + i\omega\right)^3} + \frac{\delta_{il}\delta_{jk}\epsilon_i^2\epsilon_l^2}{\left(-\boldsymbol{A}^0 + K_{il}\delta_{2\sigma} + i\omega\right)^3} \right) \left(\boldsymbol{A}_i' \boldsymbol{A}_{jk}'' + \boldsymbol{A}_{ij}'' \boldsymbol{A}_k'\right) \boldsymbol{C}_l^\sigma. \quad (2.107)$$

The sum over $|\boldsymbol{q}| = 3$ was evaluated by writing $c = 2$ symbols $\boldsymbol{A}$ with all possibilities to assign at least one of a total of three indices to each of them. The Einstein notation for the extrinsic noise components $i, j, k, l \in \{1, \ldots, M\}$ of $\boldsymbol{\eta}$ and times t_σ, $\sigma \in \{1, 2\}$ accounts for all other sums except the primed sum over $|\boldsymbol{d}^4| = 2$. The latter involves 3 tuples $(d_{12}, d_{13}, d_{23}, d_{14}, d_{24}, d_{34})$, namely all components zero but $d_{12} = d_{34} = 1$, $d_{13} = d_{24} = 1$ or $d_{14} = d_{23} = 1$. The covariances $\Gamma_{f_i^3 f_j^3}$ and the θ function evaluate according to Eq. (2.35), (2.39) and (2.49).

With $n = 3$ instead, $\boldsymbol{q}$ needs to be $(1, 1)$ and we sum over $|\boldsymbol{d}^3| = 2$ and obtain the three terms

$$\left(\frac{\delta_{ij}\delta_{ik}\epsilon_i^4}{\left(-\boldsymbol{A}^0 + K_{ik}\delta_{2\sigma} + i\omega\right)^3} + \frac{\delta_{ij}\delta_{jk}\epsilon_i^4}{\left(-\boldsymbol{A}^0 + K_{jk}\delta_{2\sigma} + i\omega\right)^3} + \frac{\delta_{ik}\delta_{jk}\epsilon_i^4}{\left(-\boldsymbol{A}^0 + (K_{ik} + K_{jk})\delta_{2\sigma} + i\omega\right)^3} \right) \boldsymbol{A}_i' \boldsymbol{A}_j' \boldsymbol{C}_k^\sigma. \quad (2.108)$$

Finally, we evaluate the terms for $u = n = 2$ and $a = 0$ for which $|\boldsymbol{d}^2| = 2$ only allows $d_{12} = 2$ and with $a = 0$ no inverse correlation times K_i are involved but $\boldsymbol{q}$ may be both (2) with $c = 1$ or $(1, 1)$ with $c = 2$ which gives

$$\frac{\frac{1}{2!}\delta_{ij}^2\epsilon_i^4}{\left(-\boldsymbol{A}^0 + i\omega\right)^2} \boldsymbol{A}_{ij}'' \boldsymbol{C}_0 + \frac{\frac{1}{2!}\delta_{ij}^2\epsilon_i^4}{\left(-\boldsymbol{A}^0 + i\omega\right)^3} \boldsymbol{A}_i' \boldsymbol{A}_j' \boldsymbol{C}_0 = \frac{\frac{1}{2}\epsilon_i^4}{\left(-\boldsymbol{A}^0 + i\omega\right)^2} \left(\boldsymbol{A}_{ii}'' + \frac{1}{\left(-\boldsymbol{A}^0 + i\omega\right)} \boldsymbol{A}_i' \boldsymbol{A}_i'\right) \boldsymbol{C}_0. \quad (2.109)$$

We note here that the matrix multiplication is non commutative so the order of terms is important. In general, using Einstein notation for fixed u and n, one writes down all possible terms $\boldsymbol{A}\ldots\boldsymbol{AC}$ with n indices ($\boldsymbol{C}_0$ without index is allowed but not so for the Taylor coefficients of $\boldsymbol{A}$) and then for each term evaluates the remaining sum over $|\boldsymbol{d}^n| = u$ in order to derive the factors containing the θ function (here c is the number of $\boldsymbol{A}$-symbols) and covariances Γ_{ij}.

The spectrum matrix $\langle \boldsymbol{P}_i(\omega)\rangle = \frac{1}{2\pi}\left(\boldsymbol{R}(\omega) + \boldsymbol{R}(\omega)^{*T}\right)$ in zero'th order is the power spectrum of the Ornstein-Uhlenbeck process for the concentrations $\mathbf{x}$ in the absence of extrinsic noise. The Ornstein-Uhlenbeck is obtained from Eq. (2.9) by setting $\boldsymbol{\eta}$ to zero. Explicitly, its power spectrum is given by the well known result[73]

$$\boldsymbol{R}(\omega) = \left(-\boldsymbol{A}^0 + i\omega\right)^{-1}\boldsymbol{C}_0 + \mathcal{O}\left(\epsilon_i^2\right) \tag{2.110}$$

$$\Rightarrow \langle \boldsymbol{P}_i(\omega)\rangle = \frac{1}{2\pi}\left(-\boldsymbol{A}^0 + i\omega\right)^{-1}\boldsymbol{B}\boldsymbol{B}^T\left(-\boldsymbol{A}^{0\,T} - i\omega\right)^{-1} + \mathcal{O}\left(\epsilon_i^2\right) \tag{2.111}$$

where we have used the Lyapunov equation (2.22) for the last equality. According to Eq. (2.23), the total spectrum matrix is $\boldsymbol{P}(\omega) = \boldsymbol{P}_e(\omega) + \frac{1}{\Omega}\langle \boldsymbol{P}_i(\omega)\rangle$.

Chapter 3

Single-cell variability of CRISPR-Cas interference and adaptation

A revised version of this chapter has been accepted for publication at *Molecular Systems Biology*:

Rebecca E. McKenzie[†], Emma M. Keizer[†], Jochem N.A. Vink, Jasper van Lopik, Ferhat Büke, Vera Kalkman, Christian Fleck, Sander J. Tans, Stan J.J. Brouns

[†]These authors contributed equally

Abstract

CRISPR-Cas defence is a combination of adaptation to new invaders by spacer acquisition, and interference by targeted nuclease activity. While these processes have been studied on a population level, the individual cellular variability has remained unknown. Here, using a microfluidic device combined with time-lapse microscopy, we monitor invader clearance in a population of *Escherichia coli* across multiple generations. We observed that CRISPR interference is fast with a narrow distribution of clearance times. In contrast, for invaders with escaping PAM mutations we show large cell-to-cell variability of clearance times, which originates from primed CRISPR adaptation. Further, by comparing cell lineage features, we determined that faster growth and cell division, and higher levels of Cascade increase the chance of clearance by interference, while slower growth is associated with increased rates of priming. Finally, through mathematical modelling we estimated the influence of target and Cascade copy numbers, as well as binding affinity of Cascade on the rate of the priming response. Our results show that the ability to adapt to an invading threat by primed CRISPR adaptation is highly stochastic, implying that only subpopulations of bacteria are able to respond to impending threats in a timely manner.

3.1 Introduction

During the last decade, important progress has been made in identifying the sequence of steps and molecular interactions required for successful adaptive immunity by the model type I-E CRISPR-Cas system.[108–117] CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) immunity involves three main stages beginning with the acquisition of a spacer, a small piece of DNA derived from a foreign invader and stored in the CRISPR array for future defence.[118,119] This array is then transcribed and processed into small CRISPR RNAs (crRNAs) which guide a surveillance complex, formed from a number of Cas (CRISPR associated) proteins, towards the invaders DNA.[120,121] For type I-E systems a 5'-CTT consensus PAM (Protospacer Adjacent Motif) sequence flanking the targeted site of the invader[122,123] allows swift recognition and ultimately degradation of the invader, through a process called direct interference.[114,124–126] However, invaders can escape direct interference via mutation within the seed region of the target site or PAM.[122,127,128] In response, the I-E system can initiate priming, which promotes accelerated acquisition of new spacers due to a pre-existing partial match to the invader.[108,111] Primed adaptation is much faster than naïve adaptation,[129] and is required for the insertion of a new matching spacer with a consensus PAM allowing subsequent invader degradation, which we here refer to as primed interference.

At the level of individual cells however, much more is unknown. Interference is a tug-of-war between invader replication and degradation, which could result in complex and stochastic dynamics within single cells. Replication and degradation themselves may also display variability between cells in the population. For instance, invader degradation rates can be affected by stochastic processes such as the expression of CRISPR components, target localization, and nuclease recruitment.[110,130] Priming also depends on many processes in which the dynamical interplay is unclear, including the production of suitable fragments of DNA for spacer acquisition (pre-spacers), the assembly of adaptation complexes required for further spacer selection, and the processing and insertion of these pre-spacers into the CRISPR array.[113,117,131,132] Elucidating the cellular dynamics and heterogeneity of the CRISPR response is critical to understanding interference and adaptation mechanistically, and of direct importance to its natural function. For instance, upon invasion, cells are thought to have a limited time window to respond in order to escape invader replication, protein production, and cell death.[133–136]

A number of studies have investigated the interference process by collecting either population averages, or single-cell data on short time scales ($<$1 s).[110,112,128,137–139] However, averaging within a population can conceal the variation between cells, as well as the dynamics within single-cells over time,[7,140] thus masking the underlying dynamics of CRISPR-Cas interference. In addition, investigations into the adaptation process have provided great insight into the diversity of spacers acquired,[137,141] possible mechanisms of target destruction,[108,142] and conditions

under which adaptation most frequently occurs within a population,[143–145] however these studies could not observe any variation existing in each step of the adaptation process within individual cells.

Recently, developments in the field have begun to include the use of time-lapse microscopy to investigate invader establishment and degradation in single cells.[146,147] Here we set out to further these techniques and investigate and quantify the dynamics and variability of both the interference and adaptation processes in single-cell lineages. Using time-lapse microscopy and microfluidic devices, we followed individual cells over multiple rounds of division while simultaneously monitoring CRISPR-Cas protein expression and DNA degradation. Hence, we obtained individual lineages, the genealogical relations between them, as well as real-time data on the DNA clearance process, instantaneous growth rates, cell sizes, and division frequencies of individual cells. We determined that while direct interference occurs quickly and consistently, clearing the target from all cells within hours, priming is highly variable and much slower, taking up to several tens of hours. Further, through stochastic modelling we were able to define the adaptation and clearance stages of priming and identified primed adaptation as the source of the variation observed. Finally, we corroborated our findings with a minimal agent-based model, that accurately replicated our data and provided further insights into the dynamics of the primed adaptation process.

3.2 Results

3.2.1 Time-lapse microscopy of the CRISPR response

Using two strains, KD615 (WT) and KD635 ($\Delta cas1,2$) (Supplementary Table 3.3), we investigated priming and direct interference respectively. The strains contain an array with a leader, two repeats and a single previously characterised spacer, spacer8 (SP8)[111,112] (Fig. 3.1a-c). In addition, these strains are engineered to control *cas* gene expression using arabinose and IPTG induction, and hence initiation of the CRISPR response. Target plasmids were engineered to encode a constitutively expressed YFP or CFP fluorescent protein[148] and contain a target sequence that is complementarity to SP8 in the CRISPR array, allowing direct monitoring of target DNA presence in individual cells over time (Fig. 3.1a-c) (Supplementary Table 3.3). In order to investigate the direct interference process, we flanked the target sequence with a 5'-CTT consensus PAM[123] (Fig. 3.1a,b). Further, to investigate the priming response we mutated the PAM to 5'-CGT (Fig. 3.1b,c), a mutation known to allow mobile genetic elements (MGE) to escape interference, and invoke a primed adaptation response.[108,112,127]

Use of a microfluidic device (Wehrens2018) enabled fluorescence time-lapse imaging for over 36 hours with the option for media exchange (Fig. 3.1d). The device contained chambers allowing observation of a single layer of cells, constant

medium supply, removal of cells that no longer fit the chamber due to growth, and control of intracellular processes via induction. Image analysis software was used to segment and track all cells and their fluorescence signals, thus allowing the re-construction of lineage trees in a defined region at the bottom of the chamber (Fig. 3.1d,e).[149–151]

3.2.2 Direct interference is fast and synchronous

We first investigated the direct interference response (Fig. 3.1a). Prior to *cas* gene induction, the images showed high YFP fluorescence in all cells, confirming the presence of the target plasmid (Fig. 3.2a) which decreased upon induction, indicating CRISPR mediated degradation of the target DNA (Fig. 3.2a, Supplementary video 1). When the plasmid did not contain a target sequence (pControl) YFP levels did not decrease for over 35 hours (Supplementary Fig. 3.1), indicating targeting by CRISPR-Cas is required for plasmid loss in this set-up. The mean YFP fluorescence per cell unit area (which estimates the YFP concentration) showed the decrease started after about 1 hour of induction, and then exhibited a smooth monotonic decline without substantial fluctuations (Fig. 3.2b). Note that traces end upon the cells exiting the observation chamber. CRISPR mediated degradation of the target was thus efficient and synchronous, and in the case of a 5-copy plasmid could overcome the plasmid replication and copy number control. Hence, we surmised that the YFP fluorescence may decrease exponentially, as the YFP proteins are diluted exponentially due to volume growth upon clearance of the last plasmid. Indeed, we found the fluorescence decrease to be exponential (Supplementary Fig. 3.2).

Direct interference variability between cells also appeared limited (Fig. 3.2b). To address it more directly, we quantified the moment all plasmids are cleared by determining the YFP production rate as the change in total cellular fluorescence per unit of time.[152] The production rate scales with the number of target DNA copies, and shows the expression timing more precisely by suppressing slow dilution effects. Indeed, the YFP production rate decreased rapidly, and reached zero (the background level of cells not expressing YFP) when the mean fluorescence was still close to its maximum (Fig. 3.2c). This moment was identified as the plasmid loss time (PLT) (Fig. 3.2c). PLT was narrowly distributed between about 1 and 2.5 hours (Fig. 3.2d, $CV^2 = 0.055$). Hence, in all cells the target was cleared. The clearance was rapid taking between 1 and 3 generations, and sometimes occurred in the same generation in which the CRISPR response was initiated by induction (Supplementary Fig. 3.3).

3.2.3 Primed adaptation is highly variable

Next, we studied plasmid clearance after adaptation from a target with a mutated PAM (Fig. 3.1c). Most notable in these priming experiments was the heterogeneity

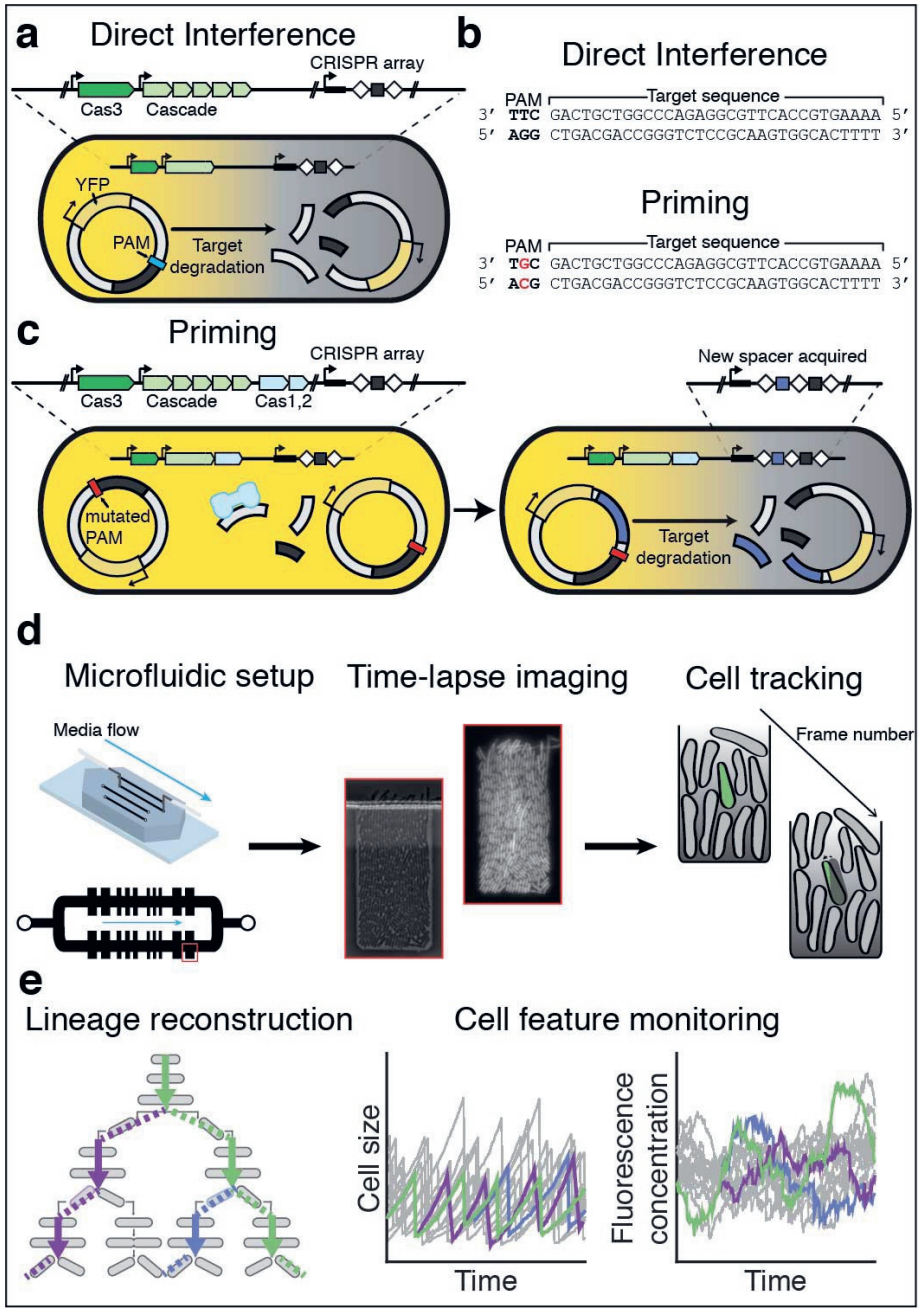


Figure 3.1: Figure caption on next page.

Figure 3.1: Investigating single-cell behaviour during CRISPR defence using time-lapse microscopy a, Schematic of the direct interference process. The cell contains a I-E CRISPR system, as well as the CRISPR array with a single spacer targeting the plasmid (grey box). The plasmid encodes YFP and contains a sequence matching the spacer (grey), flanked by a consensus PAM (blue). Immediate targeting by the CRISPR system resulting in degradation of the plasmid and loss of the YFP in the cell. b, To invoke priming the 5'-CTT consensus PAM, flanking the target sequence located on the plasmid, is mutated by one nucleotide to a non-consensus PAM 5'-CGT. c, Schematic of the priming process. (Left cell) A mutation of the PAM (red) flanking the target sequence means the spacer in the CRISPR array can no longer initiate direct interference. Fragments in the cell can be captured and processed by Cas1,2 (light blue). (Right cell) The Cas1,2 complex integrates the fragment into the CRISPR array as a new spacer (purple), which matches the target plasmid resulting in degradation and loss of YFP in the cell. d, To allow long term imaging cells are grown in a microfluidic chip that allows constant media supply. Cells within a single well are imaged frequently in phase contrast and fluorescence allowing segmentation and tracking of lineage history across frames. e, Variation in features of reconstructed single-cell lineages (left) such as size (middle) and fluorescence concentration (right) are continuously monitored enabling further investigation.

between lineages, with the clearance process ranging from 2-30 cellular generations (Supplementary Fig. 3.3). Upon induction, some lineages showed a decreasing trend in fluorescence as early as 4 hours (Fig. 3.2e-f, Supplementary video 2), while others remained fluorescent after 35 hours (Fig. 3.2f). The PLTs were indeed broadly distributed and displayed a long tail towards large values (Fig. 3.2g, $CV^2 = 0.458$). Of note, we did not observe plasmid clearance in the same generation in which the CRISPR system was induced (Supplementary Fig. 3.3).

The shapes of the YFP declines were exponential, similar to the direct interference data (Fig. 3.2b and f, Supplementary Fig. 3.2). When aligned at the PLT, the average profile of all production rate traces for direct interference and priming show a similar trend both right before and after plasmid loss is detected (Supplementary Fig. 3.4). In these data, the onset of the decrease is about 60 min before PLT in both cases, thus estimating the clearance time (CT), the duration of the target clearance process. In priming, clearance thus contributes much less to PLT variability than the preceding processes (Fig. 3.2g). These observations suggest that new spacers preceded by a consensus PAM are indeed acquired, and that the CRISPR adaptation phase is responsible for the observed temporal variability (Fig. 3.2g).

Spacer acquisition in the population was indeed confirmed by PCR of the

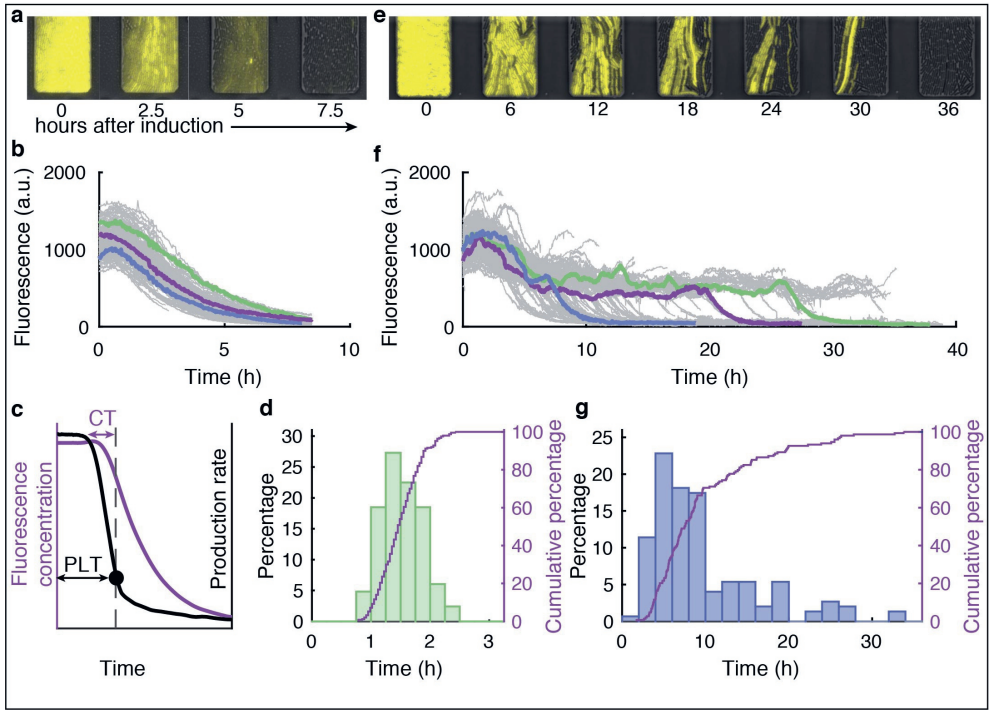


Figure 3.2: Variation in target plasmid clearance times is much larger when CRISPR adaptation is required a-b, Depict clearance of a target with a consensus PAM by direct interference a, Overlay of fluorescent and phase contrast time-lapse images. Presence of the target plasmid is tracked by its YFP production. Images are shown at 2.5 h intervals starting from induction of cas gene expression. b, Reconstructed lineage traces of the imaged population (a) from induction of the CRISPR system over time (grey) lineages show some variation in plasmid clearance times (coloured). c, Production rate (black line) of the YFP is used to determine the plasmid loss time, PLT, (black dot, black arrow) allowing earlier detection than using the mean fluorescence (purple line). The time from first targeting of a single plasmid to the PLT is defined as the clearance time (CT, purple arrow). d, Distribution of PLTs determined by the production rate during direct interference (n=250). e-f, Depict clearance of a matching target with a mutated PAM via priming e, Overlay of fluorescent and phase contrast time-lapse images. Presence of the target is tracked by YFP production. Images are shown at 6 h intervals. f, Reconstructed lineage traces of the imaged population (e) from induction of the CRISPR system over time (grey). Lineages show large variations in the time taken to clear the plasmid (coloured). g, Distribution of plasmid loss times calculated with the production rate during priming (n=149).

CRISPR array in cells collected from the microfluidic device (Supplementary Fig. 3.5). Spacer acquisition was not observed with the $\Delta cas1,2$ strain, consistent with Cas1 and Cas2 being required for acquisition.[153] In the absence of Cas1 and Cas2 however, low frequency plasmid loss was observed in 1.4% of the lineages over a 35-hour period (Supplementary Fig. 3.6). Hence, complete clearance is possible with a mutated PAM, even if highly inefficient.

3.2.4 Genealogical relations impact the CRISPR response

To study the role of genealogy in the CRISPR response, we took a more in depth look at the lineage history before plasmid loss (Fig. 3.3a). For primed adaptation, some subtrees showed all plasmid loss events occurring close together, however most subtrees showed a wide PLT variability (Fig. 3.3b, black dots), in line with lineages responding independently. However, statistical analysis showed that sisters cleared their plasmids within the same cell cycle more frequently than expected at random, and more strongly so for priming than for direct interference (Fig. 3.3c). Hence, inheritance plays a role in the CRISPR response (Fig. 3.3c).

These data led us to hypothesise that in priming, plasmid loss times in sisters correlate because spacer acquisition occurs in the mother, after which plasmid degradation (primed interference) continues into the daughters. If true, the detection of plasmid loss in each daughter will likely be close in timing, with the moment in the cell cycle for both daughters determined by when spacer acquisition occurred within the mother's cell cycle. This would result in a random distribution of loss times throughout the cell cycle for each pair of daughters in the experiment. Conversely, when loss in sisters was not correlated (*i.e.* only one sister cleared the plasmid) we believe both acquisition and clearance managed to occur in the same cell cycle. In this case, we would expect clearance to occur at the end of the mother's cell cycle. We base this on our earlier finding that on average 60 min (CT) is required for the interference process (Supplementary Fig. 3.4), indicating adaptation must occur at the beginning of the cell cycle and be directly followed by swift interference. To test this hypothesis, we divided the cell cycles into five equal fractions, and tabulated the observed loss event for each fraction. Indeed, loss events in just one sister occurred from frequently towards the end of the cell cycle (Fig 3.3d), while the moment of loss was more randomly distributed when both sisters lost the plasmid (Fig. 3.3d). Altogether this indicated that loss likely takes place more frequently in sisters than cousins (Fig. 3.3c) because adaptation occurred in the mother.

3.2.5 The growth rate has opposing effects on adaptation and interference

To study if stochastic variations in cell cycle parameters affect the CRISPR-Cas response, we developed a ranking analysis to rank each 'loss-lineage' that successfully

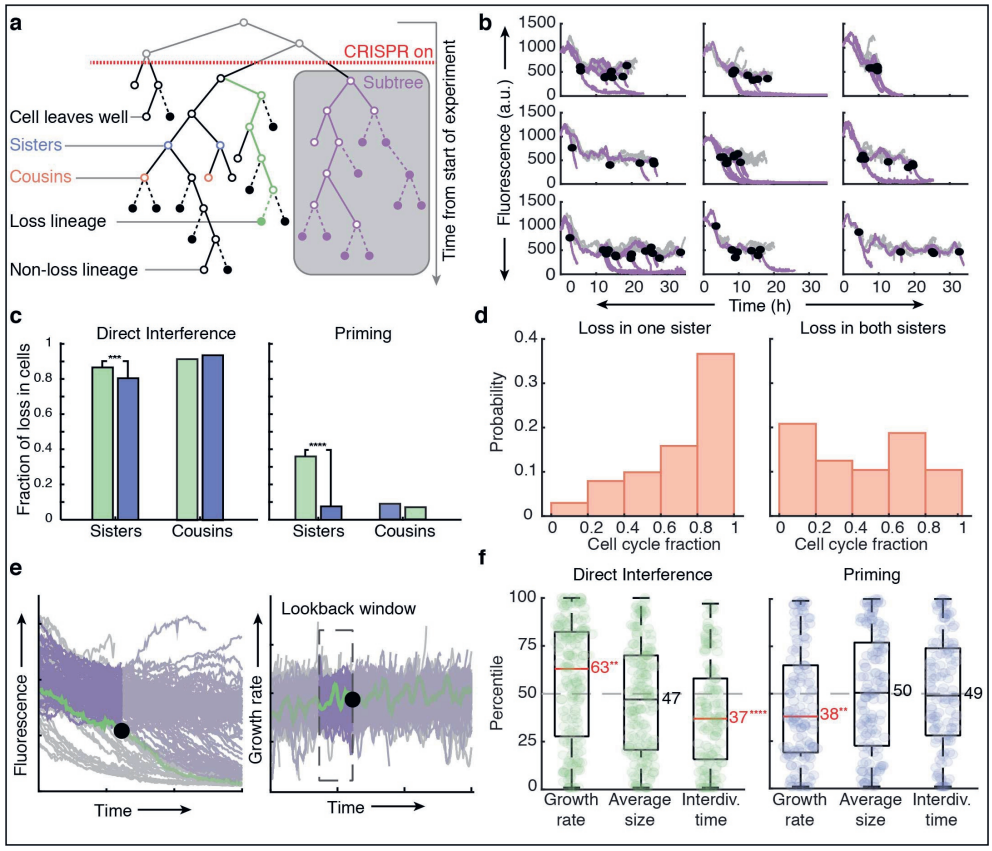


Figure 3.3: Figure caption on next page.

Figure 3.3: Growth rate and interdivision times have an influence on direct interference and priming a, Schematic of key analysis structure and terminology. b, A comparison of 9 subtrees constructed from induction. c, The observed fraction of loss in cells (green) during direct interference (left) or priming (right) related as either sisters (DI: n=171, P: n=98) or cousins (DI: n=130, P: n=138) is plotted against the fraction of expected loss events (blue) in related cells when the events are randomized in the same time window. d, The cell cycle was divided into 5 equal fractions and plasmid loss times are plotted in the corresponding fraction where one sister alone cleared the plasmid (left, n=101) or both sisters cleared the plasmid (right, n=24) e, Schematic explaining the rank-based analysis approach. For each detected loss event (left, black circle) the cell feature i.e. growth rate for that lineage (right, green) is averaged over a lookback window (right, dashed rectangle), and then ranked amongst all averages of non-loss lineages in the same window (violet, right). f, Boxplots of percentile rankings of all loss lineages that cleared a consensus target via direct interference (green, left, n=250), or a mutated target via priming (blue, right, n=149), for growth rate, birth size and interdivision respectively over a lookback window of 30 minutes. The median percentile ranking of loss lineages is indicated by a line and value, categories in which this value was significantly different from a ranking in the 50th percentile as computed by a 2-sided binomial test are indicated in red followed by asterisks. (****$p<0.0001$, ***$p<0.001$, **$p<0.01$)

cleared the plasmids relative to the 'non-loss lineages' that had not cleared the plasmids at that same moment in time (Fig. 3.3e). As cellular features such as the growth rate or concentration of proteins might not be in steady state due to changes in the environment, comparing loss-lineages that cleared the plasmids at different times over the course of the 36 hour experiment could result in detection of a trend in growth not related to the CRISPR-Cas response. The ranking was based on growth rate averaged over a 30 min 'lookback window' (Fig. 3.3e), determined using autocorrelation times which are a measure of the rate of change of a time series. The autocorrelation coefficient of the growth rate is no longer significant beyond 30 mins, thus indicating measurements more than 30 mins apart are unlikely to be correlated (Supplementary Fig. 3.7). In direct interference, the 'loss lineages' exhibited a higher median growth rate than 'non-loss lineages', with their growth rate ranking in the in the 63rd percentile ($p=0.01$) (Fig. 3.3f). These lineages also showed shorter interdivision times ($p=0.0001$), but not a difference in cell size (Fig. 3.3f). These results were robust over a range of lookback window sizes (see Supplementary Fig. 3.8). We stress that growth is likely only one of the many factors affecting the CRISPR response, which is also reflected by the broad ranking distributions (Fig. 3.3f). Overall, the analysis indicated that faster growth

in coordination with more frequent cell division has a positive effect on the rate of clearance of a consensus target.

Primed adaptation showed a different picture. To probe the effects on spacer acquisition, which occur about 60 min before plasmid loss, we used a lookback window between 90 and 60 min before the PLT. While cell size and interdivision time did not show an effect (no significant deviation from the 50th percentile) the growth rate did, with loss lineages growing more slowly compared to non-loss lineages (38th percentile, p=0.01) (Fig. 3.3f). This was robust to changes in the lookback window (Supplementary Fig. 3.9). Altogether, these findings indicated that, on average, slower growing cells achieved faster plasmid clearance through priming.

3.2.6 Cascade concentrations impact the CRISPR response

Apart from physiological determinants like growth, Cascade expression levels may influence the speed of CRISPR defence through growth rate fluctuations or the random partitioning of molecules at division.[19] To investigate this, we fused mCherry (RFP) to the N-terminus of the Cas8e subunit of Cascade[110] (Fig. 3.4a). Using single particle fluorescence calibration, we estimated that the cells contain on average about 200 Cascade molecules/μm^2 (Fig. 3.4b and Supplementary Fig. 3.10). Hence, we quantified the (stochastic) variations in Cascade abundance within single-cell-lineages upon induction (Fig. 3.4b).

Cascade levels fluctuated on a longer timescale than the cell cycle (200 min, Supplementary Fig. 3.11) and were strongly correlated between sisters and cousins (R=0.89 and 0.62 respectively, Supplementary Fig. 3.12) indicating that Cascade levels are stable over several generations. We reasoned that lineages with high Cascade concentrations may target and clear the plasmids faster. Hence, to test this notion, we investigated the correlation between Cascade search hours and the PLT for each lineage at time points onward from induction. Cascade search hours can be described as the sum of hours spent by all Cascades in the cell searching for the target and are determined from the cumulative RFP, *i.e.* the area under the RFP concentration curve of each lineage from induction until a point of interest (Supplementary Fig. 3.13). One may expect that a single-cell lineage which has a high number of Cascades for a short period of time close to induction could undergo adaptation earlier than a cell which has a lower number of Cascades over a longer period of time, or vice versa (Supplementary Fig. 3.13). To this end, we carried out this analysis to determine if spacer acquisition may be governed by a requirement for a number of Cascade search hours rather than a peak in copy number in the cell.

At 0-2 hours post induction, PLT and Cascade search hours indeed correlated negatively for direct interference but not for priming, indicating cells with a higher sum of Cascade search hours lost the plasmid earlier (Fig. 3.4d). This result is in

line with our earlier findings (Fig. 3.4c) and supports that stochastic variations in Cascade expression levels affect direct interference. For the priming process, the impact of Cascade levels appeared weaker and no significant correlation was found between PLT and Cascade search hours prior to loss (Fig. 3.4d). This suggests that neither the total search hours of Cascade nor the instantaneous expression levels play a detectable role in the determination of when plasmid loss occurs during priming. We hypothesise this could be due to the underlying processes being less synchronised in time in comparison to direct interference, and hence masked by other stochastic variations in our set up.

3.2.7 Low Cascade-target binding affinity generates CRISPR response variability

To gain insight into the variability and dynamics of the CRISPR-Cas defence we developed an agent-based simulation framework. Adaptive immunity in bacterial populations has been modelled previously[154–156] but to our knowledge none describe variability or single-cell behaviour. Briefly, we simulated 100 cells, their growth and division, plasmid maintenance, stochastic protein production and partitioning at division, spacer acquisition, and target DNA degradation (see Supplementary Methods for details). We found that with these minimal model ingredients, and by only changing the Cascade-target binding affinity due to the PAM mutation, the model could reproduce both the low variability of direct interference (Fig. 3.5a,b and Fig. 3.2b,d), and the high variability of priming (Fig. 3.5c,d and Fig. 3.2f,g) from the experimental conditions. Specifically, we found a Cascade-target binding affinity reduction of two orders of magnitude for the PAM mutation, which is consistent with previous work[157,158] (Supplementary Table 3.2).

The priming process can be conceptually understood as a two-step process, adaptation followed by interference, where a highly reduced rate of the first step is able to recreate the broadness of the PLT distribution (see Supplementary Methods for details). We hypothesised that variation of the primed adaptation response could originate from the low-affinity target search of Cascade, or the spacer integration. In the agent-based model, we were able to vary the rates of these two processes by a factor of 100, while keeping the Cas3-mediated target destruction constant, we find that slow spacer integration alone is not enough to explain the observed variability (Fig. 3.5h). Conversely, reduced Cascade-target binding affinity is both necessary and sufficient to reproduce the observations (Fig. 3.5e-h) and is required to generate pre-spacers.

3.2.8 Competition between adaptation and low-level interference

In priming, low Cascade-target affinity and resulting sporadic target degradation can yield a low-level interference prior to adaptation, which in turn provides a

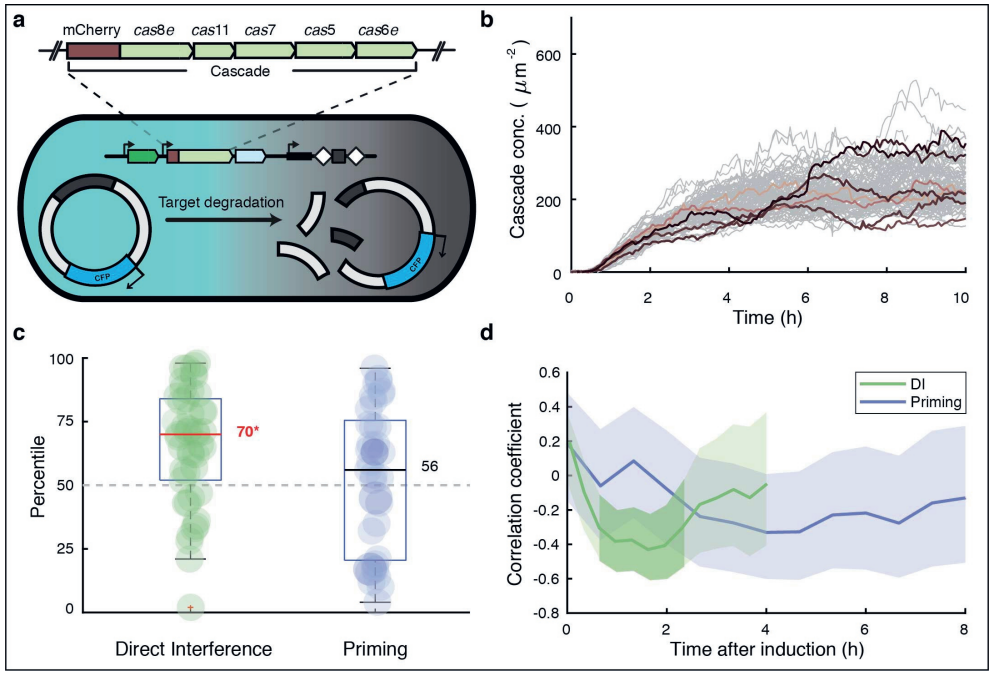


Figure 3.4: Growth rate and interdivision times influence direct interference and priming a, Schematic of the experimental set up adapted to allow visualization of target presence (CFP) and Cascade levels (mCherry) simultaneously. The expansion indicates the mCherry fluorescent tag was integrated upstream of the cas8e subunit. b, Cascade concentration of single-cell lineages over time from induction. c, Cascade concentrations were averaged over a 30-minute lookback window from the plasmid loss event for all loss lineages during direct interference (green) or priming (blue). The Cascade concentration of the loss lineages were ranked as percentile amongst the non-loss lineages and plotted here. The median percentile ranking of loss lineages is indicated by a line and value, categories in which this value was significantly different from a ranking in the 50th percentile as computed by a 2-sided binomial test (*$p<0.05$) are indicated in red followed by an asterisk. d, The Pearson correlation coefficient of plasmid loss time versus total cumulative Cascade concentration at that moment is plotted every 5 minutes (DI) or 10 minutes (Priming) starting from induction of the CRISPR system. The plotted line for both a target with a consensus PAM (green) and target with a mutant PAM (blue) are enveloped by a 95% confidence interval. Darker shading indicates where the correlation coefficient is significantly different from zero ($p<0.05$).

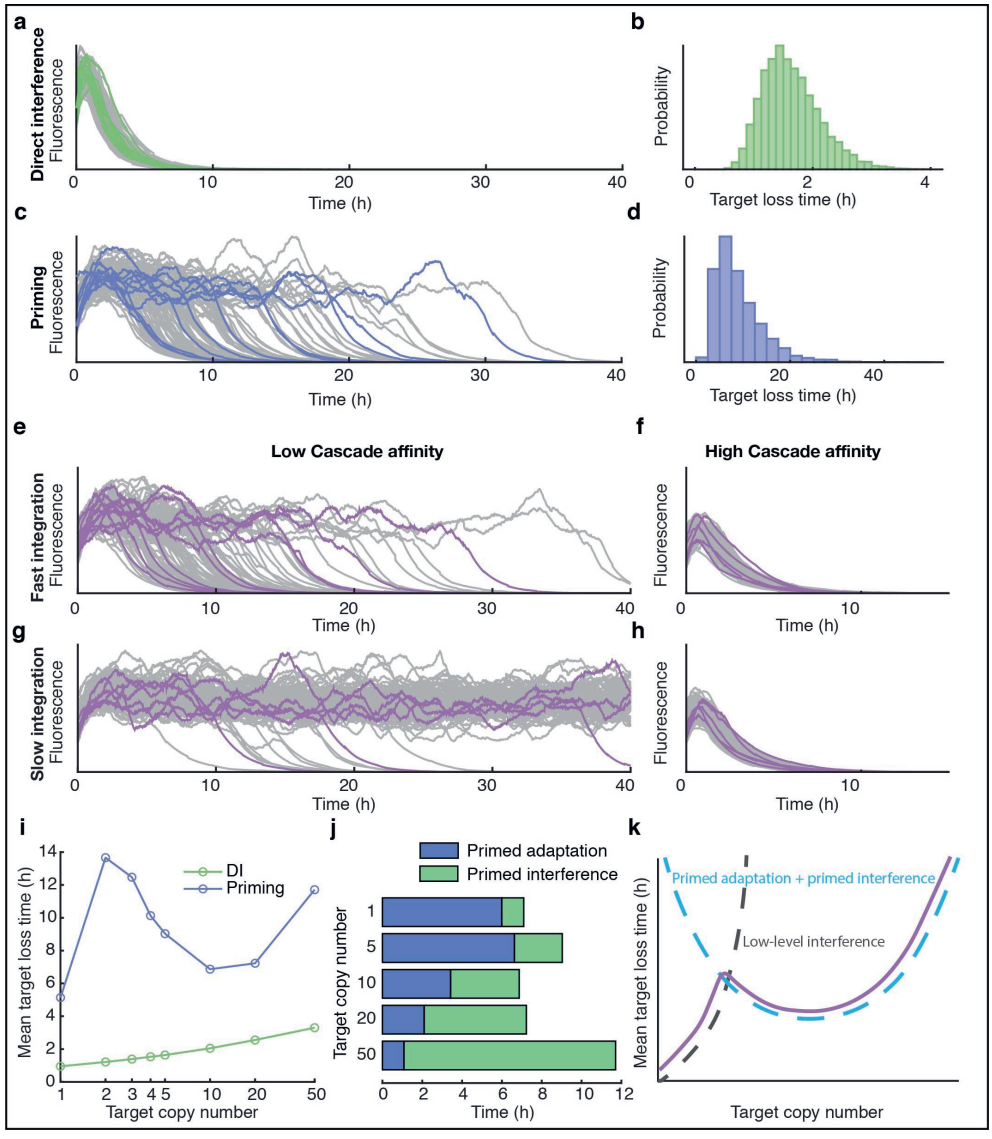


Figure 3.5: Figure caption on next page.

Figure 3.5: Results from the stochastic agent-based model of CRISPR adaptation and interference a-d, Example trajectories showing fluorescence concentration produced by target plasmids simulated with the agent-based model for the (a) direct interference and (c) priming condition, and corresponding target loss distribution (b,d respectively). e-h, Example trajectories from 4 different parameter combinations. High Cascade affinity (f,h) corresponds an increase in target binding by a factor 100 as compared to low Cascade affinity (e,g), slow integration (g,h) represents a 100-fold reduction in the spacer integration rate as compared to fast integration (e,f). i, Mean target loss time of the population as a function of the average target copy number per cell for direct interference (green) and priming (blue). j, Breakdown of average time spent on primed adaptation (blue) and primed interference (green) for cells that clear targets through priming, for target copy numbers in the range 1-50. k, Schematic of alternative target loss pathways. At low copy numbers, targets can be completely cleared through low-level interference, which becomes increasingly rare as copy numbers increase. The priming process shows a u-shaped relationship with the target copy number, as a result of adaptation becoming faster as target copy numbers increase, and time required for interference increasing with target copy number.

continuous source of target DNA fragments that can act as pre-spacers.[109] Hence, we wondered whether target abundance affects this process. For direct interference, as expected, we found that the PLT increased monotonically in simulated trajectories as the average number of targets varies from 1 to 50 (Fig. 3.5i, see Supplementary Fig. 3.14 for full range of distributions). Simulations of priming did not show such a monotonic trend: the PLT first went up, then down, and finally up again (Fig. 3.5i, Supplementary Fig. 3.15). This behaviour could be explained by splitting priming into adaptation and interference (Fig. 3.5j): while primed interference logically only speeds up with fewer targets, primed adaptation initially slows down with fewer targets because of the resulting fewer pre-spacers, but then speeds up for the lowest number of targets, because low-level interference is now sufficiently efficient, in combination with unequal partitioning upon division (Supplementary Fig. 3.16). Indeed, our experiments also showed such clearance of a 5-copy target by low-level interference without spacer acquisition (Supplementary Fig. 3.3). This alternative pathway competes with priming when there are few targets (Fig. 3.5k), and might explain the trend in Fig. 3.5j showing faster loss at 1 target as compared to 5 targets. Target abundance thus affects the balance between primed adaptation and primed interference, resulting in a non-monotonous trend for the target clearance probability.

3.2.9 Cascade expression stochasticity can accelerate CRISPR adaptation

Our experiments showed that CRISPR defence is affected by Cascade expression (Fig. 3.4c-d) which is stochastic in nature (Fig. 3.4b). However, due to the inducible promoter set-up in our experiments, variability in Cascade levels is lower than it might be in a natural setting. To investigate possible implications of this, we changed the level of gene expression variability for Cascade to have 100-fold stronger expression bursts while maintaining average Cascade concentrations (see Supplementary Methods for details). For direct interference simulations, this increased variability resulted in a higher mean PLT: while some cells could clear all targets earlier, many cells required more time to clear all targets as compared to lower-variability Cascade expression (Supplementary Fig. 3.17). Surprisingly, for priming the mean PLT became lower when the Cascade variability increased (Fig. 3.6a). The primed interference phase showed a trend similar to direct interference: a broadening of the PLT distribution yielding a slow-down on average (Fig. 3.6b). However, the entire distribution shifted to lower values for primed adaptation (Fig. 3.6c), yielding an overall speed-up. For mutated PAMs, pre-spacer production critically depends on high Cascade levels, even if transient, as the cumulative probability of a pre-spacer integration event depends on the Cas concentration in a highly non-linear fashion. In Supplementary Methods Fig. 3.1, we demonstrate how this non-linear dependence results in an increased probability of adaptation for cells with high Cascade abundance over a short period of time as compared to cells which have a lower number of Cascades over a longer period of time, despite having equal average Cascade concentrations.

3.3 Discussion

In this study, we have investigated a previously unexplored question: what are the dynamics and variability of the CRISPR adaptation and interference responses in individual cells? Our time-lapse microscopy approach allowed real-time monitoring of invader presence, cell traits and inheritance in single-cell lineages. We found that direct interference, despite its dependence on various stochastic processes and poorly understood tug-of-war between replication of invading nucleic acids and degradation by CRISPR-Cas systems, is notably efficient with invader DNA clearance achieved in all cells within 1 to 3 generations. Conversely, the priming CRISPR response was highly variable, ranging from 2 to 30 generations before clearance. Our data show that direct interference and primed interference can in fact occur on comparable time scales, and identify the adaptation phase of priming as the origin of the variation. Further, our direct observation of the CRISPR-Cas action and modelling approach revealed several factors that impact CRISPR-Cas response variability. The interaction between Cascade and the target DNA, which is

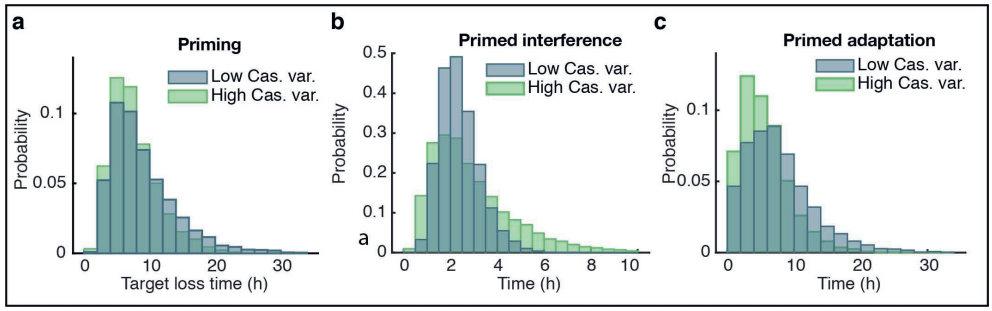


Figure 3.6: Distribution of primed adaptation and primed interference time for high and low variability in Cascade concentration a, Target loss time distribution for two different levels of Cascade concentration variability for priming. At low variability (blue) Cascade proteins are produced in frequent, small bursts, whereas at high variability (green) proteins are synthesized more sporadically in large bursts (100-fold increase), keeping average Cascade concentration constant. b-c, The variability of primed interference times (b) for high Cascade variability (green) increases as compared to low Cascade variability (blue), whereas the variability of primed adaptation times (c) decreases with higher Cascade variability.

characterised by a low affinity owing to the PAM mutation, represents a key source of heterogeneity in the adaptation process of the priming response, rather than the complex spacer integration process.

For direct interference, we found our observed degradation time of 90 mins on average for the consensus target plasmids to be on a comparable time scale to previous work[147] taking into account the differences in experimental setup including but not limited to; the number targeting spacers, copy number of targets and differences in Cas protein induction. While in agreement with the work of Guan *et al.*, we saw that cell size was not an influential factor in the speed of target degradation, we additionally found that cells that cleared the target DNA earlier, grew and divided faster than the population mean. This may be explained by the fact that faster growth is known to reduce plasmid copy numbers.[159,160]

For priming the reverse was found. Cells that adapted and cleared the target DNA earlier, grew slower than the population mean. From this we conjectured that slower growth may play a role due to copy number maintenance mechanisms, which result in higher concentrations of target plasmids in slow growing cells.[161] This hypothesis was further supported by the model which showed that adaptation can occur more efficiently when more target DNA is present in the cell. While in our set-up an effect of Cascade concentration on priming could not be detected, we note that slower growing cells had higher Cascade abundance (Supplementary Fig. 3.18) suggesting that Cascade levels may play a role in combination with other factors enhanced by slow growth. Lastly, slow growth may simply provide

more time to the cell to locate the target and facilitate spacer insertion before interruption by cell division.[144] Conversely, due to our limited ability to determine the precise moment acquisition occurred, we cannot rule out that slow cell growth could be occurring as a result of the spacer integration process itself resulting in a cellular stress response due to DNA damage.[162] However, in line with our findings that slow growth enable cells to prime earlier, spacer acquisition was found to occur more frequently on average in bacteriostatic cells,[163] specifically during the late-exponential and early-stationary phases, when cells are presumably growing slower,[143] and in slow growing populations when compared directly to faster growing populations.[144] Our findings together with these studies indicate that slow growth caused by any environmental change or cellular stress may in fact be beneficial to a cell trying to undergo adaptation.

Our finding that target copy number influences the efficiency of spacer acquisition has implications for scenarios involving invading phages. It suggests that one genome copy of a single virulent phage with an escape PAM may not lead to efficient CRISPR adaptation. However upon replication of the phage genome, it may become abundant enough, though at this point in time it is likely that primed interference with a new spacer cannot successfully eliminate a virulent phage before cell lysis.[134,164,165] Despite this, it has been shown bioinformatically that priming by type I systems is widespread in nature,[166] especially against temperate phages.[167] Such events could occur due to low-level interference, in which a cell is able to simultaneously clear the invader while present as a single copy and acquire a spacer from the fragments produced. This would result in immunisation of a single cell in the population, ultimately leading to a subpopulation of resistant cells that could limit further propagation of the same phage. Such a phenomenon may be more likely to occur when a defective phage infects the cell.[164]

The variation existing between single cells in a population is quite remarkable. In nature, stochasticity or noise in both gene expression and cellular components has been demonstrated to play a positive role in a number of cellular processes.[168–170] Stochasticity in CRISPR-Cas could contribute to bet-hedging strategies,[171] in which subpopulations develop to combat changes in the environment, such as phage exposure. A distinct subpopulation in which Cascade is highly expressed could allow faster elimination of an invading phage, and subsequent re-population. This may in turn increase the fitness of the population, by reducing the overall burden of CRISPR expression and risk of autoimmunity.[137,172] While such a strategy may not guarantee single cell survival, it is at large beneficial for the population as whole. Indeed, previous studies have shown CRISPR immunity in single cells acts to limit phage propagation throughout the population in an abortive infection like manner.[173–175] On the other hand, the survival of only a subpopulation of cells may result in population bottlenecking and an overall loss of diversity.[176] This may be disadvantageous in terms of spacer diversity, where it has been shown that populations containing a range of spacers are better able to combat and even

facilitate the extinction of new invaders.[141,154] Further, we cannot discount that a more susceptible subpopulation of cells may lead to higher overall phage titers and a larger overall threat to the population.

While a number of studies have thoroughly investigated CRISPR-Cas systems through population and single molecule based experiments,[112,114,118,137,142–144,173,174,177,178] these findings do not provide insight into the cell-to-cell variability. Our work, along with others,[146,147] has begun to bridge this gap demonstrating how important the dynamics of CRISPR-Cas systems are to their functioning and the outcome of populations facing a threat. Further investigation into different CRISPR-Cas types and classes, fluctuating environments,[179] and conditions supporting the formation of subpopulations[180] will enhance the understanding of CRISPR-Cas dynamics on both the molecular and population scale.

3.4 Methods

3.4.1 Cloning

Plasmid pTU166 targeted by KD615 and KD635, was created by amplifying the streptomycin resistance cassette from pCDFDuet-1 with primers BN831 and BN832 to add a 5'-CTT-PS8 tail. The backbone of pVenus was amplified using primers BN833 and BN834 and both products were restricted with KpnI and HindIII enzymes. Overnight ligation at 16 °C and transformation into DH5α resulted in colonies selected to contain the plasmid. Plasmids pTU190 and pTU193 were created by PCR amplification of pTU166 using primer BN911 in combination with BN912 or BN891 respectively. Products were restricted with SalI, ligated and transformed into DH5a. Target plasmids pTU389 and pTU390 were PCR amplified from plasmid pTU265 a derivative of pVenus containing CFP using primers BN2278 in combination with BN2275 or BN2276 respectively. Products were restricted with NcoI, ligated and transformed into DH5a. All plasmids were confirmed by Sanger sequencing (Macrogen). All plasmids used are listed in Supplementary Table 3.4. Primers used are listed in Supplementary Table 3.1.

3.4.2 Creation of strains KD615mCherry-Cas8e and KD635mCherry-Cas8e

Strains were created using lambda red homologous recombination.[181] Plasmid pSC020, containing both Lambda red and the Cre-recombinase, was transformed by electroporation into strains KD615 and KD635. Strains were recovered at $30°C$ for 1.5 h and plated on media containing 100 μg/ml ampicillin. Transformants were then grown overnight in liquid medium at $30°C$, with selection, and made competent the following day by inoculating 50 ml with 500 μl of overnight culture. Once the cells reached an OD600 of 0.2 a final concentration of 0.2% L-Arabinose (Sigma-Aldrich)

was added and cells were grown for another 1.5 h and subsequently washed with pre-cooled 10% glycerol. The mCherry-cas8e G-block (IDT) (Supplementary Table 3.5) based on the design used in[110] was resuspended with ddH20 to a concentration of 50 ng/μl and transformed into the competent cells by mixing 2 μl DNA with 50 μl of cells and recovering at $30°C$ for 1.5 h. After recovery cells were plated undiluted with selection for kanamycin and ampicillin. PCR verified colonies were then grown in liquid culture with 1 mM IPTG at $37°C$ to promote the loss of the kanamycin resistance cassette and pSC020. Individual colonies were screened for plasmid loss by patching each colony onto three plates containing no antibiotics, only kanamycin and only ampicillin. Colonies exhibiting no resistance were then PCR screened with primers (Supplementary Table 3.1) BN2204 and BN2205 for the presence of the mCherry-Cascade fusion. Strains were confirmed by Sanger sequencing (Macrogen).

3.4.3 Growth conditions

All strain and plasmid combinations (Supplementary Tables 3.3–3.4) used were grown at 37 °C, shaking at 180 rpm, prior to microscopy. To avoid autofluorescence under the microscope a minimal M9 media was used containing the following supplements,2% glycerol (Sigma-Aldrich), 1X EZ Supplements (M2104 Teknova), 20 μg/ml uracil (Sigma-Aldrich), 1 mM MgSO4 (Sigma-Aldrich) and 0.1 mM CaCl2 (Sigma-Aldrich), from here on called M9 media.

3.4.4 Microfluidic device

The device used was developed by D.J. Kiviet in the Ackermann lab and has been previously used in the Tans lab.[151] The device contains a main flow channel 23.5 μm high and 200 μm wide that splits into two 100 μm wide flow channels of the same height. Perpendicular to these flow channels are wells with a height of 0.75 μm, widths of 1x80 μm, 1x60 μm, 2x40 μm, 3x20 μm, 3x10 μm, 3x5 μm and depths of 60 μm, 30 μm, 50 μm and 40 μm. These well sizes are repeated 5 times and are the location where the growth of microcolonies occurs during an experiment. The PDMS devices were made by casting them into an epoxy mould, a gift from D.J. Kiviet and the Ackermann lab. The PDMS device was produced by mixing polymer and curing agent (Sylgard 184 elastomer, Dow Corning) in ratio of 1 mL of curing agent to 7.7 g of polymer. This mixture was poured into the epoxy mould and air bubbles were subsequently removed by use of a desiccator for 30 mins followed by baking at 80°C for 1 h. After baking the device can be carefully removed from the mould with aid of a scalpel and holes were punched for liquid in-and outlets. For use under the microscope, the PDMS chip was covalently bound to a clean glass coverslip. This was done by treating both the PDMS and glass surface with 5-10 sweeps of a portable laboratory corona device (model BD-20ACV,

Electro-Technic Products). After treatment the chip was placed carefully onto the glass slide and gently tapped to facilitate full contact between the PDMS and glass surface. Finally, the device was baked for another 1-2 h at 80°C and stored until the experiment was started.

3.4.5 Loading and filling of microfluidic wells

Cells were initially grown overnight (for 12 h) at 37°C, 180 rpm in 10 mL M9 media with antibiotic selection (streptomycin 50 μg/ml) for the target plasmid. The following day 500 μl of culture was passaged into fresh M9 medium (with selection for the target plasmid), approximately 3 h before microscope set up, and grown at 37°C, 180 rpm. After 3 h of growth the cells were pelleted and resuspended in 30 μl. To begin the experiment 2 μl of 0.01% Tween20 (dH2O) solution is slowly pipetted into the selected media lane to allow the removal of air and flow of liquid into the wells perpendicular to the media lane. Following this, 2 μl of concentrated bacterial culture was pipetted slowly into the same lane. Once liquid could be seen exiting at the opposite end of the media lane the syringes containing media (loaded on syringe pumps), the valve controller and the waste collection flasks were attached to the chip by metal connectors and polyethene tubing. Media was pumped into the chip at a flow rate of 0.5 mL/h allowing constant supply of nutrients to the cells. The rate of media flow was also important for removal of cells from the top of the well, to allow constant division and long-term tracking of cells located lower within the well.

3.4.6 Media switches

All experiments were carried out with precise 37°C temperature control and required the use of 2 different medias. For the first 12 h of the experiment (including loading of the chip) cells were grown in Media 1,M9 supplemented with both ahydrotetracycline (40 ng/ml) and Streptomycin (25 μg/ml) to induce the YFP and select for cells containing the target plasmid respectively. After 12 h of growth in the chip the media was switched via the valve controller (Hamilton, MPV valve positioner) to Media 2,M9 supplemented with anhydrotetracycline (40 ng/ml), 0.1% L-arabinose and 0.1 mM IPTG. This media change allowed removal of the selection for the target plasmid, continued induction of the YFP and induction of the CRISPR system after filling of the wells.

3.4.7 Spacer acquisition detection from microfluidic chip output

Over the course of the experiment, the cells that flow out of the wells and subsequently the chip were collected in a sterile Erlenmeyer flask. The cells were then concentrated by centrifuging for 5 min at 2000 g. The supernatant was removed

and cell were resuspended in 2 mL of M9 media. Colony PCR was performed with 1 μl of culture using primers BN1530 and BN1531 (Supplementary Table 3.1) and the products were run on a 2% agarose gel at 100 V for 30 mins alongside the 100-1000 bp DNA Ladder (SmartLadder-SF, Eurogentec).

3.4.8 Imaging and image analysis

For all time-lapse experiments, phase contrast images were acquired at 1 min intervals at a maximum of 2 positions. In experiments with a YFP target plasmid, fluorescent images were taken every 2 mins, with an exposure time of 500 ms. For experiments with a CFP target plasmid and the mCherry-Cascade fusion images were acquired every 4 mins with exposure times of 500 ms and 200 ms respectively. Images were acquired for the entire experiment including the first 12 hrs of growth. Cells were imaged with an inverted microscope (Nikon, TE2000), equipped with 100X oil immersion objective (Nikon, Plan Fluor NA 1.3), automated stage (Märzhäuser, SCAN IM 120 3 100), high power LED light source with liquid light guide (Sutter, Lambda HPX-L5), GFP, mCherry, CFP and YFP filter set (Chroma, 41017, 49008, 49001 and 49003), computer controlled shutters (Sutter, Lambda 10-3 with SmartShutter), cooled CMOS camera (Hamamatsu, Orca Flash4.0) and an incubation chamber (Solent) allowing temperature control. In order to obtain images with a pixel size of 0.041 μm an additional 1.5X lens was used. The microscope was controlled by MetaMorph software. A series of acquired phase contrast images were analysed with a custom MATLAB (MathWorks) program, originally based on Schnitzcells software,[150] adapted to allow for automated segmentation of cells growing in a well.[151] Segmentation was inspected and corrected manually where necessary. All segmented cells were then tracked between frames using the pixel overlap between cells allowing the formation of lineage structures.[151] Growth rates are determined by fitting an exponential function to recorded cell lengths over multiple frames and thus represent the rate of cell elongation, whereas interdivision time is calculated as the time between subsequent divisions.

3.4.9 Plasmid loss and clearance time detection using the fluorescent protein production rate

Before screening for plasmid loss, we detect cell death in lineages by applying a moving average filter to the cellular growth rate. If the cellular growth rate reached zero and did not recover again, the remainder of the fluorescence time series after this point was excluded from the analysis. For each lineage, we computed the fluorescence production rate of the plasmid-encoded fluorophore from a cell's total fluorescence, cell area, cellular growth rate, and the rate of photobleaching of the fluorophore.[152] As there is always some amount of residual fluorescence produced

by the cells, we selected an appropriate threshold for plasmid loss detection from the upper values of the distribution of production rates of plasmid-free cells. To detect plasmid loss in individual lineages we applied a moving average filter to the fluorescence production rate and detected the first instance of the production rate reaching a value below the threshold. This plasmid loss time (PLT) can be seen as an upper bound estimate, as some processes (transcription, translation, fluorophore maturation) still carry on for some time after the last plasmid has been cleared but could not be measured in our set up. The onset of the clearance time (CT), which signifies the start of the destruction of all plasmids through interference and ends at the plasmid loss time (PLT), is difficult to detect in individual lineages due to the naturally occurring fluctuations in the fluorescence production rate. To determine this quantity, we align all plasmid loss lineages at the PLT and compute the average trend. The CT per experimental condition is approximated as the duration from the point where the average production rate starts to decrease until the PLT.

3.4.10 Sister and cousin statistics

For each lineage that lost the plasmid, we wanted to compare the probability of loss in an unrelated cell and in a related cell. For related cells we counted the frequency of loss and non-loss in sister and cousin cells of the loss cell, but only if the sister or cousin divided (contained a complete cell cycle). For unrelated cells we counted the total number of loss events (i) that occurred throughout the cell cycle of the related cell. For each loss event we counted how many cells (c_i) still contained the plasmid up to that point. The probability of plasmid loss happening in an unrelated cell during the lifecycle of the related cell was subsequently calculated recursively using the following equations:

$$p_0 = 0 \tag{3.1}$$

$$p_i = (1 - p_{i-1})/c_i + p_{i-1} \tag{3.2}$$

Where p_i is the probability of loss occurring within an unrelated cell given i plasmid loss events occurred within the cell cycle of the related cell and c_i stands for the number of cells still containing the plasmid at the same time as the i-th plasmid loss event.

3.4.11 Cascade copy number determination

The control strain KD614 mCherry-Cas8e containing plasmid pTU265 (Supplementary Tables 3.3–3.4) was prepared and loaded into the microfluidic chip as above. After 12 h a sterile tube was connected to the waste tubing and output from the chip was collected for 30 mins. The media was then switched to induce Cascade. Approximately 5 h after induction when Cascade levels are considered to

be stabilized the output from the chip was again collected for 30 mins. To improve counting, cells were subsequently fixed with 2.5% paraformaldehyde solution at 22 °C for 45 mins.[182] Slides were cleaned by sonication in subsequent steps with MilliQ, acetone and KOH (1M). Next, 1% agarose pads containing the M9 medium were prepared and hardened between two slides within 20 mins of measuring to prevent desiccation. The fixed cells were then spun down and resuspended in 5 μl of which 1 μl was pipetted onto a pre-prepared agarose pad.

The cells were imaged using a TIRF microscope (Olympus IX81,Andor Ixon X3 DU897 EM-CCD camera) using a high power 561 nm laser, which quickly bleached most mCherry molecules within a couple of frames. Intensity of single molecules were measured with Thunderstorm starting from the thirtieth frame.[183] The total cell fluorescence was measured by segmenting the cells from the phase contrast image and sum fluorescence counts of all cell pixels (with background subtracted). The copy number was calculated by dividing the total cell fluorescence in the first frame by the average fluorescence intensity of the single molecules. We could then calculate the Cascade concentration 200 Cascade molecules/μm2 by dividing the population average of the mean summed RFP per cell by this copy number, which was applied to the cells in our time-lapse data.

3.4.12 Model implementation

Stochastic simulations were performed using the adapted Extrande algorithm[184] implemented in C++. Each data point in Fig. 3.5i-j and Fig. 3.6a-c was obtained from 100 simulated experiments of up to 104 min. The population size of each simulation was fixed at 100 cells. See Supplementary Methods for model details and parameters.

3.4.13 Acknowledgements

The authors would like to thank Martijn Wehrens for his help and advice throughout the project and all of the members of the Tans and the Brouns groups for input during group discussions. We acknowledge the group of Konstantin Severinov for the gift of strains KD615 and KD635. R.E.M. is supported by the Frontiers of Nanoscience (NanoFront) program from NWO/Ministry of Education (OCW). C.F. received funding from FET-Open research and innovation actions grant under the European Union's Horizon 2020 research and innovation programme (CyGenTiG; grant agreement 801041). Work in the group of S.J.T. is supported by the Netherlands Organization for Scientific Research (NWO). S.J.J.B. has received funding from the European Research Council (ERC) CoG under the European Union's Horizon 2020 research and innovation programme (grant agreement No. [101003229]) and from the Netherlands Organisation for Scientific Research (NWO VICI; VI.C.182.027).

3.4.14 Data availability

Data analysis was performed using custom MATLAB scripts, which can be found at `https://github.com/TansLab/Tans_Schnitzcells`. Scripts for lineage analysis and plotting were implemented in MATLAB and are available upon request. An implementation of the agent-based model is available at `https://git.wur.nl/Biometris/articles/CRISPR_ABM`.

Appendices

3.A Master Equation description of the probability of plasmid loss

In order to test whether the distribution of the target clearance times by direct interference can be reproduced by a simple one-step process, we consider a model using a compound probability for binding of Cascade to the target and subsequent target removal from the system. In bacteria the number of targets is subject to maintenance which delays the removal of M_0 targets. For sake of simplicity we ignore this additional step, which has the advantage that the number of unknown parameters is kept to an absolute minimum. Because direct interference is a fast process, one can assume that target maintenance does not have a strong effect on the clearance time distribution. The Cascade number is not constant, but rather Cascade production is induced at the beginning of the experiment. This simplified model only depends on five parameters: the delay after induction for production of Cascade τ_c, the Cascade production rate σ, the turn-over rate of Cascade λ, the number of targets per cell M, and the probability of a target removal event p_d. The number of targets in individual cells will be in general stochastic, however due to target maintenance one can assume that this distribution will be quite narrow. For this reason, we set $M_0 = 5$.[185]

The time dependent Cascade copy number is modelled as a production-degradation process with a delay τ_c and zero initial amount of Cascade: The bulk mean $\mu(t)$ is given by:

$$\mu(t) \quad = \quad \frac{\sigma}{\lambda}\theta(t-\tau_c)\left(1-e^{-\lambda(t-\tau_c)}\right).$$

By fitting this equation to Cascade concentration data for the bulk mean (Fig. 3.4b), we estimate: $\tau_c = 34$ min, $\sigma = 3$ min^{-1}, and $\lambda = 0.0061$ min^{-1} to obtain an average copy number of almost 500 Cascades per cell at steady state.

The removal of M_0 targets from the system is a First-Passage-Time problem.

We formulate the simple Master Equation (ME) for the conditional probability $P_M(t)$ to find M targets in a cell at a given time t:

$$\frac{dP_M(t)}{dt} = \mu(t)p_d(M+1)P_{M+1} - \mu(t)p_d M P_M,$$

where p_d is the compound probability that within the time interval Δt a Cascade molecule binds to a target and the target is subsequently removed from the system.

To obtain the First-Passage-Time distribution we need to determine the survival probability S to find at least one target, which is simply given by $S = 1 - P_0$. P_0 is obtained by solving the above ME with the initial condition $P_M(t=0) = \delta_{MM_0}$:

$$P_0(t|M_0) = \left[1 - e^{-p_d \int_0^t \mu(t')dt'}\right]^{M_0}.$$

$P_0(t|M_0) = 0$ for $t < \tau_c$ and because the state $M = 0$ is naturally an adsorbing boundary we readily find $\lim_{t\to\infty} P_0(t|M_0) = 1$. The First-Passage-Time distribution $FP_r(t|M_0)$ for target removal is given by the $FP_r = -dS/dt = dP_0/dt$:

$$FP_r(t|M_0) = M_0 p_d \mu(t) \left[1 - e^{-p_d \int_0^t \mu(t')dt'}\right]^{M_0-1}.$$

Fitting this distribution to the empirical data (Fig. 3.2d) gives rise to $p_d = 4.4 \times 10^{-4}$ min^{-1}. The average target removal time τ is given by:

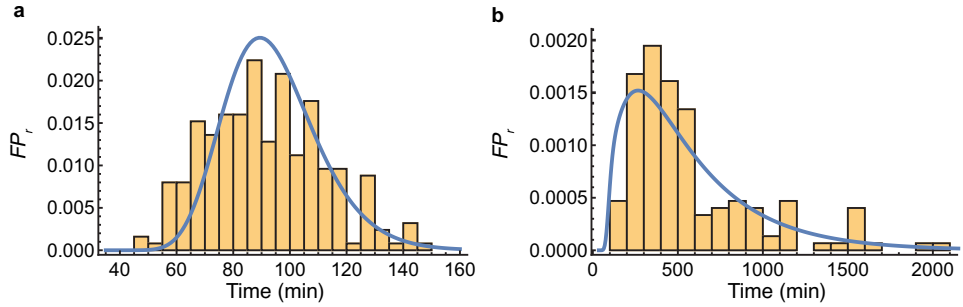
$$\tau = \int_0^\infty t' FP_r(t'|M_0)dt'.$$

Using the estimates for p_d, σ, λ, τ_c, and $M_0 = 5$ we obtain $\tau \approx 94$ min. The fit of FP_r to the data can be seen in Supplementary Methods Fig. 3.1a.

The simplified model yields a decent fit to the direct interference data. What about the target clearance during priming? To investigate whether this can be conceptually understood by a two-step process, first spacer acquisition subsequently followed by primed interference, we condition FP_r on the time τ_p needed for spacer acquisition:

$$FP_r(t|M_0, \tau_p) = M_0 p_d \theta(t-\tau_p)\mu(t) \left[1 - e^{-p_d \int_{\tau_p}^t \mu(t')dt'}\right]^{M_0-1}.$$

The rationale behind this is that a Cascade molecule needs to bind to a target to produce the pre-spacers necessary for spacer acquisition before primed interference can happen. It follows $FP_r(t|M_0, \tau_p) = 0$ for $t < \tau_p$. Note that $\tau_p \geq \tau_d$, since in



Supplementary Methods Figure 3.1: The fits of the simple one or two-step model to the data. **a**, fit of FP_r (solid line) to the target removal time in the case of direct interference. **b**, fit of FP_r (solid line) to the target removal time in the case of priming.

the absence of Cascade the probability of spacer acquisition is negligibly small. The distribution for τ_p is given by the First-Passage-Time distribution for the passage $M_0 \rightarrow M_0 - 1$: $FP_p = -dP_{M_0}/dt$:

$$FP_p(\tau_p|M_0) \quad = \quad M_0 p_p \mu(\tau_p) e^{-M_0 p_p \int_0^{\tau_p} \mu(t')dt'},$$

where p_p is the compound probability that within the time interval Δt one Cascade binds to a target, pre-spacers are produced and a spacer is integrated.

The distribution for the target removal times is given by:

$$FP_r(t|M_0) \quad = \quad \int_0^\infty FP_r(t|M_0 - 1, \tau_p) FP_p(\tau_p|M_0) d\tau_p.$$

The integral cannot be done analytically. Fitting $FP_r(t|M_0)$ to the experimentally obtained data for the distribution of target loss times during priming (Fig. 3.2e) yields $p_p = 10^{-6}$ min^{-1}. The fit of FP_r to the data can be seen in Supplementary Methods Fig. 3.1b.

3.B An agent-based model for stochastic biochemical kinetics of cell populations in microfluidic wells

Although a highly simplified description of our system, the results from the ME description show that the Cascade copy number is an important determinant in creating the variability in the PLT distribution in the case of direct interference. For priming, the distribution could be replicated by considering the process as the

result of two subsequent steps, of which the spacer acquisition process creates the wide PLT distribution. However, this model of primed adaptation is highly simplified and does not give any mechanistic insight into the process of adaptation and interference in a growing cell population. To better understand how cell-to-cell variability and population dynamics affect CRISPR-Cas defense, we have developed a stochastic, agent-based simulation framework to analyse the kinetics of spacer acquisition and target loss. An agent-based approach allows us to keep track of the biochemical composition of individual cells in a growing population, as well as the inheritance of molecules and other cellular features in lineages. In this type of model, each cell is an agent, and there is no interaction between cells. For computational efficiency and to emulate the experimental set-up, the size of the cell population is kept constant. Results for this type of set-up, where the population size is constant, are identical as for a population experiencing exponential population growth, as long as the population size is sufficiently large (100-1000 cells).[17] The intracellular reactions are governed by stochastic reaction kinetics which can be described by the Chemical Master Equation (CME). As an exact solution to the CME exists only for a handful of simple reaction networks, we use the stochastic simulation algorithm (SSA),[5] which provides trajectories which are consistent with the CME provided the rate constants are time-independent. When a reaction involves more than one molecular species, the propensity for this reaction to take place in some small time interval depends on the cellular volume. In our application, we are dealing with cells that are continuously in the exponential growth phase which violates the assumption of the SSA of constant propensities between reaction events. For this reason, we use the *Extrande* extension by Voliotis *et al.*, which allows us to efficiently simulate the reaction network containing time-dependent propensities.[184]

3.B.1 Model assumptions

Since the detailed mechanism of primed spacer acquisition in type I-E CRISPR-Cas systems is not yet completely known, we start out with a simplified model to see if this is sufficient to explain our data. Because primed adaptation is much more efficient than naive adaptation,[129] we assume that the rate of naive adaptation is negligibly small over the time course of the experiment. The spacer composition of the CRISPR array is not modeled in detail. Rather, we assume that we start out with a crRNA sequence that matches the target, but is flanked by a non-consensus PAM. The effector complexes containing this spacer can still bind to the target DNA,[112,127] but with a binding affinity that is decreased up to a factor $100 - 150$ as compared to binding with a consensus PAM.[157,158] Once the effector complex is bound to the target, Cas3-catalysed destruction of the target takes place.[186] Thus, the level of interference is associated with the level of effector complex binding.[157]

Cas3-mediated destruction of targets is a source of substrates for spacer acquisition machinery, the Cas1-Cas2 complex, during primed adaptation.[109,130]

Intermediates of target DNA degradation are transient and quickly degrade after an initial burst. Abundant levels of Cas1 and Cas2 lead to robust spacer acquisition, by allowing Cas1-Cas2 to capture the transient intermediates of Cas3 action.[130] Since in our system Cas3, Cas1, and Cas2 are highly expressed, we assume the levels of these proteins are not rate-limiting within the scope of our model and thus do not explicitly model their abundances. Furthermore, in agreement with previously published work, we assume cells have a target maintenance system that is controlled by logistic dynamics in order to keep the target concentration at its target level.[165] In addition, targets and target-containing configurations are actively partitioned between daughter cells[187,188] according to a multi-hypergeometric distribution, with each daughter receiving on average half of the mother cell's targets. All other proteins are partitioned according to a Binomial distribution, where the ratio of daughter cell sizes determines the probability of each molecule ending up in one of two daughter cells. We model synthesis of CRISPR proteins as a Poisson process, in which proteins are produced in geometrically distributed bursts to capture the effect of transcriptional bursting.[189] We assume all molecular species are stable on the timescale of the experiment (i.e. no degradation), with the exception of the free crRNAs (not loaded in Cascade) and the DNA fragments that are the result of interference, which have a short lifetime.

3.B.2 Algorithm outline

For the agent-based model, we have adapted the First-Division Algorithm by Thomas[23] to include the *Extrande* extension to the SSA. Furthermore, we keep the population size constant by randomly selecting a cell to be removed from the population in the event of a cell division. The steps to replicate our experimental set-up are described below.

1. **Population initialization:** At time $t = 0$, initialize N cells by assigning to each cell an age $t_i \sim \mathrm{U}(-\log(2)/\mu_p, \log(2)/\mu_p)$, a growth rate $\mu_i \sim \mathrm{Lognormal}(\mu_p, \sigma_p^2)$ and molecule count x_i. Select division size $V_{d,i} \sim \mathrm{Lognormal}(\mu_{V_D}, \sigma_{V_D}^2)$ and compute generation time $t_{gen,i}$ as $\log(V_{d,i}/V_{b,i})/\mu_i$, where $V_{b,i}$ is the birth size. This determines the division time of the cell which is defined as $t_{d,i} = t_i + t_{gen,i}$.

2. **Biochemical reactions:** Determine the next dividing cell: $j = \mathrm{argmin}_i(t_{d,i} - t_i)$. Determine Δt from $\min(t_{d,j} - t_j, L)$, where L is *Extrande*'s look-ahead horizon. Advance the molecule numbers of each cell independently from age t_i to $t_i + \Delta t$ using the *Extrande* algorithm and advance time from t to $t + \Delta t$.

3. **Cell division:** When $t = t_{d,j}$, replace the dividing cell by two newborn daughter cells of zero age. The birth size of both daughters is determined as $V_{b,D_1} = \mathrm{Normal}(\mu_{V_B}, \sigma_{V_B}^2)V_{d,j}$ and $V_{b,D_2} = V_{d,j} - V_{b,D_1}$. Assign to one of

these a molecule number distributed according to the Binomial distribution (proteins) and the Multi-hypergeometric distribution (targets and target configurations), depending on the mother's molecule count x_j and the daughter's size ratio to the mother cell $\frac{V_{b,D_1}}{V_{d,j}}$, and assign the remaining molecules to the other daughter. Assign to each daughter independently a growth rate μ_i, division volume $V_{d,i}$, and compute corresponding division time. To ensure a constant population size, randomly select a cell to be deleted from the population.

4. **Repeat:** Repeat from 2. until $t = t_{\text{final}}$.

3.B.3 Molecular mechanism and model parameters

Each cell in the population contains a pool of biochemical species that can interact with each other through biochemical reactions, as described in step 2. We distinguish between the targets P, the CRISPR array A, which codes for a spacer $crRNA$ matching a sequence on the target, and the surveillance protein $Cascade$. Plasmids P have a maintenance mechanism in order to keep the plasmid copy number at the target concentration p^*/V_B, where V_B is the average cell volume at birth. Together with the crRNA, the Cascade protein makes up the effector complex E. When the effector complex encounters a target it can bind, albeit with a low affinity in the case of a non-consensus PAM on the target, forming a complex EP. Destruction of the target can then take place, producing DNA fragments F. One of these fragments can be integrated into the CRISPR array A as a new spacer, transforming the array to A^* which can now also express the newly acquired crRNA, $crRNA^*$, in addition to the spacer that was already present. The effector complex containing the new spacer, E^* has a higher binding affinity for the target. These biochemical reactions are governed by the equations described in Supplementary Table 3.1.

The size of individual cells increases exponentially with a constant elongation rate throughout the cell cycle. Cellular length is used as a measure for cell size, as *E. coli* cell width remains approximately constant throughout the cell cycle and thus the cellular volume is linearly proportional to the cell length.[190] Growth parameters were chosen to be representative for our experimental data. As no kinetic data are available on individual reactions of the adaptation and interference processes, these parameters were calibrated to qualitatively agree with the experimentally determined target loss time distributions from the direct interference and priming conditions and previously published abundances of *cas* abundances.[191] Unless stated otherwise, the growth parameters used were $\mu_p = \log(2)/70$, $\sigma_p = 0.2$, $\mu_{V_B} = 0.5$, $\sigma_{V_B} = 0.07 \cdot \mu_{V_B}$, $\mu_{V_D} = 3.9$, $\sigma_{V_D} = 0.11 \cdot \mu_{V_B}$, $p^* = 5$. To simulate the direct interference condition with the same model, we simply modify the initial state of the system such that the spacer array consists of crRNA^*, which is flanked by the consensus PAM sequence.

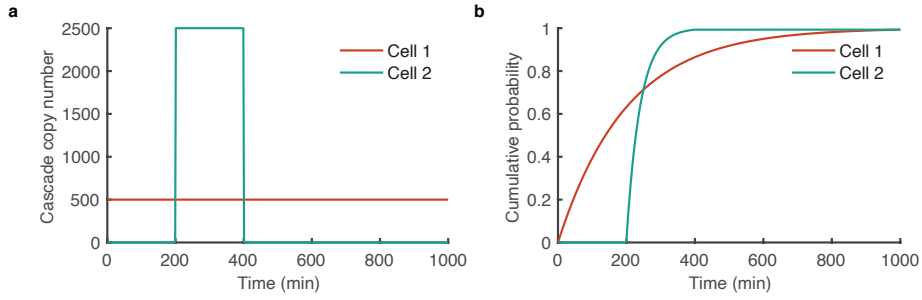
3.B.4 Cascade variability impacts the probability of spacer acquisition

In the main text of the manuscript we have shown that in priming, increased variability in the expression of Cascade can lead to faster spacer acquisition on average (Fig. 3.5c). In simulations of the agent-based model, variability of the Cascade protein concentration is controlled through the protein production rate k_1 in coordination with the average protein burst size b_c: to modify Cascade variability while maintaining a constant concentration, b_c is multiplied by a factor a while k_1 is multiplied by its inverse, $\frac{1}{a}$. In Fig. 3.5, $a = 100$ which leads to an increase of the coefficient of variation of the Cascade concentration at steady state from $CV = 0.02$ (low Cascade variability) to $CV = 0.42$ (high Cascade variability).

We will now illustrate how higher Cascade variability can lead to faster spacer acquisition by considering two scenarios, and comparing the cumulative probability of the time until spacer acquisition for the simplified two-step model, which is given by

$$FP_{SA}(t|M_0) = 1 - e^{-M_0 p_p \int_0^t \mu(t')dt'}.$$

First, we consider a cell which has a constant Cascade level of 500 copies at any point in time between $t = 0 - 1000$ min, and plot the corresponding cumulative spacer acquisition probability (Supplementary Methods Fig. 3.2a). Second, we consider a second cell in which Cascade is not constant but rather appears as a shorter 'burst' of 2500 copies from $t = 200$ min until $t = 400$ min, and 0 copies at any other time (Supplementary Methods Fig. 3.2a). The cumulative spacer acquisition probability for the second cell reaches 1 faster than for the first cell (Supplementary Methods Fig. 3.2b), despite the two cells having the same average Cascade concentration over the course of 1000 minutes. This suggests that the effects of upwards fluctuations can outweigh the downward fluctuations.



Supplementary Methods Figure 3.2: **a**, Cascade copy number of two cells with the same average over time. Cell 1 has a constant copy number of 500 Cascades, while in cell 2 Cascade is present transiently at 2500 copies between 200 and 400 minutes. **b**, Cumulative probability of time until spacer acquisition for the two cells with Cascade copy numbers as described in panel a ($M_0 = 1$, $p_p = 0.00001$).

Phase	Reactions
Target replication	$P \xrightarrow{k_0/(1+((P/V_t)/p_0)^2)} 2P,$ $p_0 = \frac{p^*}{V_B}/\sqrt{(\frac{k_0}{\mu} - 1)}$
Expression	$G \xrightarrow{k_1(t)} G + b_P \cdot Cascade$ $k_1(t) = \frac{k_1}{1+exp(-k_d t)}$ Before spacer integration $A \xrightarrow{k_2} A + b_c \cdot crRNA$ After spacer integration $A^* \xrightarrow{k_2} A^* + b_c \cdot crRNA + b_c \cdot crRNA^*$ $crRNA + Cascade \xrightarrow{k_3} E$ $crRNA^* + Cascade \xrightarrow{k_3} E^*$
Interference	$E + P \underset{k_5}{\overset{k_4}{\rightleftharpoons}} EP$ $E^* + P \underset{k_7}{\overset{k_6}{\rightleftharpoons}} EP^*$ $EP \xrightarrow{k_8} E + b_F \cdot F$ $EP^* \xrightarrow{k_8} E^* + b_F \cdot F$ $F \xrightarrow{k_9} \emptyset$
Primed adaptation	$F + A \xrightarrow{k_{10}} A^*$

Supplementary Table 3.1: Overview of the reactions in the model for primed adaptation.

Reaction	Parameter	Value
Target replication	k_0	0.125 min^{-1}
$Cascade$ production	k_1	2.4 min^{-1}
$crRNA/crRNA^*$ transcription	k_2	10 min^{-1}
$crRNA/crRNA^*$ degradation	k_3	0.014 min^{-1}
$crRNA - Cas/crRNA^* - Cas$ effector complex formation	k_4	0.01 $\mathrm{M}^{-1}\mathrm{min}^{-1}$
$E - P$ binding affinity	k_5	$1e^{-5}$ $\mathrm{M}^{-1}\mathrm{min}^{-1}$
EP dissociation	k_6	$1e^{-4}$ min^{-1}
$E^* - P$ binding affinity	k_7	$1e^{-3}$ $\mathrm{M}^{-1}\mathrm{min}^{-1}$
EP^* dissociation	k_8	$1e^{-4}$ min^{-1}
Target degradation	k_9	1 min^{-1}
Fragment degradation	k_{10}	1 min^{-1}
Spacer integration	k_{11}	0.25 $\mathrm{M}^{-1}\mathrm{min}^{-1}$
$Cascade$ burst size	b_P	3
$crRNA/crRNA^*$ burst size	b_c	3
DNA fragment burst size	b_F	5
Post-induction delay of protein production	k_d	0.025 min^{-1}

Supplementary Table 3.2: Reaction rates used in simulations

3.C Strains and plasmids used in this study

Strain	Description	Source
KD615	*E. coli* K12, F+, ΔaraBAD, araBp8-*cse1*, lacUV5-*cas3*, CRISPR I R-SP8-R, ΔCRISPR II+III	(Datsenko et al., 2012; Musharova et al., 2019)
KD635	*E. coli* K12, F+, ΔaraBAD, araBp8-*cse1*, lacUV5-*cas3*, CRISPR I R-SP8-R, $\Delta cas1,2$, ΔCRISPR II+III	(Datsenko et al., 2012; Musharova et al., 2019)
KD615mCherry-Cas8e	*E. coli* K12, F+, ΔaraBAD, araBp8-*cse1*, lacUV5-*cas3*, CRISPR I R-SP8-R, ΔCRISPR II+III, mCherry-*cas8e*	This study
KD634mCherry-Cas8e	*E. coli* K12, F+, ΔaraBAD, araBp8-*cse1*, lacUV5-*cas3*, CRISPR I R-SP8-R,$\Delta cas1,2$, ΔCRSIPR II+III, mCherry-*cas8e*	This study
pTarget (pTU166)	pSC101, StrepR, TetR mVenus PS8 flanked by 'CTT' PAM	This study
pMutant (pTU190)	pSC101, StrepR, TetR mVenus PS8 flanked by 'CGT' PAM	This study
pControl (pTU193)	pSC101 ori, StrepR, TetR-mVenus, no target	This study
pVenus	pSC101 ori, KanR, mVenus-YFP	Bokinsky lab
pCDFDuet-1	pCloDF13 ori, StrepR	Lab collection
pTU265	pSC101, StrepR, TetR-Cerulean, no target	This study

Supplementary Table 3.3: Strains used in this study

Plasmid	Description	Source
pTU389	pSC101, StrepR, TetR-Cerulean, PS8 flanked by 'CGT' PAM	This study
pTU390	pSC101, StrepR, TetR-Cerulean, PS8 flanked by 'CTT' PAM	This study
pSC020	Derivative of pKD46 containing Lambda red and the Cre-recombinase	Lab collection

Supplementary Table 3.4: Plasmids used in this study

3.D Oligonucleotides used in this study

Supplementary Table 3.5: Oligonucleotides used in this study. PAM sequences are indicated in bold, and restriction sites are underlined.

Name	Description	Sequence
BN831	Streptomycin resistance and PS8 insertion into pVenus, Fw	TTTTGGTACCTTATTT GCCGACTACCTTGGTG ATCTC
BN832	Streptomycin resistance and PS8 insertion into pVenus, Rv	TTTTAAGCTTAAAAG TGCCACTTGCGGAGA CCCGGTCGTCAGCTT ACATTCAAATATGTA TCCGCTC
BN833	Backbone amplification pVenus, Rv	TTTTGGTACCGGACTC TGGGGTTCGAG
BN834	Backbone amplification pVenus, Fw	TTTTAAGCTTCGAAAC GATCCTCATCCTG
BN891	Streptomycin resistance insertion (no target), Rv	TTTTAAGCTTACATTC AAATATGTATCCGCTC
BN911	Modify pTU166 PAM universal, Rv	TTTTGTCGACACATTC AAATATGTATCCGCTC ATGAGAC
BN912	Modify pTU166 CTT PAM to CGT	TTTTGTCGAC**ACG**CTG ACGACCGGGTC
BN1494	To amplify pTU193 Backbone minus yfp, Rv	TTTCTCGAGTAAGGAT CTCCAGGCATC
BN1495	To amplify pTU193 Backbone minus yfp, Fw	TTTCTCGAGTAAGGAT CTCCAGGCATC
BN1507	To amplify Cerulean from p15A, Fw	TTTGAATTCCAGAATT CAAAAGATCTAGGAGG
BN1508	To amplify Cerulean from p15A, Rv	TTTCTCGAGAGGATCC TTATTTATACAGCTCAT CC
BN1513	To check Cerulean insertion and confirm pTU265 by sequence, Fw	CCTCATTAAGCAGCTC TAATGCGCTG

Continued on next page

Supplementary Table 3.5: Oligonucleotides used in this study. PAM sequences are indicated in bold, and restriction sites are underlined. (Continued)

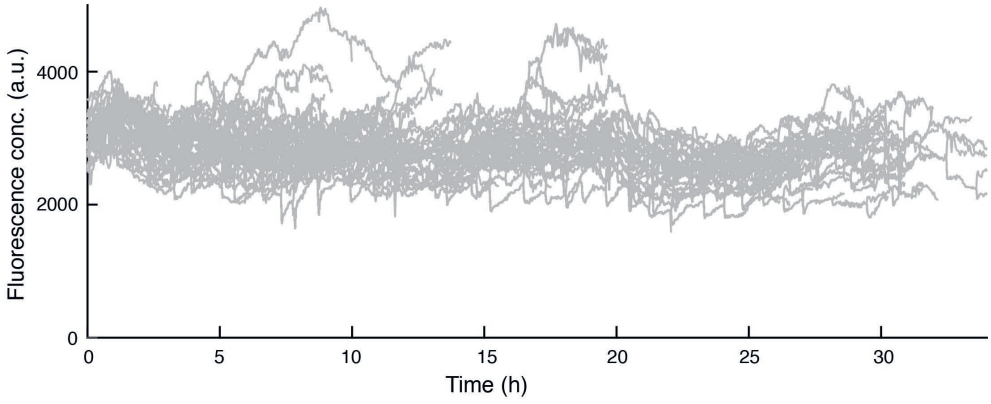
BN1530	To screen for CRISPR array amplification, Fw	GGTTTGAAAATGGGAG CTCG
BN1531	To screen for CRISPR array amplification, Rv	GTTACATTAAGGTTGG TGGGTTG
BN2202	To amplify mCherry-Cas8e gblock, Fw	ACAGAATCTGGATGGA TGG
BN2203	To amplify mCherry-Cas8e gblock, Rv	CTGATCTCTACTGCAGT ATAGC
BN2204	Screen for mCherry-cas8e knock in, Fw	GCGCTTGCACTTAATCGC
BN2205	Screen for mCherry-cas8e knock in, Rv	ACCAGCAGTGCTAAAGCG
BN2206	Screen for mCherry-cas8e knock in, Fw	CTTTCCGTCCGGTGTC AGG
BN2275	Insertion PS8 CGT PAM into pTU265, Fw	TTTCCATGGAAAAGTG CCACTTGCGGAGACCC GGTCGTCAG**CGT**ACA TTCAAATATGTATCCGC TCAT
BN2276	Insertion PS8 CTT PAM into pTU265, Fw	TTTCCATGGAAAAGTG CCACTTGCGGAGACCC GGTCGTCAG**CTT**ACAT TCAAATATGTATCCGCT CAT
BN2278	Insertion of PS8 and PAM universal, Rv	TTTCCATGGCCTCATCC TGTCTCTTGATC

3.E Synthetic DNA G-block used in this study

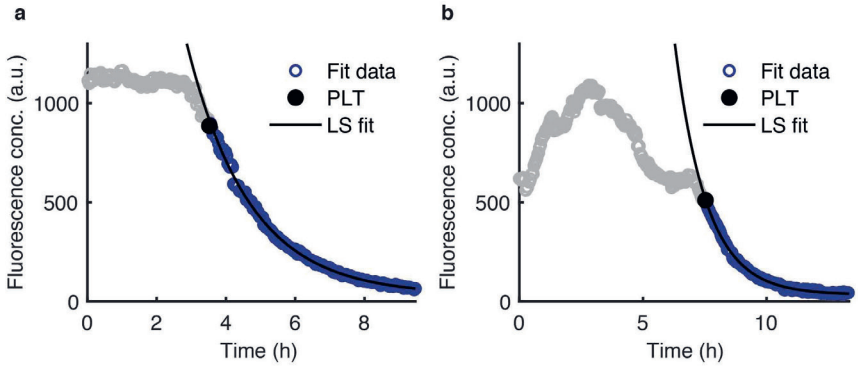
Name	Sequence
cas8e-mCherry insert	ACAGAATCTGGATGGATGGGTCTGGCAGGGTAACAGTA TTGTTATTACCTATACAGGGGATGAAGGGATGACCAGAG TCATCCCTGCAAATCCCAAATAACCTGGAGCTGCAGATA CCGTTCGTATAATGTATGCTATACGAAGTTATAGATCTCTA TTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTC ATGAGACAATAACCCTGATAAATGCTTCAATAATATTGA AAAAGGAAGAGTATGAGCCATATTCAACGGGAAACGT CTTGCTCTAGGC CGCGATTAAATTCCAACATGGATGC TGATTTATATGGGTATAAATGGGCTCGCGATAATGTCGG GCAATCAGGTGCGACAATCTATCGATTGTATGGGAAGC CCGATGCGCCAGAGTTGTTTCTGAAACATGGCAAAGGT AGCGTTGCCAATGATGTTACAGATGAGATGGTCAGACTA AACTGGC TGACGGAATTTATGCCTCTTCCGACCATCAA GCATTTTATCCGTACT CCTGATGACGCATGGTTACTCAC CACTGCGATCCCCGGGAAAACAGCATTCCAGGTATTAG AAGAATATCCTGATTCAGGTGAAAATATTGTTGATGCGCT GGCAGTGTTCCTGCGCCGGTTGCATTCGATTCCTGTTT GTAATTGTCCTTTTAACAGCGACCGCGTATTTCGTCTCGC TCAGGCGCAATCACGAATGAATAACGGTTTGGTTGATGC GAGTGATTTTGATGACGAGCGTAATGGCTGGCCTGTTGA ACAAGTCTGGAAAGAAA TGCACAAACTTTTGCCATTCTC ACCGGATTCAGTCGTCACTCATGGTGATTTCTCACTTGAT AACCTTATTTTTGACGAGGGGAAATTAATAGGTTGTATTG ATGTTGGACGAGTCGGAATCGCAGACCGATACCAGGAT CTTGCCATCCTATGGAACTGCCTCGGTGAGTTTTCTCCTT CATTA CAGAAACGGCTTTTTCAAAAATATG- GTATTGATAAT CCTGATATGAATAAATTGCAGTTTCATTTGATGCTCGATGA GTTTTTCTAAGTCGACATAACTTCGTATAATGTATGCTATAC GAACGGTAGAAATTGCAATGCATCTGCCGAATGCCGTGTG GACGTAAGCGTGAACGTCAGGATCACGTTTCCCCGACCC GCTGGCATGTCAACAATACGGGAGAACACCTGTACCGCC TCGTTCGCCGCGCCACCATAAATCACCGCACCGTTCATC AGTACTTTCAGATAACACATCG

Supplementary Table 3.5: Synthetic DNA G-block used in this study

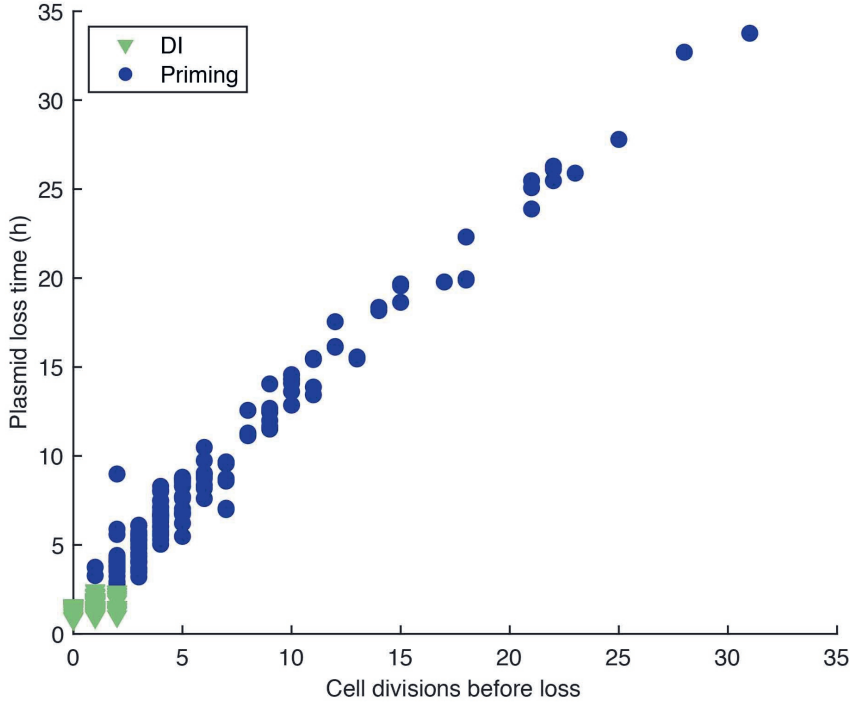
Supplementary Figures



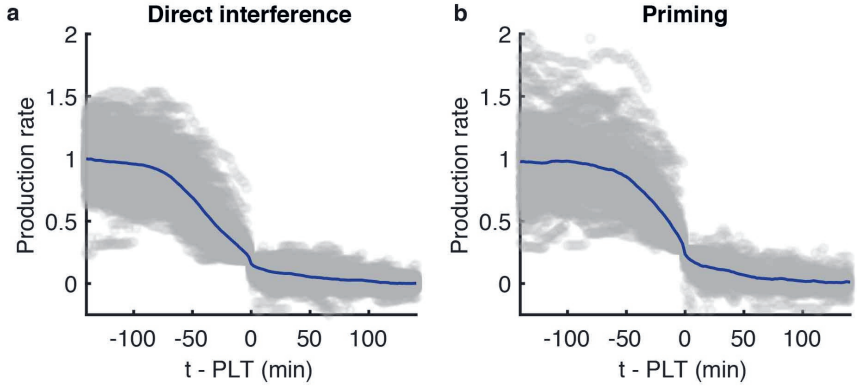
Supplementary Figure 3.1: Plasmid loss is CRISPR-dependent The YFP fluorescence traces in arbitrary units (a.u.) of the WT strain harbouring pControl a plasmid with no target for the CRISPR-Cas system. Time-lapse imaging was carried out for 35 hours post induction of the *cas* genes, and showed no plasmid loss.



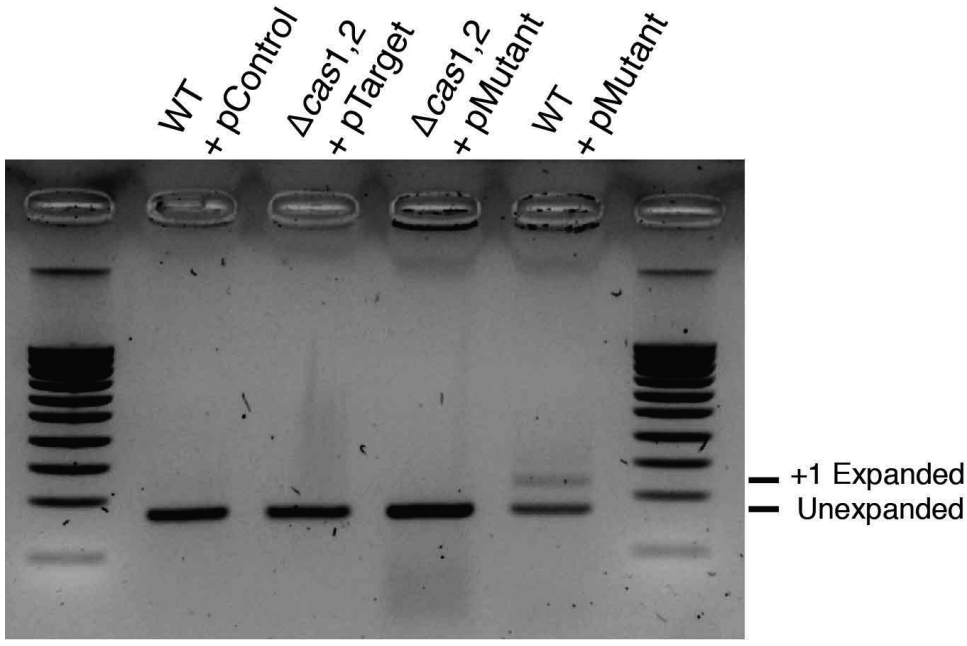
Supplementary Figure 3.2: Decay of YFP fluorescence in both direct interference and priming follows exponential decay The fluorescence concentration (open circles) of (a) direct interference and (b) priming lineages can be described by exponential decay. This was evaluated by performing a least-squares (LS) fit of the fluorescence concentration data (purple open circles) after the moment of plasmid loss (PLT, black circle) to an exponential curve (LS fit, black line).



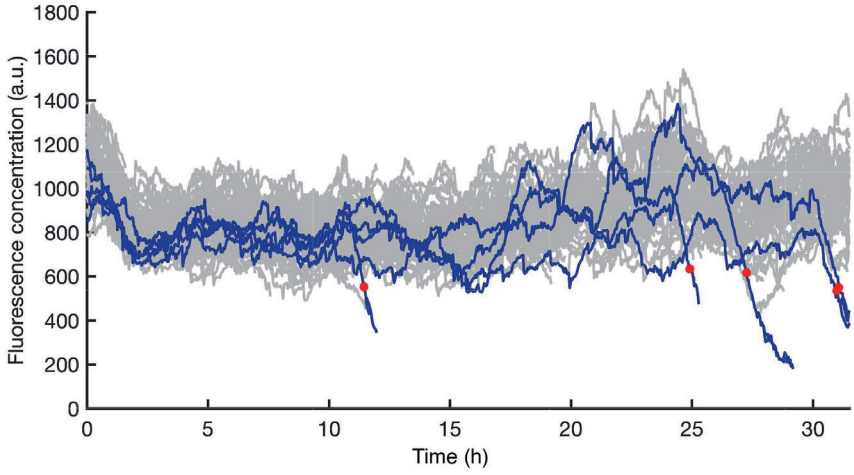
Supplementary Figure 3.3: Cell divisions before plasmid loss highly correlates with plasmid loss time The number of cell divisions from the moment of induction until plasmid loss are plotted against the PLT in hours. Both consensus target clearance by direct interference (green) and mutant target clearance by priming (blue) are shown.



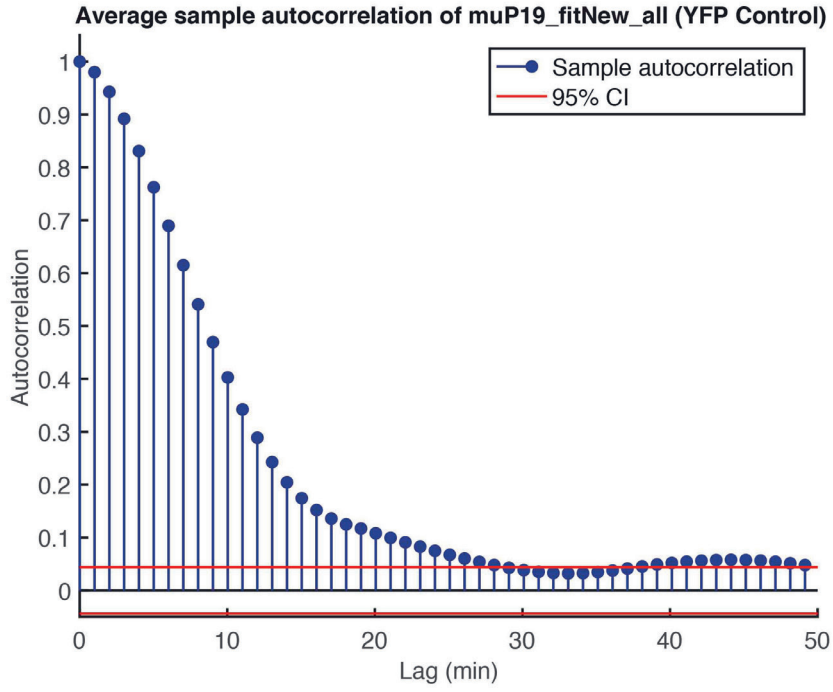
Supplementary Figure 3.4: Plasmid loss during Direct interference and primed interference processes occur on a comparable timescale All production rate traces (grey) starting from 140 minutes prior to the detected plasmid loss time PLT from (a) direct interference and (b) priming were aligned at the PLT (t-PLT=0) and the average trend (navy) normalized for comparison. From the average trend, we estimate the clearance time (CT), time taken from the initiation of plasmid clearance until elimination of all copies, to be in the order of 60 minutes for both direct interference and priming from the onset of the production rate decrease.



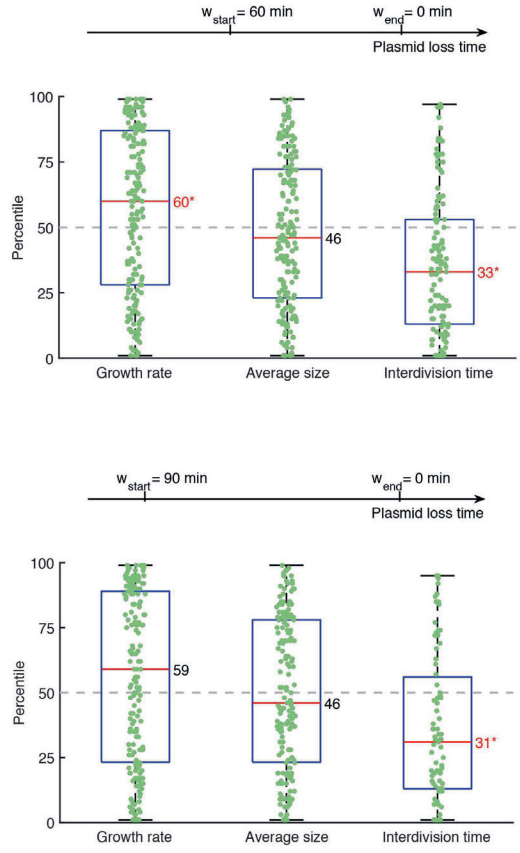
Supplementary Figure 3.5: Spacer acquisition was only seen in the WT strain in the presence of a mutated PAM triggering priming Cells from the chip output were collected in a flask for each experiment and the CRISPR arrays were screened for expansion due to spacer acquisition by PCR amplification using primers BN1530 + BN1531 (Supplementary Table 3.1). The gel shows PCR amplified CRISPR arrays from each experiment a, WT + pControl (Control) b, $\Delta cas1,2$ + pTarget (direct interference) c, $\Delta cas1,2$ + pMutant d, WT + pMutant (priming). The presence of a larger band indicates array expansion and therefore successful spacer acquisition.



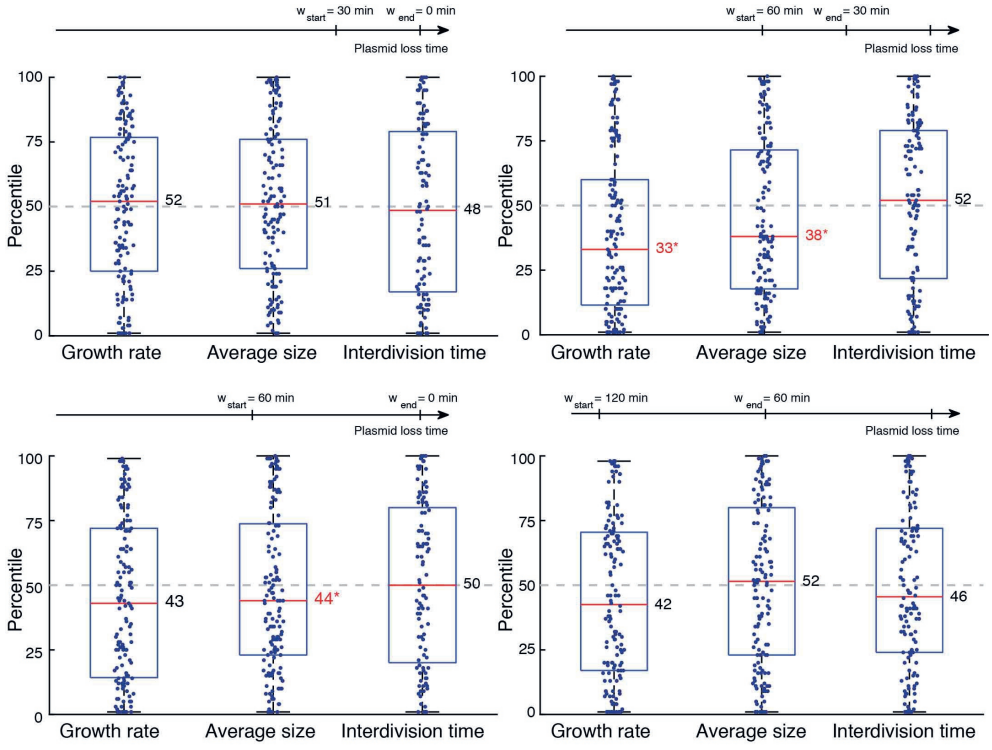
Supplementary Figure 3.6: In the absence of Cas1 and Cas2 clearance of a target with a non-consensus PAM mutant occurs rarely The YFP fluorescence of the $\Delta cas1,2$ strain containing pMutant was imaged for 34 hours after induction. Lineages that were able to clear the plasmid are highlighted in blue with the red dot indicating the moment of detection. 1.4% of lineages (5 unique events, red dot) cleared the plasmid.



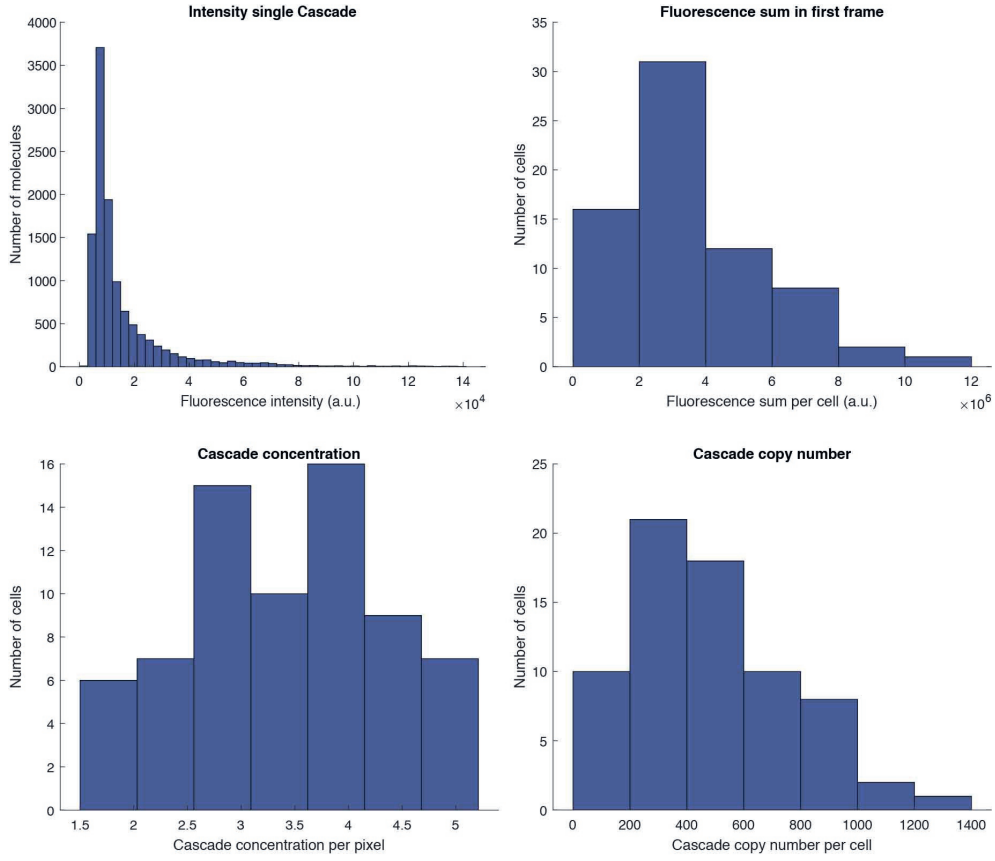
Supplementary Figure 3.7: Autocorrelation time of cellular growth rate The autocorrelation time was calculated for the cellular growth rate of the WT strain containing pControl by averaging the autocorrelation of cell growth as a function of time in individual lineages. After 10 minutes the autocorrelation of cellular growth has decreased to 0.4. After approximately 30 minutes the autocorrelation has decreased to zero, as indicated by 95% confidence intervals (red lines).



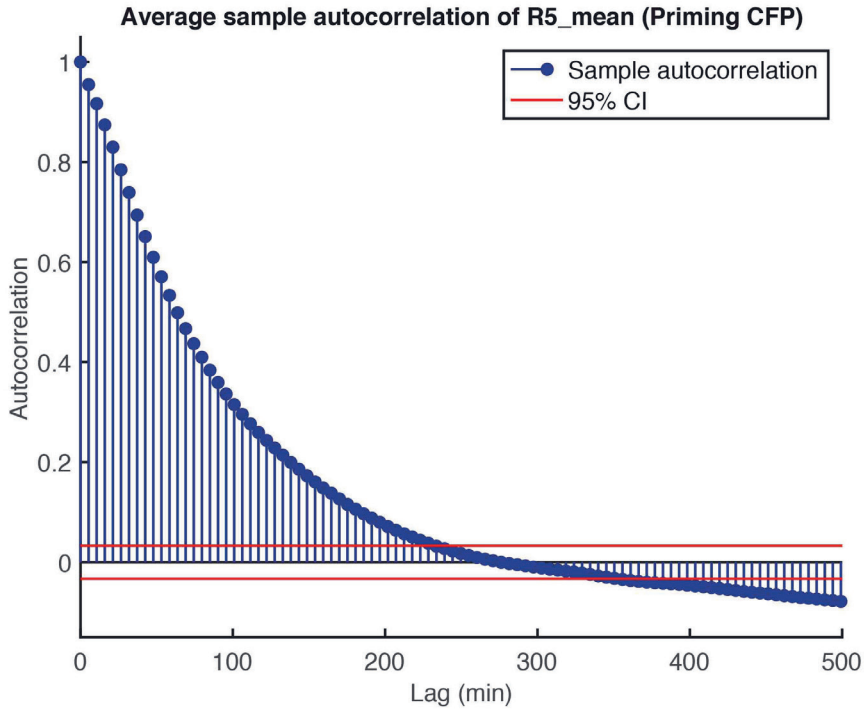
Supplementary Figure 3.8: Growth rate, cell size and interdivision time of direct interference with different lookback windows Boxplots of growth rate, average cell size and interdivision time presented as the percentile rankings of all plasmid loss lineages (green) that cleared a known target via direct interference. The cell feature of interest (e.g. growth rate) was averaged over a lookback window chosen in relation to the time from plasmid loss of the lineage of interest. The same cell feature was then averaged for all non-loss lineages in the population at that same moment. The cell feature of interest was then ranked amongst the non-loss population as a percentile. We considered lookback windows of 60 minutes prior to plasmid loss (top) and 90 minutes prior to plasmid loss (bottom). The median percentile ranking of loss lineages is indicated by a red line and black text, categories in which this value was significantly different from a ranking in the 50th percentile (p-value<0.05) are indicated in red text followed by an asterisk.



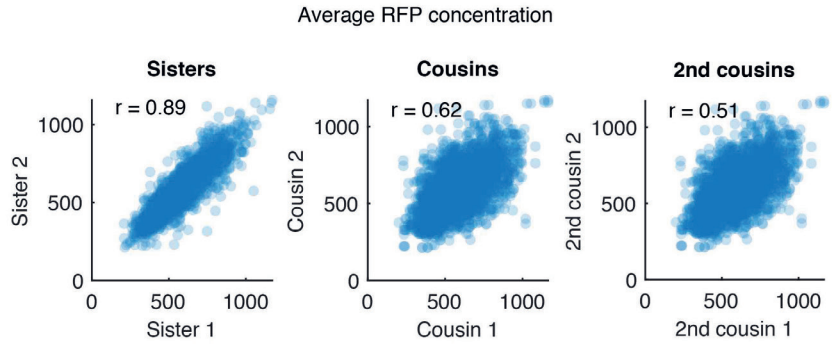
Supplementary Figure 3.9: Growth rate, cell size and interdivision time of priming with different lookback windows Boxplots of growth rate, average cell size and interdivision time presented as the percentile rankings of all plasmid loss lineages (navy) that cleared a known target via priming. The cell feature of interest (e.g. growth rate) was averaged over a lookback window chosen in relation to the time from plasmid loss of the lineage of interest. The same cell feature was then averaged for all non-loss lineages in the population at that same moment. The cell feature of interest was then ranked amongst the non-loss population as a percentile. We considered a range of lookback windows. The median percentile ranking of loss lineages is indicated by a red line and black text, categories in which this value was significantly different from a ranking in the 50th percentile (p-value <0.05) are indicated in red text followed by an asterisk.



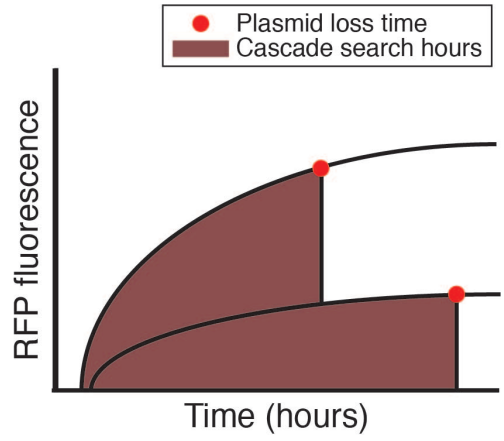
Supplementary Figure 3.10: Cascade copy number determination a, The fluorescence sum (RFP) of each cell in the first frame was determined. b, The RFP molecules were then bleached until it was possible to determine the fluorescence intensity of a single molecule (representing a single Cascade). c, The Cascade copy number per cell was then determined by dividing the average fluorescence sum by the average intensity of a single Cascade molecule. d, The Cascade concentration per pixel was determined by dividing the fluorescence sum by the area of the cell in pixels.



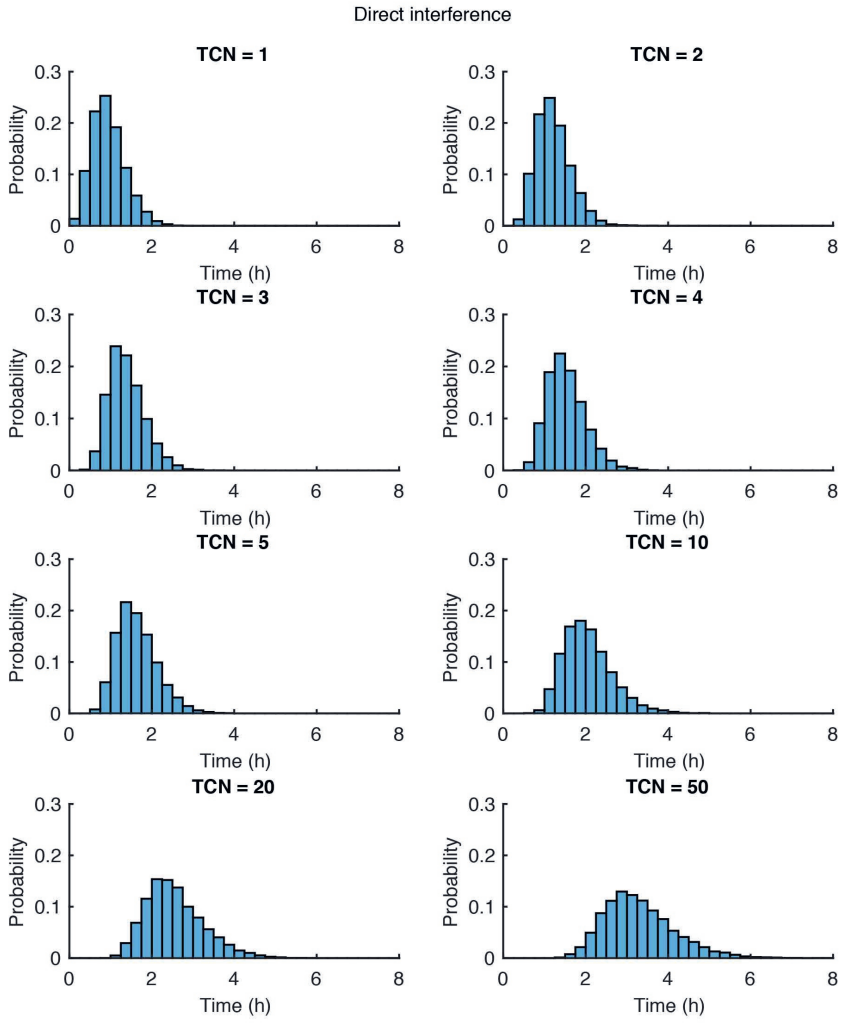
Supplementary Figure 3.11: Autocorrelation of RFP (Cascade) concentration The autocorrelation was calculated by averaging over the autocorrelation of RFP concentration of the WT-mCherry strain in individual lineages. After approximately 200 minutes the autocorrelation has decreased to zero, as indicated by 95% confidence intervals (red lines). The long decay time of the autocorrelation function indicates that Cascade protein levels fluctuate on a time scale longer than the cell cycle.



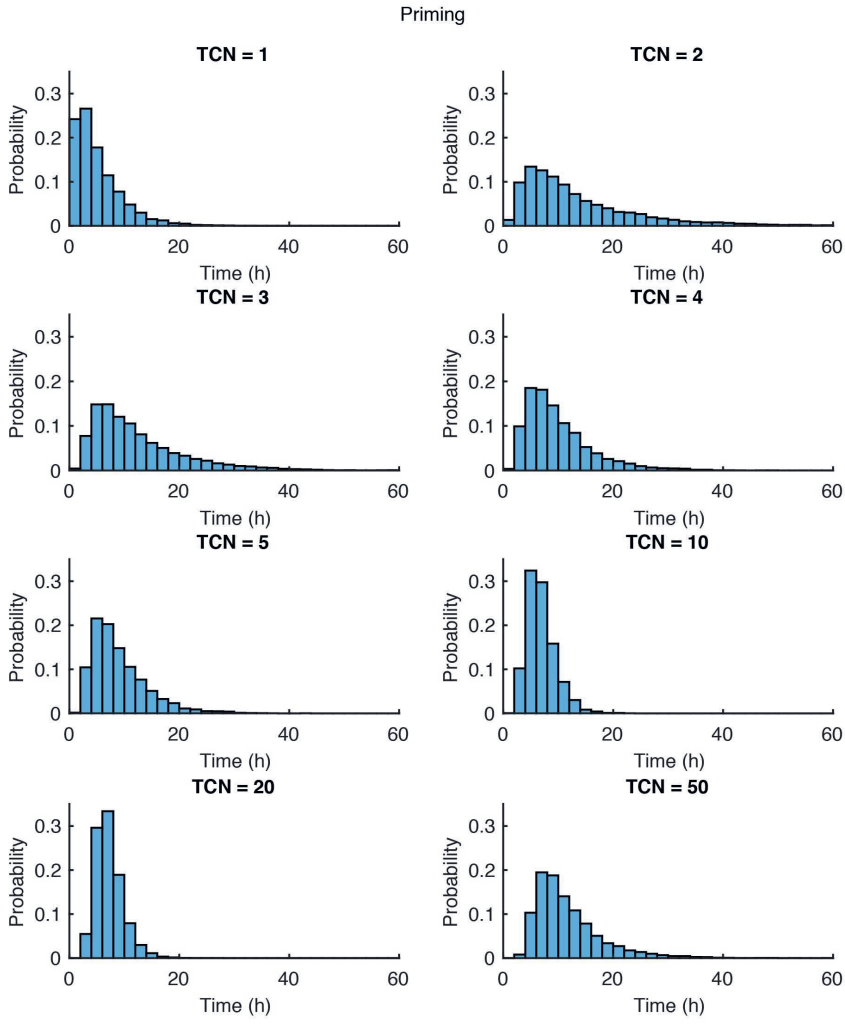
Supplementary Figure 3.12: Correlation of RFP levels between cells related as sisters, cousins, and second cousins The levels of RFP (Cascade) are strongly correlated between sister, cousins, and second cousins. The correlation coefficient r decreases as the cells become less closely related.



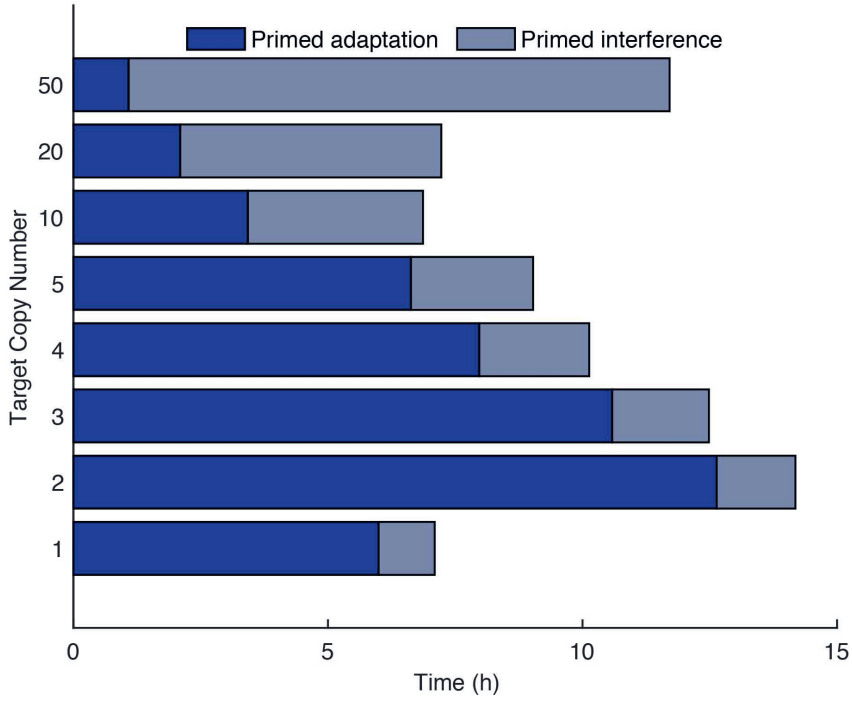
Supplementary Figure 3.13: Cascade search hours are calculated from the cumulative RFP Cascade search hours are the sum of all hours all Cascades have been searching in the cell. This can be calculated from the cumulative RFP or the area under the RFP curve as shown here in maroon. Two lineages with different RFP expression profiles are shown (black curves). The cell which has a higher RFP fluorescence and therefore copy number of Cascade loses the plasmid earlier as indicated by the red dot, while the cell with a lower copy number of Cascade loses the plasmid later. The two cells however, both lose the plasmid after approximately the same number of Cascade search hours i.e. the same area under the RFP curve.



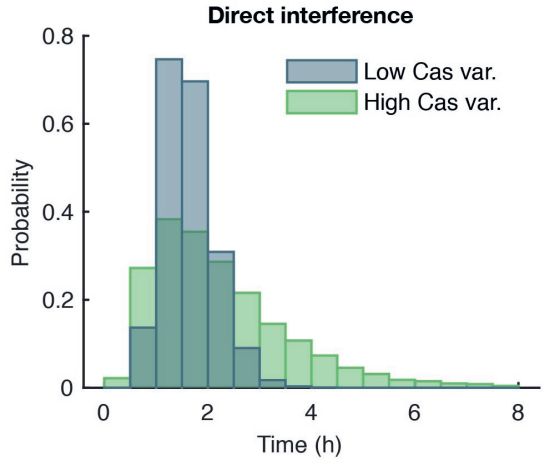
Supplementary Figure 3.14: Distribution of target loss times resulting from simulations of the direct interference condition for average target copy numbers (TCN) per cell ranging from 1-50



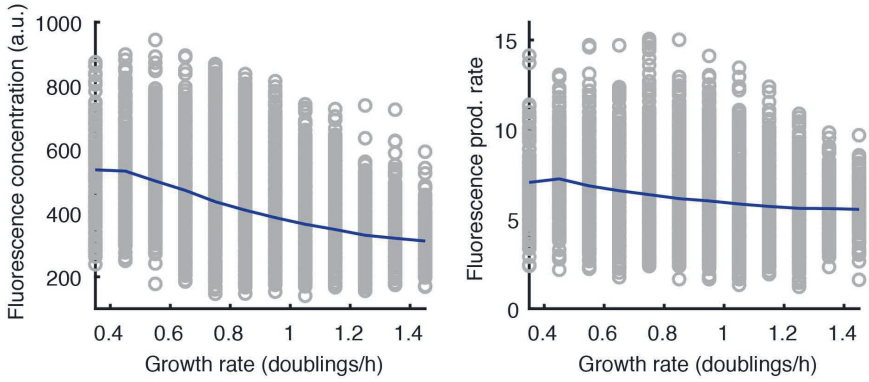
Supplementary Figure 3.15: Distribution of target loss times resulting from simulations of the priming condition for average target copy numbers (TCN) per cell ranging from 1-50



Supplementary Figure 3.16: Target loss time as a function of the target copy number (TCN) as computed from simulated trajectories by the agent-based model for the priming condition Bar charts representing the time spent on primed adaptation (navy) and primed interference (grey) for cells clearing targets through priming with an average plasmid copy number ranging from 1-50.



Supplementary Figure 3.17: Distribution of plasmid loss times in direct interference for high and low variability in Cascade concentration Target loss time distribution for two different levels of Cascade concentration variability resulting from simulated trajectories of the direct interference condition. At low variability (blue) Cascade proteins are produced in frequent, small bursts, whereas at high variability (green) proteins are synthesised more sporadically in large bursts (100-fold increase), keeping average Cascade concentration constant. The variability of PLT interference times for high Cascade variability increases as compared to low Cascade variability.



Supplementary Figure 3.18: Slower growing cells have higher RFP (Cascade) concentrations Cascade concentration (left) and Cascade production rate (right) show an inverse relationship with cellular growth rate (grey circles), revealing slower growing cell on average (navy line) have a higher concentration of Cascade.

Chapter 4

CRISPR-Cas interference and adaptation are PAM-dependent

Emma M. Keizer, Rebecca E. McKenzie, Christian Fleck, Jaap Molenaar, Sander J. Tans, Stan J.J. Brouns

Abstract

Bacteria can clear invaders such as phages and viruses from their system through an RNA-guided defence mechanism called CRISPR-Cas. However, invaders can escape CRISPR immunity by developing spontaneous mutations within the seed region or PAM region of the target site. It has been shown that mutations in the PAM sequence affect target recognition by the surveillance complex Cascade, which has implications for effective clearance of both known and unknown invaders. However, the exact mechanism by which Cascade initiates primed adaptation, the acquisition of a new immunological memory, is largely unknown. Here, we characterise the dynamics of CRISPR-mediated target clearance for three different PAM variants at the single-cell level, and compare these to simulated trajectories from two mechanisms of primed adaptation proposed in the literature. We show that features of the data are consistent with the interference-independent model for adaptation, and consider which factors might explain the observed cell-to-cell variability. Our results show that the CRISPR-response depends strongly on the PAM variant, which has important implications for the bacterial response to invading DNA elements.

4.1 Introduction

In the previous chapter, we discussed the cell-to-cell variability present in the response of CRISPR-Cas systems to invading nucleic acids. While clearance of invading DNA is dependent on various stochastic processes and competition between replication of foreign invaders and degradation by CRISPR-Cas systems, destruction of a previously encountered target is efficient. However, significant variation exists in the clearance dynamics of targets carrying mutations in the PAM region, which allows them to escape direct interference. We showed that spacer acquisition is required for timely plasmid loss, as clearance of the invader is rare in the absence of the Cas1-Cas2 protein complex responsible for the integration of new spacers into the CRISPR array. For type I-E systems a consensus protospacer adjacent motif (PAM) flanking the targeted site of the invader allows swift recognition and ultimately degradation of the invader, through a process called direct interference.[114,124–126] We have previously shown that for the 5′-CTT consensus PAM, spacer acquisition is not required for successful plasmid clearance, as all cells in the population manage to clear all plasmids within hours in the absence of Cas1-Cas2. However, invaders can escape direct interference by developing spontaneous mutations within the seed region of the target site or PAM.[122,127,128] In response, the I-E system can initiate priming, which promotes accelerated acquisition of new spacers due to a pre-existing partial match to the invader.[108,111] Primed adaptation is much faster than naive adaptation,[129] which occurs during infection by an invader that has not been previously encountered, and is required for the insertion of a new matching spacer with a consensus PAM allowing subsequent invader degradation. We were able to define the adaptation (primed adaptation) and clearance (primed interference) stages of priming and identified primed adaptation as the source of the variation in plasmid loss time observed, which is characterised by a low affinity of the surveillance complex Cascade to the target plasmid owing to the PAM mutation. Although the heterogeneity in plasmid loss times between cells is affected by variations in cellular growth rate and Cascade expression, the interaction between Cascade and target DNA represents the key source of heterogeneity.

The molecular mechanism by which Cascade initiates priming remains elusive.[114] The number of cells in the population that obtain new spacers as a result of primed adaptation has been shown to be very low in the case of a fully matching target spacer and consensus PAM.[108,130,138] Mutations in the PAM or target spacer sequence decrease the efficiency of interference by the nuclease Cas3, but have been shown to stimulate primed adaptation.[127,165] Yet, a functional Cas3 protein is required for primed adaptation, suggesting the mechanisms of interference and primed adaptation are connected.[109,116,128] A systematic screening of all 64 possible trinucleotide PAM sequences has shown that while there appears to be a trade-off between interference and priming efficiency, some PAM sequences are proficient at both.[112] So far this has only been studied in bulk assays, masking

information on the timing and variability of invader clearance within the population. Previously we have explored the single-cell dynamics of the CRISPR-Cas response to targets harbouring a consensus PAM (5′-CTT), as well as a priming-proficient non-consensus PAM (5′-CGT). Here, we will now extend our analysis to the 5′-AAT PAM, a sequence that is known to promote both interference and priming.[112] We will track the dynamics of cell populations harbouring target plasmids with these 3 different PAMs, and monitor their ability to promote interference and primed adaptation under different conditions.

Currently, two competing mechanistic models exist in the literature describing how Cascade, Cas1-Cas2, and Cas3 interact to generate new spacers during primed adaptation. The first model hypothesises that priming is the consequence of interference, as Cas3 recruited to DNA-bound Cascade produces short DNA fragments in the cell which can captured by Cas1-Cas2 for integration into the CRISPR array.[109,130,137,186] The alternative model proposes that Cascade, Cas3, and Cas1-Cas2 assemble into a complex, allowing DNA fragments that are cleaved by Cas3 to be directly taken up by Cas1-Cas2.[116,192,193] By adapting the agent-based simulation framework which was developed in the previous chapter, we aim to uncover the molecular mechanism behind primed adaptation. To this end, we simulate the plasmid loss dynamics of a cell population using the two different models for spacer generation during primed adaptation, and compare features of the simulated trajectories to time-lapse data on target clearance in individual lineages for the three PAM variants.

4.2 Results

Using six strains, we monitor the processes of interference and priming for 3 PAM variants. For each PAM variant we test a wild-type (WT) and $\Delta cas1,2$ strain. The interference process is studied by using $\Delta cas1,2$ strains, lacking the *cas1* and *cas2* genes coding for the adaptation proteins. Wild-type (WT) strains encode for the full range of CRISPR-associated (Cas) proteins, including Cas1 and Cas2, enabling the acquisition of new spacers. As in the previous chapter, the strains contain a CRISPR array with a leader, two repeats and a single previously characterised spacer, spacer8 (SP8).[111,112] In addition, these strains are engineered to control *cas* gene expression using arabinose and IPTG induction, and hence initiation of the CRISPR-Cas response. Target plasmids were engineered to encode a constitutively expressed YFP fluorescent protein and contain a target sequence that is complementary to SP8 in the CRISPR array. This allows us to monitor target DNA presence in individual cells over time. The direct interference (DI) process was monitored by flanking the target sequence with a 5′-CTT consensus PAM. Further, to investigate the priming (P) response we mutated the PAM to 5′-CGT, a mutation known to allow mobile genetic elements (MGE) to escape interference, and invoke a primed

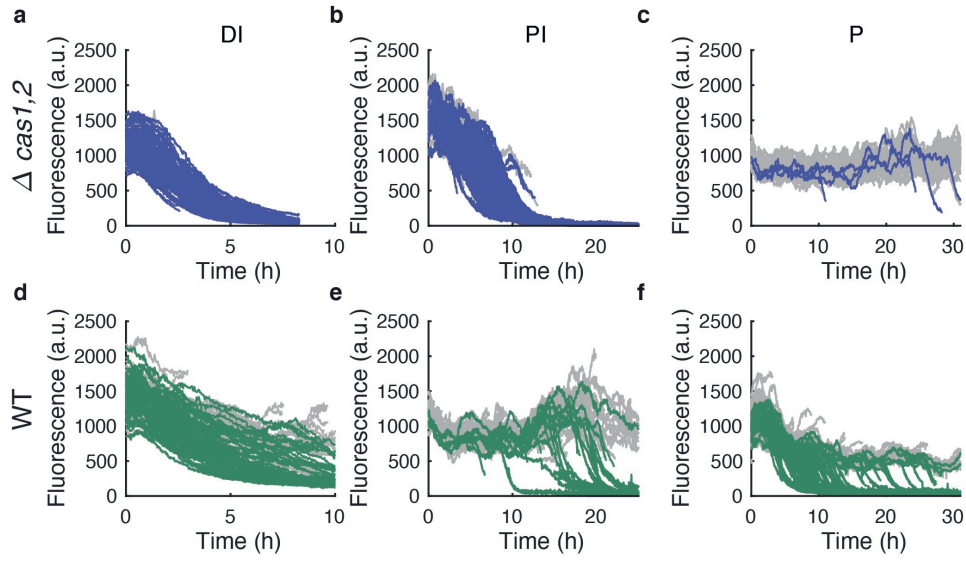


Figure 4.1: Reconstructed lineage traces of the imaged population from induction of the CRISPR-Cas system over time (grey) for 3 PAM variants. **a,d,** direct interference (DI), **b,e,** priming interference (PI), and **c,f,** priming (P). Coloured traces show lineages that successfully clear all target plasmids in the absence ($\Delta cas1,2$, navy lineages) and presence (WT, green lineages) of Cas1-Cas2.

adaptation response. In addition, we flank the target sequence with a 5′-AAT PAM, which has previously been shown to be capable of both priming and interference (PI).[112] In the remainder of the text, we refer to these respective PAMs as direct interference (DI), priming (P), and priming interference (PI).

As described in Chapter 3, we use a microfluidic device enabling fluorescence time-lapse imaging for over 36 hours with the option for media exchange. The device contained chambers allowing observation of a single layer of cells, constant medium supply, removal of cells that no longer fit the chamber due to growth, and control of intracellular processes via induction. The microscope set-up combines phase contrast imaging at 1 min intervals and fluorescence imaging at 2 min intervals. Phase contrast images are segmented and tracked using custom MATLAB software, based on the Schnitzcells software[150] which allows for the reconstruction of individual cells into lineage trees. The moment all plasmids are cleared, which we term the plasmid loss time (PLT, Fig. 3.2c), is quantified using the YFP production rate as defined by Levine *et al.*[152] For a more detailed description of these methods we refer to Section 3.4.9.

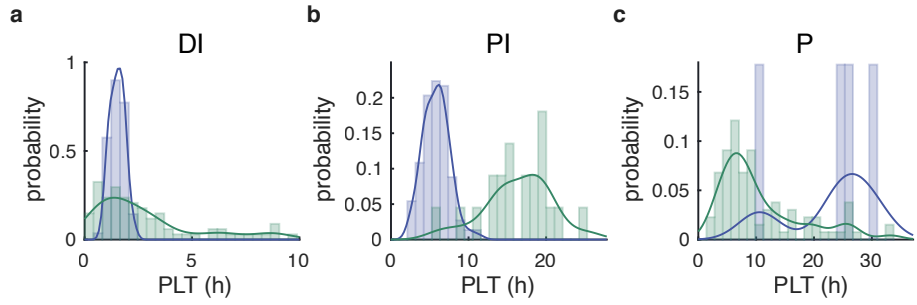


Figure 4.2: Histogram of plasmid loss times (PLT) in the absence ($\Delta cas1,2$, navy) or presence (WT, green) of Cas1-Cas2 for **a,** direct interference (DI; $\Delta cas1,2$: $n = 82$, WT: $n = 135$), **b,** priming interference (PI; $\Delta cas1,2$: $n = 140$, WT: $n = 21$), and **c,** priming (P; $\Delta cas1,2$: $n = 4$, WT: $n = 94$) PAM variants. Solid lines show kernel density estimation (KDE) of underlying histogram data.

4.2.1 The rate of Cas3-mediated target degradation is PAM-dependent

Fig. 4.1 shows the target fluorescence concentration over time for cell populations with plasmids where the target sequence is flanked by the direct interference (Fig. 4.1a,d), priming interference (Fig. 4.1b,e), and priming (Fig. 4.1c,f) PAM sequence. Lineages with plasmid loss, as detected by the plasmid-encoded YFP production rate, are highlighted (navy and green lines for $\Delta cas1,2$ and WT respectively), whereas lineages in which plasmids were not successfully cleared before exiting the well or before the end of the experiment are shown in grey.

As reported in Chapter 3, for DI the target was cleared in all cells (Fig. 4.1a). The PLT has a narrow distribution, with plasmid loss taking place between 1-2.5 hours after induction (Fig. 4.2a). The average PLT was 1.5 hours ($CV^2 = 0.055$). Plasmid loss sometimes occurred in the same generation in which the CRISPR-Cas response was initiated by induction, and clearance took a maximum of 3 generations (Supplementary Fig. 3.3). For priming interference, the plasmid was also cleared in the whole population (Fig. 4.1b) with plasmid loss taking between 2.5 and 11.5 hours with an average PLT of 5.8 hours (Fig. 4.2b, $CV^2 = 0.083$), or between 2 and 7 generations from the moment of induction (data not shown). In the absence of Cas1-Cas2, the priming PAM showed very sporadic plasmid loss (Fig. 4.1c), with only 4 out of the 60 cells present at the start of the experiment successfully clearing the plasmids after being monitored more than 30 hours since induction (Fig. 4.2c). The fastest plasmid loss event took place 10.5 hours after induction of *cas* gene expression. This indicates that although for the priming PAM target clearance through interference can take place, the rate of interference can rarely be

overcome by the rate of plasmid replication.

4.2.2 Cas1-Cas2 can stimulate or attenuate plasmid clearance

Next, we monitor plasmid loss of targets with the three PAM variants flanking the target spacer sequence in cells expressing the full range of *cas* genes, including *cas1* and *cas2*. As previously reported, cells containing escape mutant targets are able to acquire new spacers and clear the plasmid through primed interference. Although there is much variation in the timing of this process, the probability of plasmid loss for the WT strain with priming PAM is highly increased, with plasmid loss being observed as early as 1.7 hours after induction (Fig. 4.2c). Plasmid loss was highly variable, and loss events could be observed throughout the experiment with a maximum recorded loss time of 33.8 hours (Fig. 4.1f), or 31 generations (Supplementary Fig. 3.3). The average time of all recorded plasmid loss events was 9.6 hours ($CV^2 = 0.46$). After monitoring the cells for 36 hours, some cells in the population still had high fluorescence production rates indicating the presence of plasmids. Thus, plasmid clearance was not successful for the whole population (Fig. 4.1f).

Surprisingly, in the case of the direct interference and priming interference PAMs, we see that plasmid loss speed is reduced in WT cells as compared to the $\Delta cas1,2$ strains (Fig. 4.1a,b compared to Fig. 4.1d,e). For the plasmids with the DI PAM, WT cells showed plasmid loss minutes after induction, but with a maximum time of 9.9 hours (Fig. 4.2a). Although all cells in the population cleared the plasmid within 10 hours or 3 generations, the average plasmid loss time was increased to 2.8 hours as compared to 1.5 hours in the absence of Cas1-Cas2 (Fig. 4.2a). Slower plasmid loss was also seen in the WT strain when the PI PAM was introduced on the target, with plasmid loss taking between 6.3 - 25.3 hours (6-28 generations) with a mean PLT of 16.4 hours as compared to 5.8 hours in the absence of Cas1-Cas2 (Fig. 4.2b). In addition, the frequency of plasmid loss was decreased, with a number of cells still harbouring plasmids after 30 hours of monitoring (Fig 4.1e). Hence, we conclude that for the PAM variants monitored here, both the interference and primed adaptation process are affected. Unsurprisingly, the presence of Cas1-Cas2 when priming is required for timely plasmid loss, and results in a higher probability of clearance. However, surprisingly the presence of Cas1-Cas2 decreases the probability of plasmid clearance in all cells where priming is not required to clear the plasmid from the cell, as observed for targets containing the DI and PI PAM.

4.2.3 An agent-based model to study the molecular basis of primed adaptation

In the previous chapter, we have set up an agent-based stochastic simulation framework with which we could simulate the dynamics of a cell population for

the direct interference ($\Delta cas1,2$) and priming ($\Delta cas1,2$ and WT) conditions (Appendix 3.B). Hitherto we have made no assumptions on the interactions between Cascade, Cas1-Cas2, and Cas3. However, since the data presented here suggest that Cas1-Cas2 affects the speed of interference we adapt the previous model to include molecular interactions involving Cas1-Cas2 and Cas3. While it has been shown that the *cas1* and *cas2* genes are essential and specific to the adaptation phase and both are relatively well conserved in almost all CRISPR-Cas systems,[114] the molecular details of the adaptation process are not fully known. In the literature, there are two proposed pathways for primed adaptation. In the first, referred to as the interference-dependent (ID) pathway (Fig. 4.3a), target degradation by Cas3 supplies substrates for primed adaptation.[109,130,137,186] Following target binding by the surveillance complex, Cas3 is recruited and degrades the target DNA. The interference products form a pool of potential spacer substrates that can be captured by Cas1-Cas2. In the second proposed mechanism, the interference-independent (II) pathway (Fig. 4.3b), Cas3 and Cas1-Cas2 are recruited by Cascade into a priming complex following target recognition.[116,192,193] This complex translocates along the DNA and directly excises double-stranded pre-spacers from the target DNA. For both pathways, primed adaptation is triggered only upon PAM-dependent target recognition by the Cascade surveillance complex, ensuring that pre-spacers are derived only from the invader. Using the agent-based simulation framework, we set out to generate data for both models and assess how they agree with our experimental data. As the available data is not sufficient to accurately estimate all model parameters, which increase in number with the model complexity, we aim that the model qualitatively replicates essential features of the experimental data rather than quantitatively matches the plasmid loss time distributions.

Interference-dependent spacer acquisition

We set out with the simplest model, in which interference products form a pool of spacer candidates. As the model described in Chapter 3 does not include any direct interaction between Cascade and Cas1-Cas2, this can be viewed as an implementation of the interference-dependent pathway. In order to replicate all six experimental conditions, we adapt the model by explicitly including synthesis of Cas1-Cas2 and Cas3, the production of which, like Cascade, can be switched on at the start of the experiment. Previously, we have quantified the copy number of Cascade in our cells.[194] As the genes encoding Cascade, Cas1, and Cas2 are under the control of the same promoter, we will assume equal abundances of Cascade and Cas1-Cas2. We will also assume the same abundance for Cas3, although its copy number was not directly measured. At this copy number Cas3 is not rate-limiting, as increasing the concentration by a factor 10 did not affect the simulation outcomes. We simulate a population of 100 exponentially growing and dividing cells, their plasmid maintenance, stochastic protein production and partitioning at division.

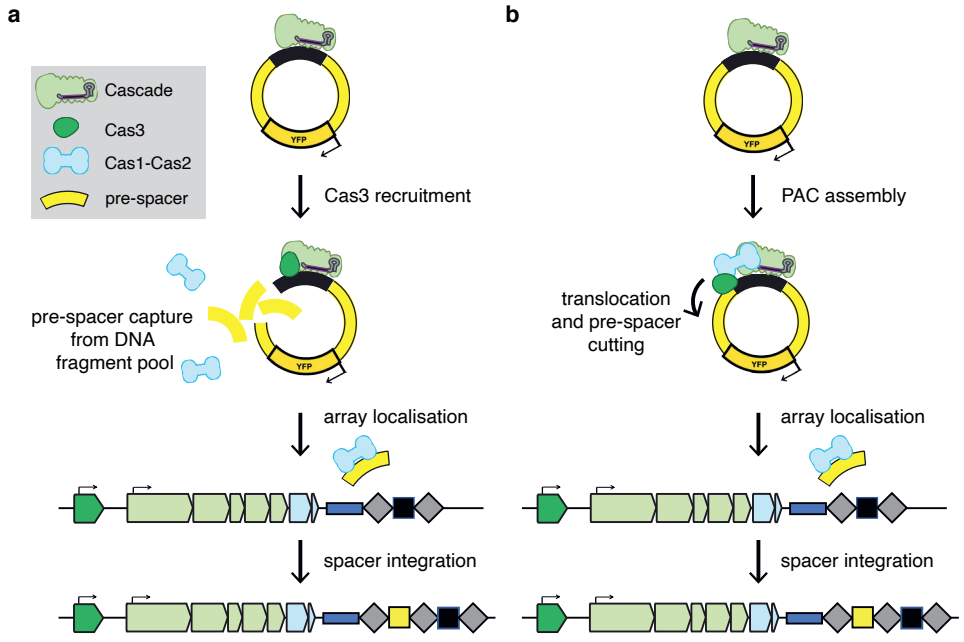


Figure 4.3: Two alternative molecular mechanisms underlying primed adaptation: **a,** Interference-dependent (ID) pathway, **b,** Interference-independent (II) pathway with the primed acquisition complex (PAC).

Protein expression, spacer acquisition and target DNA degradation are governed by the equations described in Supplementary Table 4.1. Note that the molecular mechanism is the same as described in Chapter 3, with the exception of molecular interactions involving Cas3 and Cas1-Cas2 which are explicitly incorporated.

In the previous chapter, it was postulated that mutations in the PAM region of the target affect the ability of Cascade to form a long-lasting bond with the target. We estimated that the Cascade-target binding affinity was reduced by two orders of magnitude in the priming case as compared to the consensus PAM and were able to replicate corresponding plasmid loss distributions for both conditions (Fig. 4.4a and Fig. 4.4f respectively). For the target with the PI PAM, we estimate the Cascade-target binding affinity by manually fitting the PLT distribution from simulated trajectories in the absence of Cas1-Cas2 (Fig. 4.4b), resulting in a binding affinity that is 5 times lower than is the case for the consensus PAM. For targets containing the P PAM, plasmid loss is very sporadic in simulated trajectories in the absence of Cas1-Cas2 with on average 4 events among 100 cells over 40 hours (Fig. 4.4c), which is consistent with the experimental data.

We then switch on Cas1-Cas2 synthesis at the start of the simulation to see the effect of adaptation on the plasmid loss time according to the ID model. For direct

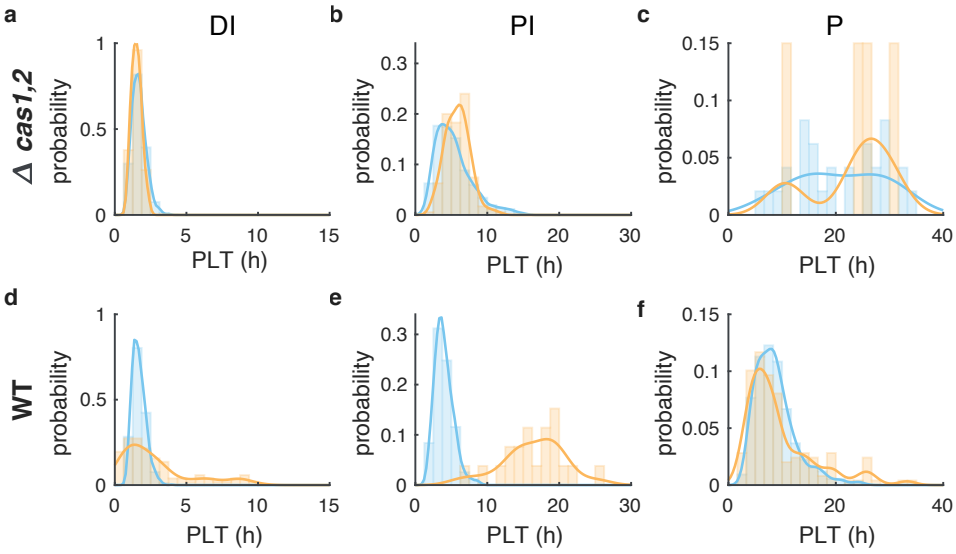


Figure 4.4: Experimental plasmid loss time distributions (orange) versus simulated plasmid loss time data (blue) with the interference-dependent (ID) model for 3 PAM variants: **a,d,** direct interference (DI), **b,e,** priming interference (PI), and **c,f,** priming (P) in the absence ($\Delta cas1,2$, top row) and presence (WT, bottom row) of Cas1-Cas2. Solid lines show kernel density estimation (KDE) of underlying histogram data.

interference, there is no difference in the plasmid loss distribution in the presence of Cas1-Cas2 as compared to the $\Delta cas1,2$ scenario as both distributions have a mean of 1.7 hours (Fig. 4.4d). This indicates that all targets have been destroyed before spacer acquisition can take place. For priming interference, the effect is a reduction in the plasmid loss time: 4 hours on average (Fig. 4.4e) versus 5.7 hours without Cas1-Cas2. In 18% of all plasmid loss events, cells are able to clear all plasmids through interference alone, while the majority (82%) of cells reach plasmid clearance through the acquisition of a new spacer. It is unknown, however, if these percentages are representative as the experimental data does not provide information on the dominant plasmid clearance pathway. In the population of cells containing targets with the priming PAM, this results in plasmid loss being greatly increased as compared to the $\Delta cas1,2$ case, with a PLT distribution resembling the experimental data and a mean plasmid loss time of 9.7 hours (Fig. 4.4f).

From this we conclude that while the ID model is able to qualitatively describe a subset of the experimental conditions (Fig. 4.4 a,b,c,f), it cannot reproduce the observed effect of Cas1-Cas2 in the direct interference and priming interference conditions. According to the interference-dependent primed adaptation mechanism, Cas1-Cas2 has no molecular involvement in the plasmid destruction process. For this reason, the addition of Cas1-Cas2 does not slow down the rate of target

degradation and we are not able to replicate the experimental observations of slower loss for the DI and PI PAM sequences in the presence of Cas1-Cas2. While we do not rule out the possibility that new spacers can be acquired from the pool of interference products, the ID mechanism alone is not able to explain our experimental observations. A model for primed adaptation which describes an additional mechanism in which Cas1-Cas2 affects the speed of interference might better replicate our data.

Interference-independent spacer acquisition

In the interference-independent pathway, it is hypothesised that the binding of Cascade to the target triggers the recruitment of both Cas1-Cas2 and Cas3 which leads to the assembly of a primed acquisition complex (PAC) as shown in Fig 4.3b. The PAC then translocates along the target DNA in search of a spacer substrate, which can be excised and transported to the spacer array. We adapt our stochastic simulation model to include these molecular interactions. Notably, our experiments have shown that while rare, plasmid degradation is also possible in the absence of Cas1-Cas2 (Supplementary Fig. 3.6). This is reflected in the II model, as upon target binding the Cascade surveillance complex can either recruit Cas3 to directly degrade the plasmid, or recruit Cas1-Cas2 and Cas3 to form the PAC. The PAC can then excise a spacer from the plasmid, which is transported by Cas1-Cas2 to the spacer array.

Various factors might affect the balance between the Cas3-mediated degradation pathway and the PAC assembly pathway. In our model, these are limited to the following: the concentration of Cas3 and Cas1-Cas2, and the recruitment rates of either Cas3 (direct degradation) or Cas1-Cas2 (PAC assembly) by the target-bound effector complex. No quantitative data is available on *in vivo* kinetic recruitment rates of Cas proteins by target-bound Cascade, and we have thus assumed that the recruitment rates of Cas3 and Cas1-Cas2 are equal. The Cascade-target binding affinities for DI, PI, and P are identical to the rates used to simulate the ID model. While this results in identical PLT distributions in the $\Delta cas1,2$ case (Fig. 4.5a-c), in the presence of Cas1-Cas2 we now have two different adaptation pathways which results in a change in the plasmid loss rate when production of Cas1-Cas2 is switched on (Fig. 4.5d-f). When Cas1-Cas2 is expressed, overall we see a slowing down of plasmid loss rates for the DI and PI conditions. For direct interference, the mean of the PLT distribution is increased from 1.7 hours to 6.5 hours (Fig. 4.5d). This effect of Cas1-Cas2 is larger than what is observed in the experimental PLT distribution. For priming interference, the average PLT is increased from 5.8 hours to 9 hours Fig. 4.5e), whereas the experimental plasmid loss distribution is shifted towards even larger plasmid loss times. In 1% of the simulated plasmid loss events, clearance happens without the acquisition of a new spacer. Conversely, for the priming PAM the probability of plasmid loss is greatly increased compared to the

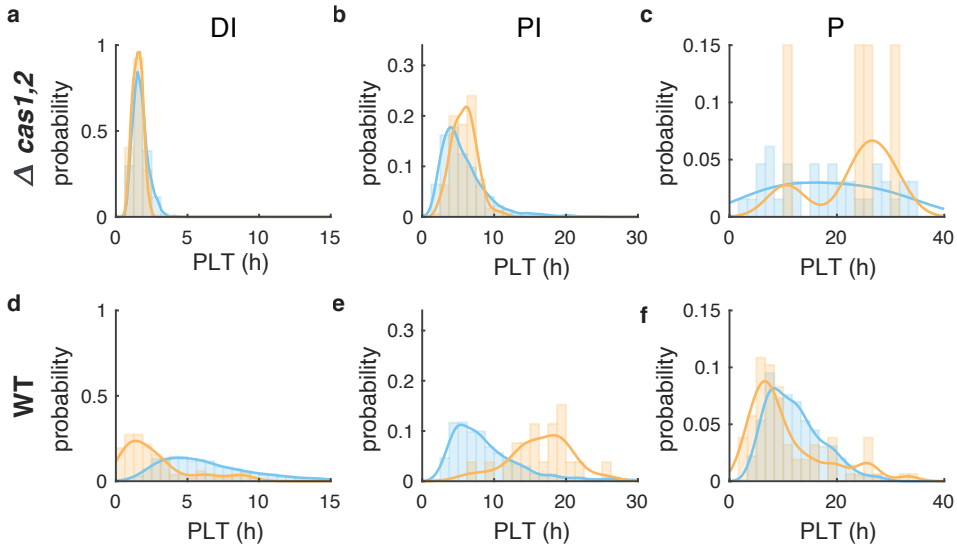


Figure 4.5: Experimental plasmid loss time distributions (orange) versus simulated plasmid loss time data (blue) with the interference-independent (II) model for 3 PAM variants: **a,d,** direct interference (DI), **b,e,** priming interference (PI), and **c,f,** priming (P) in the absence ($\Delta cas1,2$, top row) and presence (WT, bottom row) of Cas1-Cas2. Solid lines show kernel density estimation (KDE) of underlying histogram data.

$\Delta cas1,2$ condition, and the plasmid loss dynamics quantitatively agree with the experimentally observed loss events (Fig. 4.5f).

In contrast to the ID mechanism, the II pathway is able to mechanistically explain the slower clearance rate of plasmids in which the target sequence is flanked by the DI or PI PAM in the presence of Cas1-Cas2. Upon successful target identification, Cascade forming a bond with the target can initiate the formation of a PAC, in which Cas3 is less processive than in the Cascade-Cas3 complex, and thus leads to less efficient target destruction.[116] For the model parameters used in these simulations, in 50% of Cascade-target binding events result in the formation of the PAC as recruitment rates of Cas1-Cas2 and Cas3 are assumed to be equal. Further analysis is required to determine if different recruitment rates, or even PAM-dependent recruitment rates might result in better qualitative agreement with the experimental data.

[†]Statistics may not be representative due to small number of loss events.

		Experimental		ID model		II model	
		$\langle PLT \rangle$	CV^2_{PLT}	$\langle PLT \rangle$	CV^2_{PLT}	$\langle PLT \rangle$	CV^2_{PLT}
DI	$\Delta cas1,2$	1.5 h	0.049	1.7 h	0.083	1.7 h	0.092
	WT	2.8 h	0.77	1.7 h	0.080	6.5 h	0.30
PI	$\Delta cas1,2$	5.8 h	0.083	5.4 h	0.23	5.8 h	0.32
	WT	16 h	0.071	4.0 h	0.11	9.0 h	0.33
P	$\Delta cas1,2^{\dagger}$	22 h	0.14	21 h	0.16	18 h	0.3
	WT	11 h	0.47	8.7 h	0.19	12 h	0.18

Table 4.1: Statistics of simulated plasmid loss time distributions for the interference-dependent (ID) model and the interference-independent (II) model as compared to the experimental plasmid loss data for all 6 experimental conditions.

4.3 Discussion

In this study, we have looked at the processes of interference and primed adaptation at the single-cell level. These cells harbour plasmids engineered to contain various PAMs which flank the target sequence matching the CRISPR array, allowing direct monitoring of target DNA presence in individual cells over time. The three PAM variants used, here referred to as direct interference, priming interference, and priming, have been shown to differ in their ability to invoke an interference or priming response.[130] In the literature, two competing molecular mechanisms exist for primed adaptation. We have demonstrated that the ID mechanism, in which adaptation relies on DNA fragments produced by the interference machinery, was unable to reproduce the slower rate of plasmid loss observed in the presence of Cas1-Cas2 for the DI and PI PAM sequences (Fig. 4.4). According to this mechanism, Cas1-Cas2 does not physically associate with Cas3 or Cascade, and consequently addition of Cas1-Cas2 does not result in slower plasmid loss for DI and PI PAMs in WT cells as compared to $\Delta cas1,2$ cells. Instead, the observation of attenuated plasmid loss can be explained by the existence of a primed acquisition complex that includes Cascade, Cas3, and Cas1-Cas2, which generates pre-spacers by moving along target DNA.[193] We have shown that trajectories simulated according to this molecular mechanism qualitatively agree with the experimental data, though substantial discrepancies remain between experimental PLT distributions and PLT distributions simulated with this model (Fig. 4.5). We will now explore possible explanations for these discrepancies.

Although single-molecule studies have produced experimental data in support of the PAC, only the interaction between Cas1-Cas2 and Cascade was demonstrated *in vivo* for the type I-E CRISPR system of *Thermobifida fusca*.[116] Several details on the assembly of the PAC, such as the order in which proteins are recruited into the complex after target recognition by Cascade, remain elusive. It has been suggested

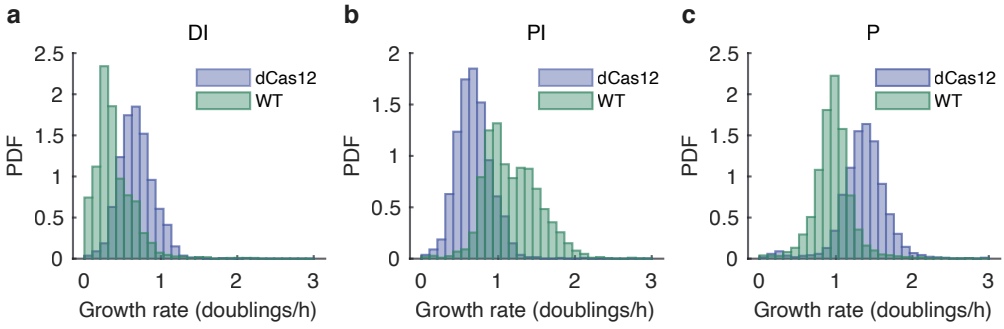


Figure 4.6: Distributions of growth rate (doublings/hour) of cells for 3 PAM variants: **a,** direct interference, **b,** priming interference, and **c,** priming in the absence (navy, $\Delta cas1,2$) and presence (green, WT) of Cas1-Cas2.

that there are two assembly routes that include initial recruitment of either Cas3 or Cas1-Cas2 to target-bound Cascade, followed by the addition of the remaining sub-complex.[116] This would result in different probabilities for interference and adaptation, as opposed to a scenario where a specific order is required for the PAC assembly. Alternatively, it has been proposed that Cas3 and Cas1-Cas2 might assemble into a pre-complex which is then recruited by Cascade.[192,195] A possible indication for this mechanism can be found in type I-F CRISPR systems, where Cas2 is traditionally fused to Cas3.[196] Moreover, this could provide an additional explanation for the decreased plasmid loss speed for DI and PI plasmids in the presence of Cas1-Cas2, as the assembly of a Cas1-Cas2-Cas3 complex would reduce the amount of free Cas3 available for interference.

In the previous chapter, we concluded that cellular growth rate has a small yet significant effect on the speed of the CRISPR-Cas response of cells. In our experimental set-up it is not possible to control the cellular growth rate, and there are differences across experiments in the distributions of growth rates of cells (Fig. 4.6). We now turn to the question of whether differences in the cellular growth rate, which affect the distribution of target copy number and Cas protein abundance, could explain the distinct plasmid loss time distributions of the WT and $\Delta cas1,2$ strains. We point out that the agent-based model only includes effects of growth rate on molecule concentrations due to dilution, as the relationship between growth rate, protein synthesis rate, and plasmid replication rate is not well characterised.

To recapitulate, in Chapter 3 we showed that lineages which successfully clear all plasmids ('loss lineages') through direct interference in the absence of the Cas1-Cas2 adaptation machinery exhibit a higher median growth rate relative to lineages that had not lost their plasmids at that moment ('non-loss lineages'). Conversely, priming (WT) was more successful in cells growing relatively slowly. This can be explained

by considering the effect of cellular growth on both the concentration of Cas proteins and the target plasmid copy number. Indeed, slower cell growth correlates with higher concentrations of Cascade (Fig. 4.7a, Supplementary Fig. 3.18), which is a determinant in the speed of plasmid loss (Fig. 3.4). In addition, in previous studies faster growing cells have been shown to have lower plasmid abundance[159,160] which correlates with faster interference, whereas in slow growing cells plasmid maintenance mechanisms increase plasmid abundance,[161] resulting in higher DNA availability as substrate for pre-spacers (Fig. 4.7a).

While for the three additional experimental data sets presented in this chapter (DI (WT), PI ($\Delta cas1,2$; WT)) no significant deviations in growth rate between loss-lineages and non-loss lineages were detected (Supplementary Figures 4.1, 4.2 and 4.3), this does not exclude the possibility that growth rate has an influence on the plasmid loss time. For the data presented in this chapter, Cascade and plasmid abundance were not monitored during imaging so we are forced to speculate about the effects of the observed differences in growth rate between the $\Delta cas1,2$ and WT strains of each PAM. In the experiments in which the target sequence on the plasmid is flanked by the DI PAM, cells void of *cas1* and *cas2* had higher growth rates than WT cells (Fig. 4.6a). Fast growth correlates with lower Cas protein concentrations, which leads to a lower rate of interference, but could also result in lower plasmid copy numbers. On average, it takes less time to destroy a smaller number or target plasmids through interference (Fig. 3.5i). In line with the result of a lower average PLT for faster growing $\Delta cas1,2$ cells with DI targets(Fig. 3.3f), the higher average PLT in the WT strain (Fig. 4.2a) could be in part explained by the potentially higher plasmid copy number associated with slower growing cells. However, as plasmid loss still takes place multiple hours after induction, this effect might be in part counteracted by the higher Cascade levels present later in the experiment (Fig. 3.4b). Overall, it is unclear how the balance between Cascade and plasmid concentrations, and with it the rate of interference and primed adaptation, might shift in relation to changes in the growth rate (Fig. 4.7b). For the PI PAM, the WT strain grows on average faster than the $\Delta cas1,2$ strain (Fig. 4.6b), whilst exhibiting slower and more infrequent plasmid loss (Fig. 4.2b). Following previous reasoning, faster growth could result in faster plasmid loss through interference, but might also lead to a smaller fraction of cells clearing the plasmid through priming. As it is not known which is the dominant mechanism of plasmid clearance in the WT condition (interference or primed adaptation followed by primed interference), it is not straightforward to predict what would be the effect, if any at all, of a faster growth rate. Given the large difference between the PI WT and $\Delta cas1,2$ PLT distributions and frequency of plasmid loss events, combined with the wide distribution of growth rate rankings of loss lineages, it is highly unlikely that any effect of the growth rate is large enough to explain the observed differences in PLT between the two conditions. For the priming PAM, it is clear that the differences in PLT between the $\Delta cas1,2$ and WT strains is not caused by any growth rate

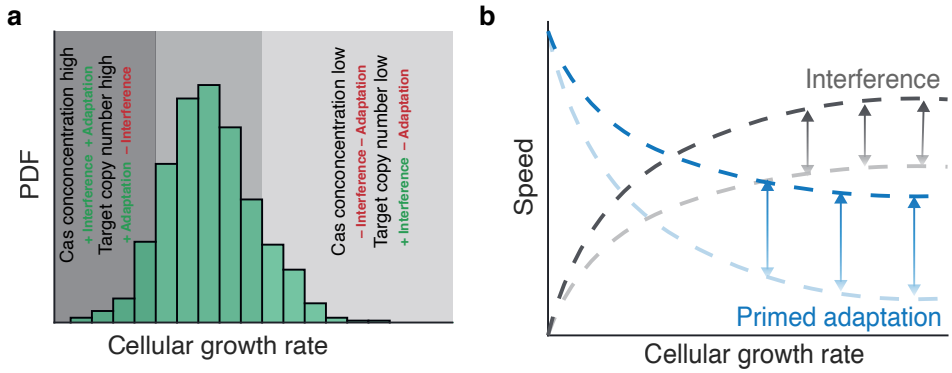


Figure 4.7: Schematic showing the hypothesised effect of cellular growth rate on target copy number concentration, Cas protein concentration, and speed of interference and adaptation. **a,** Distribution of growth rate in a cell population results in cell with various target copy number and Cas protein concentrations. Lower target copy numbers are associated with higher interference, but lower adaptation speed, whereas lower concentrations of Cas proteins result in a decrease of both interference and primed adaptation speed. **b,** Due to the effect of target copy number concentration and Cas protein concentrations, the speed of interference increases with growth rate whereas primed adaptation speed decreases. The magnitude of either effect is unknown however, as indicated by the vertical arrows. The overall effect of both on the plasmid loss speed depends on the relative change in either process.

effects (Fig. 4.6c), as Cas1-Cas2 has been shown to be required for successful spacer acquisition.[114]

We note that our data does not suggest that the interference-dependent spacer acquisition pathway should be dismissed as a model for primed adaptation. Experiments have demonstrated that degradation products from the activity of Cas3, although highly unstable, can function as pre-spacers.[130,186] It is therefore possible that pre-spacers can be taken both directly from the invading DNA as well as be selected from the pool of DNA fragments that are the result of interference. The II pathway may be the most prominent in cells due to the efficiency of recruiting Cas1-Cas2 to target-bound Cascade, however free Cas1-Cas2 in the cell might be able to spontaneously capture fragments through a facilitated diffusion mechanism as proposed by Kim *et al.*[113] The probability for the ID mechanism to result in the acquisition of a new spacer increases when interference is sustained over a longer period, such as when the rate of plasmid replication is higher than the rate of interference. As the interference rate is PAM-dependent, this might shift the balance between adaptation mechanisms for different PAM sequences in favour of interference-dependent spacer acquisition. This might result in a highly advantageous strategy for cells, as it leads to an amplification of the interference

effect while reducing the chance that invaders escape CRISPR immunity through mutations.

4.4 Future research

Future experiments are necessary to resolve the mechanism underlying the generation of pre-spacers. Sequencing of the population to determine the percentage of cells acquiring new spacers could provide insight into the dominant mechanism of plasmid loss for cells carrying targets containing PAM sequences which are proficient at both interference and primed adaptation. However, this will not provide information about the mechanism behind plasmid clearance at the single-cell level. Detection of spacer acquisition in single cells could be achieved through the use of a frame-shift system where spacer integration results in a fluorescent signal.[143,197] In addition, the present lack of data on the concentrations of CRISPR-associated proteins makes it impossible to untangle how growth rate, protein concentrations, and target copy number contribute to the probability of successful plasmid clearance. To this end, sensitive reporters of the concentrations of the various Cas proteins and the target copy number are required. In Chapter 3 a set-up in which mCherry (red fluorescent protein, RFP) is fused to Cascade (see also Vink *et al.*[110]) was successfully used to monitor Cascade concentrations in addition to plasmid loss. However, in order to additionally monitor Cas1-Cas2 and Cas3, one needs to ascertain that steric hindrance due to the presence of fluorescent tags does not block the function or activity of proteins involved in CRISPR defence, or hinder the formation of multi-protein complexes.[198] If this turns out to not be feasible, modulation of the cellular growth rate could ensure a fairer comparison between experimental conditions. However, due to differences in metabolic load this could still result in different protein concentrations.

In our current simulation framework, a model containing both adaptation pathways could be achieved by merging the reactions from both proposed mechanisms, although this would increase the number of parameters to be estimated from the data. Before performing inference, however, structural identifiability analysis should be carried out to establish the identifiability of model parameters from the available data.[199] If the model is found to be structurally non-identifiable, it may be necessary to reduce the model in complexity. In addition, systematic sampling of the parameter space is required to assess the uniqueness, robustness, and biological plausibility of parameter sets that are consistent with the experimental data.

4.5 Conclusion

In summary, in this study we have expanded on previous work (Chapter 3) to further study the role of the PAM sequence in the dynamics and variability of the

CRISPR interference and adaptation response of bacterial populations. We have shown that the PAM plays an important role in the successful clearance of invading elements, and that both interference and adaptation are strongly PAM-dependent. Surprisingly, the presence of Cas1-Cas2 appears to disrupt the degradation of targets containing PAM sequences which were previously shown to be interference-proficient. We have employed an agent-based stochastic simulation framework to show that this observation is consistent with the interference-independent model for spacer acquisition. We thus hypothesise that our analysis provides support for the existence of a priming complex which involves recruitment of Cas3 in conjunction with Cas1-Cas2 by target-bound Cascade. A mechanism in which the PAM sequence affects the relative rates of interference and priming would allow bacteria to have a diversified response to invader DNA, thus increasing the population's chance of survival. These findings open up new avenues of research on how cellular communities are able to provide robust immunity against infections by balancing rapid target destruction and acquisition of new spacers.

4.6 Methods

4.6.1 Experimental methods

Methods for cloning, growth conditions, time-lapse microscopy, image analysis, and plasmid loss detection are described in Section 3.4. Details on the KD615 (WT) and KD635 ($\Delta cas1,2$) strains and plasmids pTU166 (5′-CTT PAM, direct interference) and pTU190 (5′-CGT PAM, priming) can be found in Supplementary Tables 3.3–3.4 of Chapter 3. Plasmid pTU189 (containing the 5′-AAT PAM for the priming interference condition) targeted by KD615 (WT) and KD635 ($\Delta cas1,2$) was created by PCR amplification of pTU166 using primer BN911 in combination with BN910 (Supplementary Table 4.5).

Although for the pTU166 (DI) targeted by KD635 ($\Delta cas1,2$) and pTU190 (P) targeted by KD615 (WT) data from 2 wells from the same experiment were presented in the previous chapter, here only data from 1 well was included for a fair comparison with the other conditions, for which only 1 well was available. In Chapter 3, we have shown that data from wells within the same experiment are comparable.

4.6.2 Model implementation

Stochastic simulations were performed using the adapted Extrande algorithm[184] implemented in C++. The simulation procedure and algorithm outline is described in Appendix 3.B, for which the code is available upon request. Each data point in Fig. 4.4 and Fig. 4.5 was obtained from 10 simulated experiments of up to 7 h.

The population size of each simulation was fixed at 100 cells. See Appendix 4.A for model details and parameters.

4.6.3 Author contributions

R.E.M., S.J.J.B. and S.J.T. and conceived the project. R.E.M. performed the experiments. E.M.K. analysed the data and performed the modelling, E.M.K. wrote the chapter with input from R.E.M., S.J.J.B., S.J.T., C.F., R.W.S., and J.M.

Appendices

4.A Reaction mechanism and model parameters

Plasmids P have a maintenance mechanism in order to keep the plasmid copy number at the target concentration p^*/V_B, where V_B is the average cell volume at birth. At the start of the experiment, expression of the Cas proteins $Cas1$-$Cas2$, $Cas3$, and $Cascade$ starts. The protein expression rate k_1 is made time-dependent to emulate the experimental set-up, in which there is some delay in protein expression due to the time required for activation of the pAraBad promoter. Spacers $crRNA$ are continually transcribed from the spacer array A, and together with Cascade form the effector complex, E. The effector complex can form a reversible bond with the target plasmid.

Given the limited quantity of available data, accurate parameter estimates of the two proposed mechanistic models could not be obtained. In addition, the model structure is too complex for analytical treatment. For this reason, the parameters were altered manually and the fit to the experimental conditions was not optimised.

4.A.1 Interference-dependent (ID) model

In the ID model, the complex EP can recruit Cas3 which degrades the plasmid into small DNA fragments F, which can be picked up by $Cas1$-$Cas2$. One of these pre-spacers captured by the Cas1-Cas2 complex, $FCas12$, can be integrated into the CRISPR array as a new spacer, transforming the array to A^* which now also expresses the newly acquired crRNA, $crRNA^*$, in addition to the spacer that was already present. The effector complex containing the new spacer, E^*, has a higher binding affinity to the target plasmid. These biochemical reactions are described in Supplementary Table 4.1. Parameters are given in Supplementary Table 4.2.

Phase	Reactions
Target replication	$P \xrightarrow{k_0/(1+((P/V_t)/p_0)^2)} 2P,$ $p_0 = \frac{p^*}{V_B}/\sqrt{(\frac{k_0}{\mu} - 1)}$
Expression	$G \xrightarrow{k_1(t)} G + b_P \cdot Cas1\text{-}Cas2$ $G \xrightarrow{k_1(t)} G + b_P \cdot Cas3$ $G \xrightarrow{k_1(t)} G + b_P \cdot Cascade$ $k_1(t) = \frac{k_1}{1+exp(-k_d t)}$ Before spacer integration $A \xrightarrow{k_2} A + b_c \cdot crRNA$ After spacer integration $A^* \xrightarrow{k_2} A^* + b_c \cdot crRNA + b_c \cdot crRNA^*$ $crRNA + Cascade \xrightarrow{k_3} E$ $crRNA^* + Cascade \xrightarrow{k_3} E^*$ $crRNA \xrightarrow{k_4} \emptyset$
Interference	$E + P \underset{k_6}{\overset{k_5}{\rightleftharpoons}} EP$ $E^* + P \underset{k_8}{\overset{k_7}{\rightleftharpoons}} EP^*$ $EP + Cas3 \xrightarrow{k_9} E + Cas3 + b_F \cdot F$ $EP^* + Cas3 \xrightarrow{k_9} E^* + Cas3 + b_F \cdot F$ $F \xrightarrow{k_{10}} \emptyset$
Primed adaptation	$F + Cas1\text{-}Cas2 \xrightarrow{k_{11}} FCas12$ $FCas12 + A \xrightarrow{k_{12}} A^*$

Supplementary Table 4.1: Overview of the reactions in the interference-dependent (ID) model for primed adaptation.

Reaction	Parameter	Value
Target replication	k_0	0.125 min^{-1}
Cas proteins synthesis rate	k_1	2.4 min^{-1}
$crRNA/crRNA^*$ transcription	k_2	10 min^{-1}
$crRNA/crRNA^*$ degradation	k_3	0.014 min^{-1}
$crRNA$-$Cas/crRNA^*$-Cas complex formation	k_4	0.01 $\mathrm{M}^{-1}\mathrm{min}^{-1}$
E-P binding affinity (P)	k_5	$2e^{-5}\mathrm{M}^{-1}\mathrm{min}^{-1}$
E-P binding affinity (PI)	k_5	$2e^{-4}\mathrm{M}^{-1}\mathrm{min}^{-1}$
E-P binding affinity (DI)	k_5	$1e^{-3}\mathrm{M}^{-1}\mathrm{min}^{-1}$
EP dissociation (all)	k_6	$1e^{-4}$ min^{-1}
E^*-P binding affinity	k_7	$1e^{-3}\mathrm{M}^{-1}\mathrm{min}^{-1}$
EP^* dissociation	k_8	$1e^{-4}$ min^{-1}
Target degradation	k_9	$5e^{-3}$ min^{-1}
Fragment degradation	k_{10}	1 min^{-1}
$Cas1$-$Cas2$ picks up pre-spacer	k_{11}	$1e^{-3}\mathrm{M}^{-1}\mathrm{min}^{-1}$
Spacer integration	k_{12}	0.02 $\mathrm{M}^{-1}\mathrm{min}^{-1}$
Cas proteins burst size	b_P	3
$crRNA/crRNA^*$ burst size	b_c	3
DNA fragment burst size	b_F	5
Post-induction delay of protein production	k_d	0.025 min^{-1}

Supplementary Table 4.2: Reaction rates for the interference-dependent (ID) model.

4.A.2 Interference-independent (II) model

In the II model, the complex EP can recruit either Cas3 which directly degrades the plasmid, or $Cas1$-$Cas2$ which initiates the assembly of the primed acquisition complex PAC. $Cas3$ is recruited into the PAC, which excises a spacer from the target plasmid. This pre-spacer bound by Cas1-Cas2, $FCas12$, is integrated into the CRISPR array. From this point on, the mechanism is identical to the ID model. The biochemical reactions governing the II model are described in Supplementary Table 4.3. Parameters are given in Supplementary Table 4.4.

Phase	Reactions
Interference	$E + P \underset{k_6}{\overset{k_5}{\rightleftharpoons}} EP$ $E^* + P \underset{k_8}{\overset{k_7}{\rightleftharpoons}} EP^*$ $EP + Cas3 \xrightarrow{k_9} E + Cas3$ $EP^* + Cas3 \xrightarrow{k_9} E^* + Cas3$
Primed adaptation	$EP + Cas1\text{-}Cas2 \xrightarrow{k_{10}} PAC$ $EP^* + Cas1\text{-}Cas2 \xrightarrow{k_{10}} PAC^*$ $PAC + Cas3 \xrightarrow{k_{11}} FCas12 + Cas3 + E$ $PAC^* + Cas3 \xrightarrow{k_{11}} FCas12 + Cas3 + E^*$ $FCas12 + A \xrightarrow{k_{12}} A^*$

Supplementary Table 4.3: Overview of the reactions in the interference-independent (II) model for primed adaptation. Plasmid replication and expression phase are identical to those of the ID model described in Supplementary Table 4.1.

Reaction	Parameter	Value
Direct target degradation	k_9	$5e^{-3} M^{-1}\text{min}^{-1}$
PAC assembly	k_{10}	$5e^{-3} M^{-1}\text{min}^{-1}$
Spacer cutting by PAC	k_{11}	$1e^{-4}\text{M}^{-1}\text{min}^{-1}$
Spacer integration	k_{12}	0.1 $M^{-1}\text{min}^{-1}$

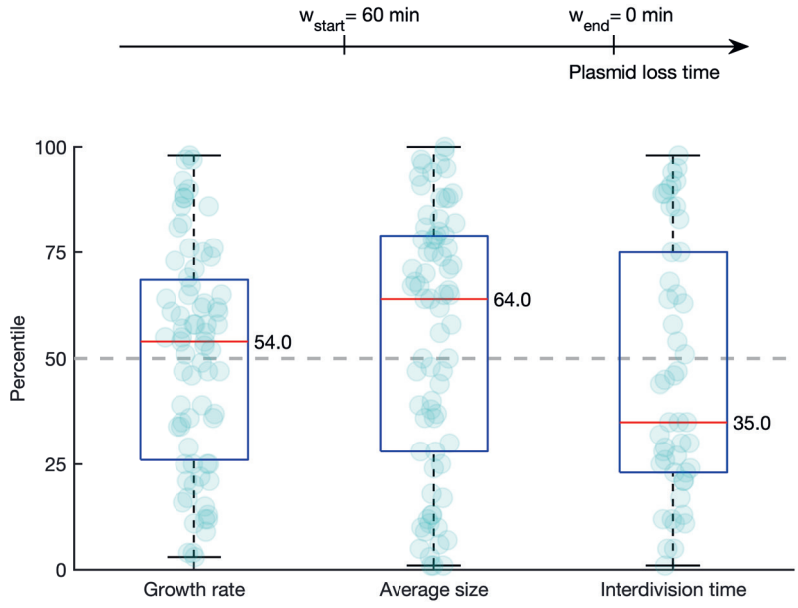
Supplementary Table 4.4: Reaction rates for the interference-independent (II) model. All other reaction rate are identical to the ID model as described in Supplementary Table 4.2.

4.B Plasmids and oligonucleotides used

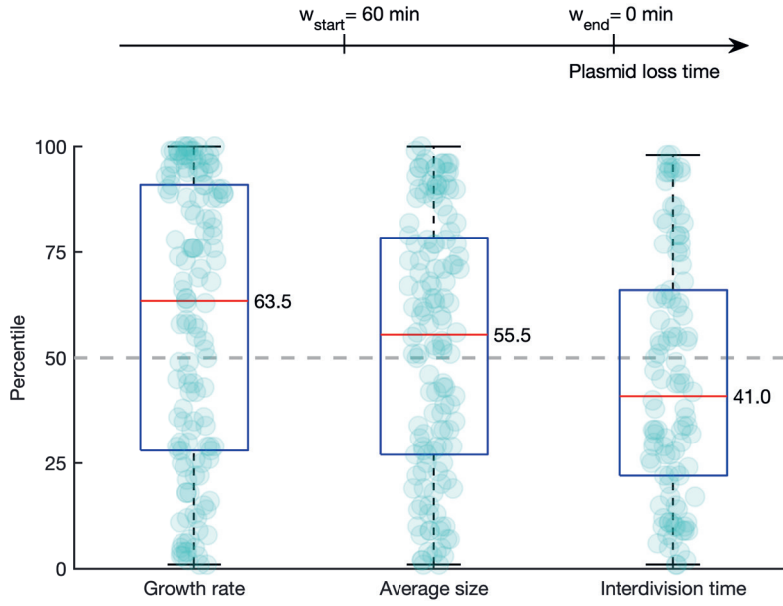
Plasmid	Description	Source
pTU189	pSC101, StrepR, TetR mVenus PS8 flanked by 'ATT' PAM	This work
Oligonucleotides	**Description**	**Sequence**
BN910	Modify pTU66 CTT PAM to AAT, Fw	TTTTGTCGACATTCTG ACGACCGGGTCTCC

Supplementary Table 4.5: Plasmids and oligonucleotides used

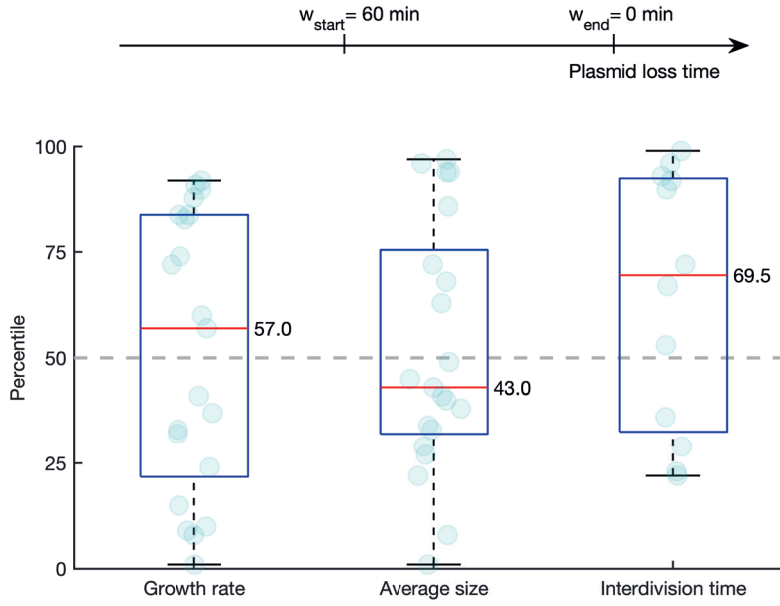
Supplementary Figures



Supplementary Figure 4.1: Growth rate, cell size and interdivision time of direct interference WT with different lookback windows Boxplots of growth rate, average cell size and interdivision time presented as the percentile rankings of all plasmid loss lineages (green) that cleared a known target via direct interference. The cell feature of interest (e.g. growth rate) was averaged over a lookback window chosen in relation to the time from plasmid loss of the lineage of interest. The same cell feature was then averaged for all non-loss lineages in the population at that same moment. The cell feature of interest was then ranked amongst the non-loss population as a percentile. We considered lookback windows of 60 minutes prior to plasmid loss. The median percentile ranking of loss lineages is indicated by a red line and black text, categories in which this value was significantly different from a ranking in the 50th percentile (p-value<0.05) are indicated in red text followed by an asterisk.



Supplementary Figure 4.2: Growth rate, cell size and interdivision time of priming interference $\Delta cas1,2$ with different lookback windows Boxplots of growth rate, average cell size and interdivision time presented as the percentile rankings of all plasmid loss lineages (green) that cleared a known target via direct interference. The cell feature of interest (e.g. growth rate) was averaged over a lookback window chosen in relation to the time from plasmid loss of the lineage of interest. The same cell feature was then averaged for all non-loss lineages in the population at that same moment. The cell feature of interest was then ranked amongst the non-loss population as a percentile. We considered lookback windows of 60 minutes prior to plasmid loss. The median percentile ranking of loss lineages is indicated by a red line and black text, categories in which this value was significantly different from a ranking in the 50th percentile (p-value<0.05) are indicated in red text followed by an asterisk.



Supplementary Figure 4.3: Growth rate, cell size and interdivision time of priming interference WT with different lookback windows Boxplots of growth rate, average cell size and interdivision time presented as the percentile rankings of all plasmid loss lineages (green) that cleared a known target via direct interference. The cell feature of interest (e.g. growth rate) was averaged over a lookback window chosen in relation to the time from plasmid loss of the lineage of interest. The same cell feature was then averaged for all non-loss lineages in the population at that same moment. The cell feature of interest was then ranked amongst the non-loss population as a percentile. We considered lookback windows of 60 minutes prior to plasmid loss. The median percentile ranking of loss lineages is indicated by a red line and black text, categories in which this value was significantly different from a ranking in the 50th percentile (p-value<0.05) are indicated in red text followed by an asterisk.

Chapter 5

Stochastic gene expression in *Arabidopsis thaliana*

This chapter is published as:
Ilka S. Araújo, Jessica M. Pietsch, Emma M. Keizer, Bettina Greese, Rachappa Balkunde, Christian Fleck, Martin Hülskamp
Nature communications 8.1 (2017): 1-9.

DOI: https://doi.org/10.1038/s41467-017-02285-7

Abstract

Although plant development is highly reproducible, some stochasticity exists. This developmental stochasticity may be caused by noisy gene expression. Here we analyse the fluctuation of protein expression in *Arabidopsis thaliana*. Using the photoconvertible KikGR marker, we show that the protein expressions of individual cells fluctuate over time. A dual reporter system was used to study extrinsic and intrinsic noise of marker gene expression. We report that extrinsic noise is higher than intrinsic noise and that extrinsic noise in stomata is clearly lower in comparison to several other tissues/cell types. Finally, we show that cells are coupled with respect to stochastic protein expression in young leaves, hypocotyls and roots but not in mature leaves. Our data indicate that stochasticity of gene expression can vary between tissues/cell types and that it can be coupled in a non-cell-autonomous manner.

5.1 Introduction

Plant development is governed by regulatory mechanisms that lead to the formation of specialized cell types and tissues in a well-organized manner. At the cellular and molecular level, however, a surprisingly high degree of stochasticity is observed.[200] One way to look at stochasticity is that it may be a problem to establish regularity. On the other hand, stochasticity might be important to break the homogeneity, which is necessary for correct pattern formation.[201,202]

In plants, stochasticity during development is best described for leaf growth. Here no correlation between growth rates and cell sizes, nuclear sizes and anisotropy was found.[203] Similarly, the length of the cell cycle and the time point at which cells switch to endoreduplication was found to be stochastic in sepals.[204] Recently, it was demonstrated that fluctuations of the transcription factor ATML1 initiate the spatial distribution of giant cells in sepals.[205] The characteristics and basis of stochastic gene expression was analysed in various organisms including bacteria, yeast, mammalian cell cultures, *Dictyostelium discoideum*, *Mus musculus* and *Drosophila melanogaster*.[7,8,206–212] The overall noise of gene expression in a given cell can be divided into two components.[7] Extrinsic noise equally affects the expression of all genes in a cell, for example, because of differences in the number of RNA polymerases or ribosomes between cells. Intrinsic noise is due to the inherent stochasticity of molecular processes influencing transcription and translation. As a consequence, the expression of individual genes fluctuates over time.

In this work we analyse the noisiness of gene expression in *Arabidopsis thaliana* with emphasis on two questions: First, is intrinsic and extrinsic noise different in different tissues or cell types? It might be expected that stochasticity changes during cell differentiation or endoreduplication. Endoreduplication leads to higher copy numbers of genomes, which could balance the fluctuation of individual gene copies and a reduction of intrinsic noise. Second, is stochasticity of gene expression coupled between cells in a tissue? This could be the case because cellular conditions are inherited during cell divisions or because plant cells are well connected with each other through plasmodesmata[213] such that they could cross regulate and balance each others transcription.

We demonstrate that gene expression fluctuates over time. In addition, we show that extrinsic noise is higher than intrinsic noise and that extrinsic noise in stomata is lower than in other tissues/cell types. Our spatial analysis of stochastic gene expression revealed coupling between cells in some but not all tissues.

5.2 Results

5.2.1 Temporal analysis of fluctuations

Fluctuations of gene expression over time have been successfully measured in single-cell systems including bacteria and human tissue cultures.[214–216] In a first experiment, we aimed to detect a temporal correlation of protein expression in intact plant leaves by determining the correlation of protein levels between different time points (auto-correlation).[217] Towards this end we developed the following experimental setup: (1) We decided to compare the protein levels at only two time points because the experiments have to be done with excised leaves and prolonged maintenance is expected to produce artefacts. (2) Under steady state expression, we had difficulties to detect relative differences of protein levels within 3 h time intervals. We therefore used the photoconvertible NLS-KikGR. KikGR can be irreversibly converted from a green fluorescent protein (KikG) to red fluorescent protein (KikR) by 405 nm illumination.[218] Using this system we determined the production of new proteins[219] by converting KikG to KikR followed by the quantification of newly produced green fluorescent KikG after 3 and 6 h. (3) We targeted the fluorescent protein to the nucleus by adding a NLS sequence to facilitate the selection of single cells. (4) We expressed NLS-KikGR under the strong ubiquitously and constitutively active cauliflower 35S and the UBIQUITIN10 (UBQ10) promoters. Fairly strong constitutive promoters were chosen to reach sufficiently high expression levels and thereby fluorescence intensities to measure fluctuations. Although this limits general conclusions, this procedure should result in a conservative estimation of intrinsic noise in our experiments as experimental data and theoretical considerations show that constitutive promoters show the lowest intrinsic noise.[7,220] Two different promoters were selected to exclude that we are exploring a specific property of one promoter. (5) We excluded that movement of NLS-KikGR between cells leads to correlation between neighbouring cells by using a KikGR protein version that forms tetramers[218] which should not move between cells. We confirmed this by expressing the KikGR protein in single epidermal *Arabidopsis* cells by biolistic transformation. In these experiments we found no fluorescence in the neighbouring cells (Supplementary Fig. 5.1d-e[218,221]).

For the temporal correlation analysis, transgenic p35S:NLS-KikGR and pUBQ10:NLS-KikGR *Arabidopsis* leaves were dissected and kept in darkness for 36 h to reduce the amount of already converted red NLS-KikR protein. NLS-KikG was converted to the NLS-KikR by confocal laser scanning microscopy (CLSM). The amount of KikG was determined at three time points (0 h, 3 h and 6 h, Fig. 5.1a, b). To minimise technical errors and to control bleaching effects we measured each nucleus at each time point two times and used the mean for further calculations. We used at least three biological replicas to determine the average Spearman and Pearson's correlation coefficients of the fluorescence levels between the 3-h intervals: p35S:NLS-KikGR

(number of leaves=4, n=393 cells, Spearman's: $r = 0.83$, Pearson's: $r = 0.88$, example leaf: Fig. 5.1c, Supplementary Fig. 5.2), pUBQ10:NLS-KikGR (number of leaves=3, n=153 cells, Spearman's: $r = 0.76$, Pearson's: $r = 0.80$, example leaf: Fig. 5.1d, Supplementary Fig. 5.3). Control experiments with p35S:NLS-KikGR plants without the 36 h dark treatment exhibited fluctuations in a similar range (number of leaves=10, n=465 cells, Spearman's: $r = 0.59$, Pearson's: $r = 0.68$, Supplementary Fig. 5.4). The finding that the correlations coefficients were always clearly below 1 (perfect correlation) indicates that we can detect fluctuations between the two 3-h time intervals.

In order to put the experimentally determined correlation coefficients into a context we used a modelling approach aiming to address two questions: Is the linear two-stage model shown in Fig. 5.2a sufficient to explain the data? How does cell-to-cell variability affect the decay of the auto-correlation? Towards this end, we analytically calculated the non-stationary auto-correlation function of the linear two-stage model with a stochastic translation rate ν_1 as a source of extrinsic noise (Appendix 5.A). We further simulated the stochastic KikGR system. In Fig. 5.2b we show example trajectories before and after the converting light pulse. In order to make a prediction for the value of the temporal auto-correlation between 6 h and 3 h after the conversion we need to obtain estimates for the model parameters. We estimated the degradation rate d_1 of KikR using the measured values of the red fluorescent protein at 3 h and 6 h after conversion (Appendix 5.A; $d_1 = 0.09\mathrm{h}^{-1} \pm 0.023\mathrm{h}^{-1}$). The other model parameters are unknown and it is not easy to obtain reliable estimates. However, we can show that the auto-correlation for the non-stationary two-stage process with extrinsic translational noise is bounded from below by the much simpler auto-correlation function of the one-stage death-birth process (taking only protein production and decay into account), which only depends on the stability of the protein (Fig. 5.2c, Appendix 5.A, Supplementary Fig. 5.24). According to this we estimated the value r for auto-correlation of the KiKG gene expression between 3 h and 6 h to be in the range $1 \geq r \geq 0.6$ (Appendix 5.A). Our experimental data are consistent with this expectation suggesting that the two-stage model provides a good estimate for the underlying noise. Cell-to-cell variability prolongs the auto-correlation time, given that the correlation time of the extrinsic noise is longer than the correlation time of the intrinsic fluctuations (an assumption underlying our analytical calculations, Appendix 5.A).

5.2.2 Extrinsic and intrinsic noise in different tissues

To enable a spatial analysis of the intrinsic and extrinsic noise we adopted a dual reporter strategy initially used in bacteria and yeast (Fig. 5.3a).[7,8] Extrinsic noise is seen when both marker values correlate and show the same variation. Intrinsic noise is recognized when the two marker values are not correlated in single-cell

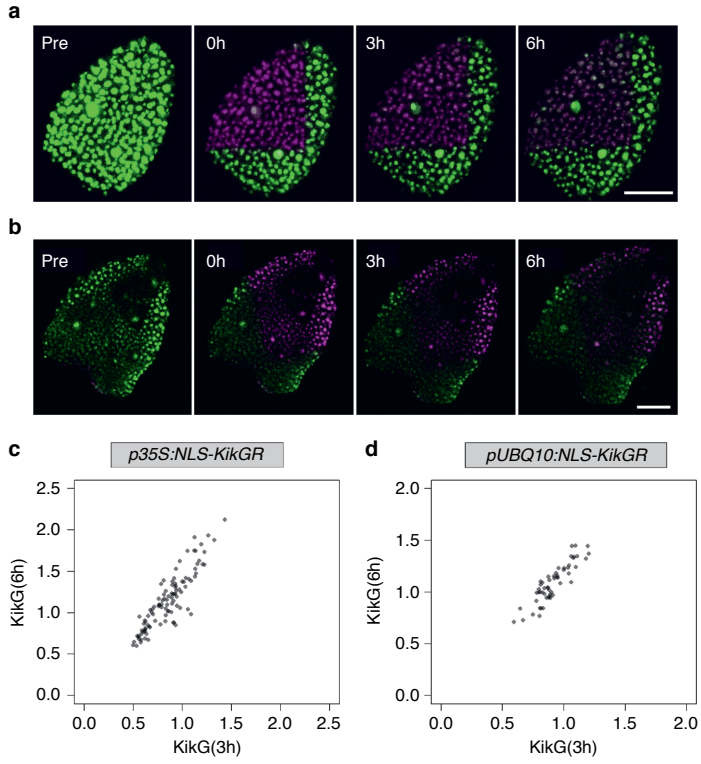


Figure 5.1: Temporal analysis of fluctuation in p35S:NLS-KikGR and pUBQ10:NLS-KikGR lines. a, Confocal laser scanning microscopy (CLSM) images of p35S:NLS-KikGR before (pre) and after conversion (0 h, 3 h and 6 h). b, CLSM images of pUBQ10:NLS-KikGR before (pre) and after conversion (0 h, 3 h and 6 h). Scale bar: 50μm. c, Scatter plot of p35S:NLS-KikG expressing cells (n=103) obtained from one representative leaf. The normalised mean fluorescence intensity of the cells at 3 h is plotted against the normalised mean fluorescence intensity of the cells at 6 h. Data points are shown in grey, overlapping data points appear black. d, Scatter plot of pUBQ10:NLS-KikG expressing cells (n=55) obtained from one representative leaf.

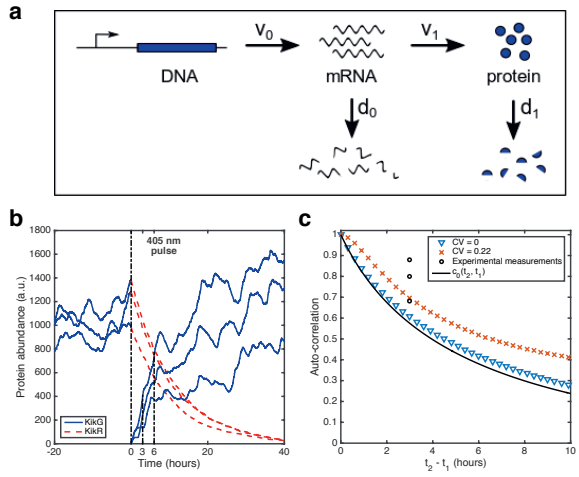


Figure 5.2: Theoretical analysis of fluctuations. a, Schematic illustration of the two-stage stochastic gene expression model. ν=production rate, d=degradation rate. b, Modelling of three stochastic realizations of the KikGR reporter. After a 405 nm pulse the green fluorescence-emitting KikG is transformed into red fluorescence-emitting KikR. The auto-correlation between KikG at 3 h and 6 h is calculated. Parameters are: $\nu_0 = 2.25h^{-1}$, $d_0 = 1.125h^{-1}$, $d_1 = 0.09h^{-1}$ and $\nu_1 = 41.825h^{-1}$, $\nu_1 = 48.506h^{-1}$, $\nu_1 = 46.069h^{-1}$ for the three different trajectories. c, Modelling of the non-stationary auto-correlation of the two-stage gene expression model in presence of extrinsic noise (crosses and triangles) as calculated from stochastic simulation of the KikGR reporter from 105 trajectories (Appendix 5.A) and the theoretical non-stationary auto-correlation of a birth-death process $c_0(t_1, t_2)$ (black solid line) as a lower bound of the non-stationary auto-correlation (Appendix 5.A). The extrinsic noise is simulated as cell-to-cell variations in the protein translation rate ν_1 with different coefficients of variation (CV). Parameters for the two-stage model are: $\nu_0 = 2.25h^{-1}$, $d_0 = 1.125h^{-1}$, $d_1 = 0.09h^{-1}$ and $< \nu_1 >= 45h^{-1}$. In the case of no extrinsic noise ($Var(\nu_1) = 0h^{-2}$, blue triangles), the auto-correlation of the two-stage model approaches that of the birth-death model. With increasing extrinsic noise ($Var(\nu_1) = 100h^{-2}$, red crosses) the auto-correlation increases. The reason for this is that the covariance and the variance become dominated by the extrinsic noise, for which a much longer correlation time was assumed.

measurements. We generated transgenic plants expressing 2xNLS-YFP and 2xNLS-CFP under the control of the 35S promoter. We used YFP and CFP fusions to two nuclear localization signals (2xNLS) for two reasons. First, by targeting the signal to one defined region in the cell, the nucleus, we improved the accuracy of measurements. Second, the targeting of the marker to the nucleus reduces the intercellular mobility that would lead to an underestimation of fluctuations[221] (Appendix 5.B, Supplementary Fig. 5.1a-c, e).

The analysis of pavement cells in young and mature rosette leaf stages revealed intrinsic and extrinsic noise. As shown in Fig. 5.3b the colour of individual nuclei ranged from green to magenta in merged YFP/CFP pictures indicating that the relative expression of the two 35S promoters driving 2xNLS-CFP and 2xNLS-YFP differs from cell to cell. This is indicative for intrinsic noise. Plotting the mean CFP values against the mean YFP values revealed intrinsic and extrinsic noise for young and mature leaves (representative leaf shown in Fig. 5.3c,d, Supplementary Figs. 5.5 and 5.6). The statistical analysis revealed significantly higher extrinsic noise than intrinsic noise (Appendix 5.C, Fig. 5.3e,f, young leaf: $p = 1.1 \times 10^{-5}$ and mature leaf: $p = 7.6 \times 10^{-5}$, Wilcoxon rank-sum test). Thus, extrinsic noise is the major source of noisy gene expression in young and mature rosette leaves. This parallels previous findings in yeast.[201,217,222] These conclusions were confirmed using an independently transformed *Arabidopsis* line carrying the *p35S:2xNLS-YFP p35S:2xNLS-CFP* constructs (Supplementary Fig. 5.7a-b, Supplementary Fig. 5.8, Supplementary Fig. 5.9). To test whether the high extrinsic noise is specific to the 35S promo $pUBQ10 : 2xNLS - CFP$ ter, we also tested the UBQ10 promoter. Stably transformed $pUBQ10 : 2xNLS - YFP$ $pUBQ10 : 2xNLS - CFP$ *Arabidopsis* plants revealed similar behaviour as described for the 35S promoter (Supplementary Fig. 5.7c-e, Supplementary Fig. 5.10, Supplementary Fig. 5.11).

These findings raised the question whether noise differs in different cell types or tissues. We therefore determined the noise additionally in stomata, epidermal hypocotyls and root tip cells of *p35S:2xNLS-YFP p35S:2xNLS-CFP* plants (Fig. 5.4, Supplementary Figs. 5.12-5.18). Intrinsic and extrinsic noise were found in a similar range in all tissues/cell types except for stomata. For stomata we found a clearly and significantly lower extrinsic noise in both independent transgenic lines (root/stomata: $p = 0.0007$, $p = 0.03$, hypocotyl/stomata: $p = 0.0002$, $p = 0.03$, pavement cells in young leaves/stomata ($p = 1.08 \times 10^{-5}$, $p = 0.006$, Wilcoxon rank-sum test). This indicates that extrinsic noise can vary in a tissue/cell type specific manner.

5.2.3 Noise in cells with different DNA contents

Next we tested the concept whether higher endoreduplication levels lead to reduced noise. We took advantage of the fact that pavement cells exhibit a wide range of ploidy levels between 2C and 64C.[203,223] This allowed us to study the correlation

between ploidy and noise for one specific cell type. As higher DNA contents lead to increased nuclear sizes, we used the maximal area of each nucleus as an estimator for the DNA content in our correlation studies.[224] We determined the maximal nuclear area of leaf epidermal cells in a stack of images and analysed the YFP and CFP values. We estimated the median of nuclear area of all pavement nuclei and considered four quartiles separately in *p35S:2xNLS-YFP p35S:2xNLS-CFP* and $pUBQ10 : 2xNLS - YFP$ $pUBQ10 : 2xNLS - CFP$ plants. Intrinsic noise levels were similar in all four quartiles (Supplementary Fig. 5.19b-e; Supplementary Fig. 5.20b). This finding is not unexpected as already two gene copies in a diploid cell might be sufficient to balance fluctuations in one of them. For extrinsic noise we observed the trend that larger nuclei have slightly more extrinsic noise (Supplementary Fig. 5.19c,f; Supplementary Fig. 5.20c). Increased extrinsic noise in larger cells could be explained by less uniform cellular states in cells with a higher DNA content or by changes due to a progression of cell differentiation.

5.2.4 Correlation of noise between neighbouring cells

Finally, we aimed to understand whether fluctuation in gene expression is coupled in neighbouring cells. In contrast to unicellular bacteria or yeast one could envision that in tissues extrinsic noise might be correlated in immediately neighbouring cells. This could result from initially similar cellular conditions in daughter cells or cellular connectivity of plant cells by plasmodesmata. To exclude that the low movement rates of the fluorescent marker protein used in this study leads to a correlation between neighbouring cells we estimated the transport coefficient for the mobility between cells. We found that the contribution to a cell-cell correlation due to mobility of the fluorescent protein is negligible (Appendix 5.B). We reasoned that the intrinsic gene expression noise is mechanistically decoupled between cells, *i.e.*, the expression of the fluorescent protein in one cell does not influence the expression in another cell. Moreover, due to cell division the stochastic gene expression in the growing tissue is never in stationary state. This would yield an extra contribution from the intrinsic noise if one would analyse the spatial correlation using a single reporter system (Appendix 5.B). However, using the dual reporter *p35S:2xNLS-YFP* $p35S : 2xNLS - CFP$plants for a cross-analysis (relating CFP to YFP and vice versa, see Fig. 5.5a) we can estimate the variance of the extrinsic noise (e.g., variability in ribosome number, transcription factor abundance[12, 225]) and the covariance between the extrinsic noise of neighbouring cells (Fig. 5.5b). The covariance between stochastically identical cells is equal to the variance of the extrinsic noise. Therefore, it is necessary to normalise the covariance using the variance of the extrinsic noise to obtain a measure between 1 and –1 for the correlation between neighbouring cells (Appendix 5.B). In this way we estimated the correlation of the extrinsic noise between nearest neighbour cells in young leaves and found a weak but significant correlation ($r = 0.34$, $p < 0.0002$, randomisation

test, Fig. 5.5d). To test whether this correlation ceases with increased distances we calculated the correlation between each nucleus and its 39 closest neighbours. We observed a drastic reduction with increasing distance. To judge over how many cell diameters extrinsic fluctuations are correlated we determined the average nearest neighbour distance and used this value to define five concentric rings (tiers) of cell distances (Fig. 5.5e). These data indicate that on young leaves correlation is mainly found between immediately neighbouring cells. By contrast, we detected no correlation between neighbouring cells on mature leaves ($r = 0.02$, $p = 0.433$, randomisation test, Fig. 5.5f). Similar results were obtained with the second *p35S:2xNLS-YFP p35S:2xNLS-CFP* line ($r = 0.423$; $p = 0.0014$; Supplementary Fig. 5.21) and with a $pUBQ10 : 2xNLS - YFP\ pUBQ10 : 2xNLS - CFP$ line ($r = 0.413$; $p = 0.0003$, randomisation test, Fig. 5.5g-i). Thus, correlation is only found in young but not in mature leaves (see confirmation in independent transformants in Supplementary Fig. 5.21). A distance dependent correlation of extrinsic noise was also found in hypocotyl and root tissues for two independent *p35S:2xNLS-YFP p35S:2xNLS-CFP* lines (Supplementary Fig. 5.22, Supplementary Fig. 5.23).

To judge to what extent inheritance of mRNA and protein content can explain the observed next-neighbour correlation we used the two-stage gene expression model shown in Fig. 5.1a under the following assumptions: At time $t = 0$ the mRNA and protein content of a mother cell expressing the dual reporter system is copied to two daughter cells. Thus daughter cells have identical initial condition for the mRNA and protein amount. Except for the translation rate, the cells inherit all parameter values from the mother cell (Fig. 5.5c). The translational rates of the daughter cells are stochastic and in general different from each other (extrinsic noise).

When cells divide the mRNA and protein content of the mother cell is inherited to the daughter cells. When calculating the CFP-YFP cross-correlation of the dual reporter system one avoids to introduce a correlation due to identical initial conditions. However, even though the initial conditions of CFP and YFP of the daughter cells are different they come from the same distribution, the underlying protein distribution of the mother cell. This induces a correlation which will decay with time. To show this we start with Eq. (5.9) and make the dependence on the extrinsic noise process z_0 of the mother cell explicit:

$$\begin{aligned}\mathrm{cov}(c_1, y_2) &= \langle\langle\langle\langle c_1|N_1, z_1, z_0\rangle\rangle_N\langle\langle y_2|N_2, z_2, z_0\rangle\rangle_N\rangle_{z_1}\rangle_{z_2}\rangle_{z_0} \\ &\quad -\langle\langle\langle\langle c|N, z, z_0\rangle\rangle_N\rangle_z\rangle_{z_0}\langle\langle\langle\langle y|N, z, z_0\rangle\rangle_N\rangle_z\rangle_{z_0}\end{aligned}$$

If we assume for simplicity that z_0, z_1 and z_2 are independent we find:

$$\mathrm{cov}(c_1, y_2) = \langle\langle\langle\langle x|N, z\rangle\rangle_N\rangle_z^2\rangle_{z_0} - \langle\langle\langle x|N, z\rangle\rangle_N\rangle_z\rangle_{z_0}^2.$$

At $t = 0$, right after cell division the expression above yields σ^2_{ext}, while for $t \to \infty$ the gene expression of the daughter cells are independent from the stochastic processes in the mother cell, therefore expect we $\text{cov}(c_1, y_2) \to 0$ for large t. It follows that for the correlation (defined by Eq. 5.10) between daughter cells $r(t) \leq 1$ with $r(t = 0) = 1$ and $\lim_{t\to\infty} r(t) = 0$ holds. To determine $r(t)$ we calculate the non-stationary covariance between two daughter cells as well as σ^2_{ext} using the two-stage gene expression model. Because we only wish to estimate the contribution of cell division to the overall spatial correlation, we employ a simple cell division model. At time $t = 0$ an exact copy of the mother cell is produced. Thereby we avoid all complication introduced by cell growth. We obtain:

$$\sigma^2_{\text{ext}}(t) = \left(\langle a^2 b^2\rangle - \langle ab\rangle^2\right)\left(1 - 2e^{-d_1 t} + 2e^{-2d_1 t}\right) \tag{5.1}$$

$$\text{cov}(c_1, y_2)(t) = \left(\langle a^2 b^2\rangle - \langle ab\rangle^2\right) e^{-2d_1 t} \tag{5.2}$$

$$r(t) = \frac{e^{-2d_1 t}}{(1 - 2e^{-d_1 t} + 2e^{-2d_1 t})}. \tag{5.3}$$

We then test the validity of the obtained expression for $r(t)$ by simulating the dual reporter system as described in Appendix 5.D.1. Typical trajectories of CFP and YFP for mother and daughter cells are shown in Supplementary Fig. 5.25A. After cell division the trajectories for CFP and YFP become different because of the different translational rates (extrinsic noise) and the stochasticity of the gene expression systems itself (intrinsic noise). From the simulation we estimate the temporal correlation between the two daughter cells using Eq. 5.11 (Appendix 5.B). The resulting time-dependent cell-to-cell correlation rate is shown in Fig. 5.6, which are in agreement with the theoretical predictions.

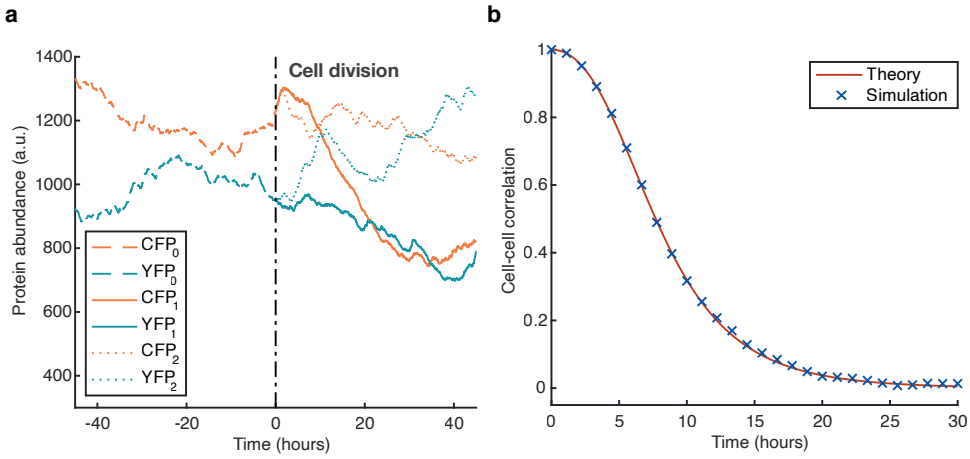


Figure 5.6: Correlation due to inheritance. Cell division with dual reporter system. At $t = 0$ an identical copy of the mother cell is produced. All parameters of the two-stage model are inherited, besides the translational rates, which are in general different between mother and daughter cells. (a) Example trajectories of the situation before and after cell division are shown. The daughter cells inherit mRNA and protein content from the mother cell. All parameters of the two- stage gene expression model are as well inherited besides the translational rates, which are different between mother ($\nu_1 = 14.581$) and daughter cells ($\nu_1 = 10.449$ and $\nu_1 = 15.917$, respectively). The other parameters are: $\nu_0 = 2.25$, $d_0 = 1.125$, $d_1 = 0.029$. (b) Correlation calculated using Eq. 5.11 between daughter cells, computed from 10^5 simulated trajectories. The solid line denotes theoretical predictions of the autocorrelation given by Eq. 5.30, crosses denote the stochastic simulation results. Parameters are: $\nu_0 = 2.25$, $\rangle\nu_1\langle = 14.5$, $Var(\nu_1) = 5$, $d_0 = 1.125$, $d_1 = 0.029$.

5.2.5 Estimating the contribution of cell division to the measured next neighbour correlation

As shown above, cell division introduces a correlation between neighbouring cells. This means that even though there are no further correlated processes, one may find a correlation due to inheritance of mRNA and protein. It is therefore important to estimate the value of this contribution. When we calculate the correlation between cells and their next neighbours we do not know which cells are daughter cells, but to make progress we reason that for any given cell two cells out of its neighbouring cells are progeny cells from the last two cell division events. Because we also do not know when after the cell division we observe the cells, we argue that all times within the intervals $[0, T]$ for the last cell division and $[T, 2T]$ from the next to last cell division are equally likely, where T is the inverse cell division rate. For young leaves the observed division rate of *Arabidopsis* leaf epidermal cells is $T = 33$ h.[226]

We average $r(t)$ over these time intervals:

$$r_1 = \frac{1}{T}\int_0^T r(t)dt, \quad r_2 = \frac{1}{T}\int_T^{2T} r(t)dt$$

and find $r_1 = 0.69$ and $r_2 = 0.11$. We used $d_1 = 0.029$ h^{-1} as the degradation rate for CFP and YFP.[227] On average the cells have five neighbouring cells. Given that two from these five cells are correlated to the center cell through inheritance of mRNA and protein content and the others are uncorrelated we arrive at our final estimate for the contribution of cell division to the measure next-neighbour correlation: $r = (r_1 + r_2)/5 \approx 0.16$

This indicates that the inheritance of mRNA and protein content is not sufficient to explain the observed spatial correlation. To estimate the maximal correlation caused by inheritance we considered the case that not only the mRNA and protein content is inherited but also all rates related to gene expression (*i.e.*, all rates of the daughter cells are identical and equal to the rates of the mother cell and do not change over time). In this case we found a correlation of $r = 0.4$ which is close to the experimentally observed value.

5.3 Discussion

As reported before in bacteria, yeast and animals we report in this manuscript fluctuations of gene expression in 3 h time intervals. Interestingly, a recent publication by Meyer et al.,[200] provided evidence that such a temporal fluctuation can be fixed and translated into different cell differentiation responses.

The use of a dual reporter system enabled us to distinguish between extrinsic and intrinsic noise. Consistent with results in yeast,[201,217,222] we found that extrinsic noise is the major source of noisy gene expression in young and mature rosette leaves. This indicates that the physiological state of plant cells equally affecting expression of all genes creates more noise than the intrinsic stochasticity of molecular processes influencing transcription and translation. It is therefore conceivable that differences of extrinsic noise in different cell types reflects different physiological properties or states of the cell types. Consistent with this, our theoretical analysis of the spatial correlation suggests, that the inheritance of mRNA and proteins is not sufficient to explain the spatial correlation and that the inheritance of cellular conditions (e.g., ribosome number, stress status) and/or cell-cell communication is required. This explanation also fits the finding that we found no spatial correlation in mature leaves as both, cell division rates[226] and the number of open plasmodesmata ceases in mature tissues.[228] In this light, spatial coupling of extrinsic noise suggests that some processes needed for gene expression co-vary. One possible explanation is

that tissues are composed of micro-domains with different physiological properties that lead, e.g., to different numbers of accessible RNAs and/or ribosomes.

5.4 Methods

5.4.1 Construct generation

pENS-YFP GW (GlyphosatR) and pENS-CFP GW (GlyphosinatR) were modified by introducing the phosphorylated linker $SalI - SV40NLS - XhoI$ (5'-CTCGAGATGCCAAA GAAGAAAAGAAAAGTTGAAGATCCTGGGTCGAC-3') into the XhoI restriction site of the vectors. For generation of pENS-2xNLS-YFP GW and pENS-2xNLS-CFP GW the ligation procedure was repeated and an additional SV40NLS was introduced into the XhoI site. In order to introduce a stop codon downstream of the YFP sequence an LR reaction (Gateway®cloning of system Invitrogen) was performed with pENTR1A-ccdB[229] and the destination vectors. *p35S:NLS-KikGR* and *pUBQ10:NLS-KikGR* were generated by LR reactions using pENTRA-NLS-KikGR and the pAMPAT plasmids. All constructs and their use in different experiments are summarized in Supplementary Table 1.

5.4.2 Transient transformation by particle bombardment

Arabidopsis leaves were transiently transformed by particle bombardment using a particle gun (gene gun). 0.8 μg DNA of each construct were used and pipetted into one reaction tube. 10 μl gold (30 mg/ml; diameter 1 μm), 20 μl CaCl2 2.5 M and 8 μl spermidine 0.1 M were added. After incubation for 10 minutes at room temperature the gold suspension was centrifuged (10 s, 10,000 r.p.m.) and the coated gold particles were resuspended in 100 μl 70% EtOH. After a second centrifugation step (10 s, 10,000 r.p.m.) 50 μl absolute EtOH were added to the gold particle pellet. After resuspending the pellet, the suspension was centrifuged again (10 s, 10,000 r.p.m.). Finally, the gold particle pellet was resuspended in 15 μl abs. EtOH and placed onto a plastic disc (macro carrier). After drying, the macro carrier was placed into the particle gun and the gold particles were used for bombardment of leek cells. Rupture disks (900 psi) and a vacuum of 26 Hg (inch of mercury, equivalent to 3.38 Pa at 0 °C) were applied during bombardments. After bombardment the samples were stored in darkness at room temperature and were analysed 16-24 h after the transformation procedure.

5.4.3 Stable transformation of *Arabidopsis thaliana*

5 ml pre-culture of agrobacteria containing the desired construct were grown over night. At the next day 200 ml were inoculated with 500 μl of the pre-culture. After 24 h 10 g sucrose and 50 μl Silwet L-77 were added to the culture. Plant flowers

were dipped into the suspension for 10 s. For double transformation 100 ml of each culture were mixed shortly before transformation as described previously.[230] 10 g and 50 μl Silwet L-77 were added and flowers were incubated in the suspension for 10 s.

5.4.4 Confocal laser scanning microscopy

Confocal laser scanning microscopy (CLSM) images were generated using Leica TCS SPE. Images were analysed and quantified using the software ImageJ. Mean grey values (0=black; 255=white) of regions of interest (ROIs) of 8 bit images were used for calculations. Analyses were always performed with overlaying maximum Z-stack projection images. Laser, gain, and detection parameters were never changed for all image acquisitions (Supplementary Table 2).

5.4.5 Measurement of photoconvertible NLS-KikGR

Single leaves (leaf number 3 or 4) of 7 days old stably transformed *p35S:NLS-KikGR* and *pUBQ10:NLS-KikGR* plants (Col-0) were imaged by CLSM (Leica TCS SPE). The leaves were placed onto 1% MS agar on a cover slip for imaging. NLS-KikG was photoconverted to NLS-KikR by a 405 nm laser line (100% laser power) exciting the sample for 5 to 10 s. NLS-KikR degradation and re-synthesized NLS-KikG were sequentially imaged two times at three different time points (0, 3, and 6 h) with the defined Z-slide distance of 3.0 μm. Each nucleus was manually selected from Z-stack projections (512px$\times$512px) of entire leaves to find the maximal cross-section. The boundary of the nucleus was always clearly seen independent of the fluorescence intensity. Subsequently, we determined the mean grey intensity of the selected region using ImageJ. Mean grey values of ROIs were used for calculation of protein degradation of NLS-KikR and protein synthesis of NLS-KikG.

To test whether NLS-KikGR can move between cells, we transformed single *Arabidopsis* leaf epidermal cells by biolistic transformation with p35S:NLS-KikGR. Among 20 transformed cells we found not a single case where fluorescent signal could be detected in the neighbouring cells (Supplementary Fig. 5.1d-e).

5.4.6 Measurement of CFP and YFP fluorescence intensities

Stably transformed *p35S:2xNLS-YFP p35S:2xNLS-CFP* and $pUBQ10:2xNLS-YFP\ pUBQ10:2xNLS-CFP$ plants (Col-0) were imaged by CLSM (Leica TCS SPE). Laser, gain and detection parameters were never changed for all image acquisitions. CFP and YFP fluorescence intensities were sequentially imaged with the defined Z-slide distance of 1.51 μm. Nuclei were selected and analysed as describe above. The mean background of each channel and image was measured separately and subtracted from the CFP and YFP mean grey values and the data

were normalised using the mean fluorescence of the data set. For raw images see Supplementary Fig. 5.26. Calculations of extrinsic and intrinsic noise were performed for each image separately and finally all noise values of each image of the same tissue were presented in a box plot including the corresponding median and mean values. We excluded samples from the analysis in which the YFP and CFP value distributions were significantly different in a Kolmogorov-Smirnov test to exclude that a skewing between the two channels influences our analysis. For all statistical analysis we confirmed that the data structure is adequate. Each root tip was virtually rotated in the $z - y$ axis to ensure a horizontal position of the root tip in the image. Only the upper 15 μm layer was analysed for calculations of noise to select only epidermal cells.

5.4.7 Code availability

The codes that support the findings of this study are available from `https://gitlab.com/wurssb/Stochastic_GE_in_Arabidopsis_thaliana.git`.

5.4.8 Data availability

Additional data that support the findings of this study are available on request.

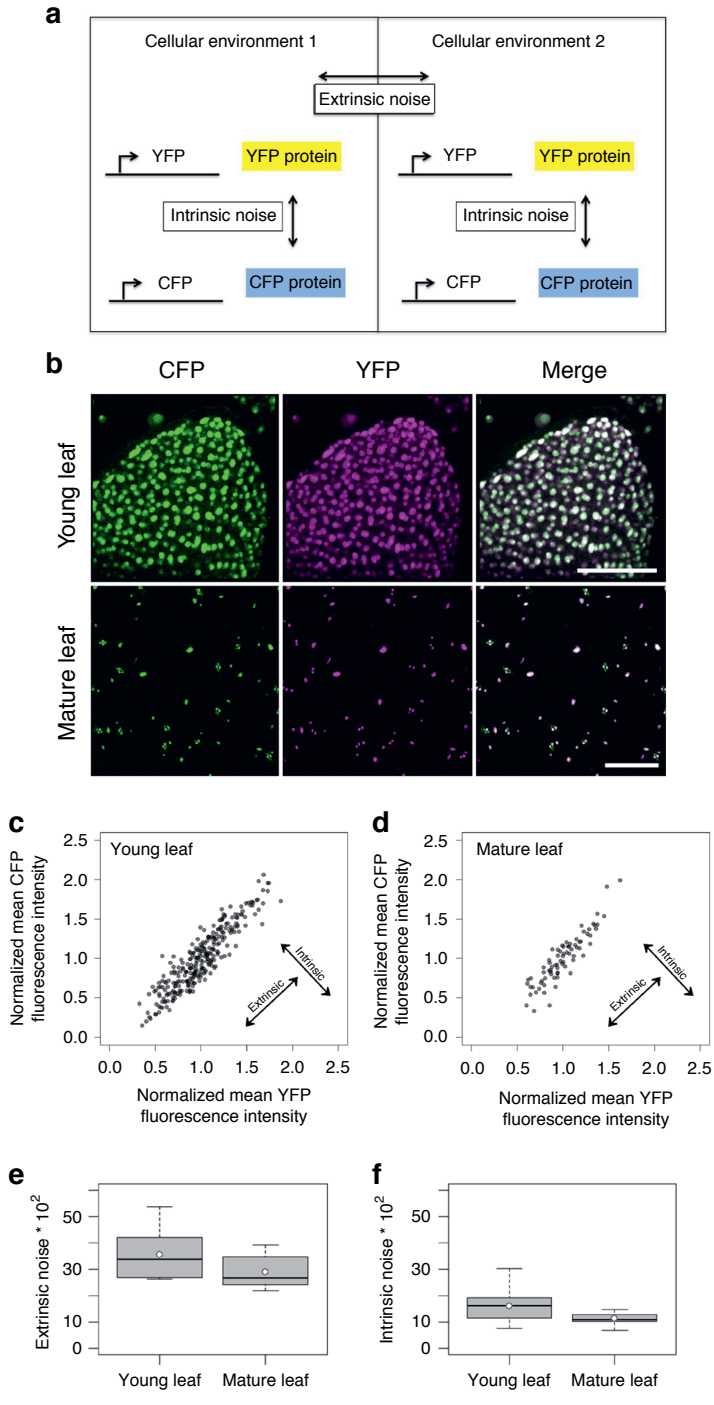


Figure 5.3: Figure caption on next page.

Figure 5.3: Intrinsic and extrinsic noise in young and mature rosette leaves of *p35S:2xNLS-YFP p35S:2xNLS-CFP* plants. a, Schematic illustration of the experimental setup to determine the intrinsic and extrinsic noise. b, CLSM images of a young, developing leaf and a mature leaf of a *p35S:2xNLS-YFP p35S:2xNLS-CFP* line. CFP is shown in green, YFP in magenta and the same fluorescence levels of both is indicated in white. Note, stomata show autofluorescence in the CFP channel. Scale bars: 50 μm (young leaf) and 100 μm (mature leaf). c, Scatter plot of the normalised CFP mean fluorescence intensity plotted against the normalised YFP mean fluorescence intensity of single cells in one representative young leaf (n=284). Pearson's correlation coefficient=0.914, Spearman's correlations coefficient=0.905. Data points are shown in grey, overlapping data points appear black. d, Scatter plot of the normalised CFP mean fluorescence intensity plotted against the normalised YFP mean fluorescence intensity of single cells in one representative mature leaf (n=76). Pearson's correlation coefficient=0.909, Spearman's correlation coefficient=0.906. e, Box plot of extrinsic noise measurements of young (n=10 leaves with a total number of 2219 cells, median=33.5) and mature leaves (n=10 leaves with a total number of 757 cells, median=26.6). The extrinsic noise was slightly but not significantly higher in young leaves as compared to mature leaves ($p = 0.075$ Wilcoxon rank-sum test). f, Box plot of intrinsic noise measurements of young (n=10 leaves with a total number of 2219 cells, median=16.1) and mature leaves (n=10 leaves with a total number of 757 cells, median=10.8). The intrinsic noise was significantly higher in young leaves ($p = 0.029$ Wilcoxon rank-sum test). Boxes show 25th and 75th percentiles and median. White dots show mean values.

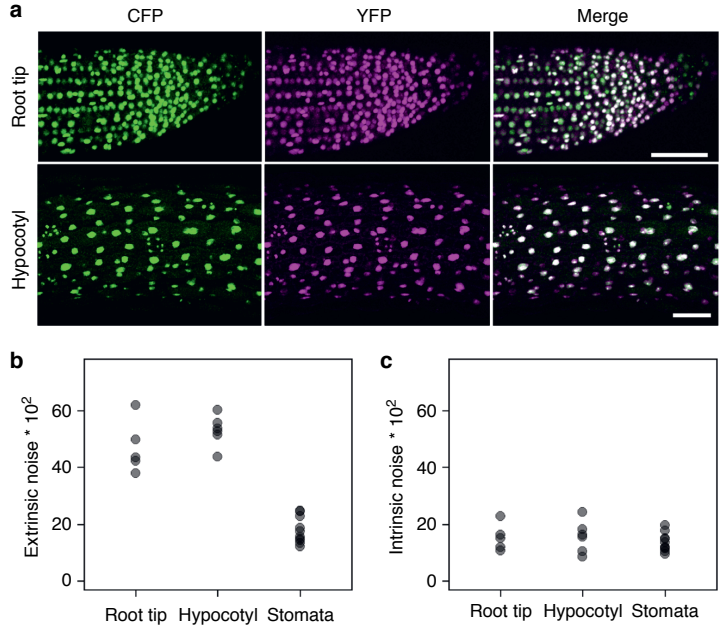


Figure 5.4: Intrinsic and extrinsic noise in stomata cells, root tip cells and hypocotyl cells of *p35S:2xNLS-YFP p35S:2xNLS-CFP* plants. a, CLSM images of a root tip and a hypocotyl of *p35S:2xNLS-YFP p35S:2xNLS-CFP* plants. CFP is green, YFP is magenta and overlay is white. b, Plot of extrinsic noise of root tip cells (n=6 roots with a total number of 463 cells, median=43.4), hypocotyl cells (n=6 hypocotyls with a total number of 690 cells, median=53.1) and stomata cells (n=10 from mature leaves with a total number of 513 cells, median=16.6). c, Plot of intrinsic noise of root tip cells (n=6 roots with a total number of 463 cells, median=15.1), hypocotyl cells (n=6 hypocotyls with a total number of 690 cells, median=15.9) and stomata cells (n=10 from mature leaves with a total number of 513 cells, median=12.8). The extrinsic noise in root tip cells and hypocotyl cells was significantly higher than in stomata cells ($p = 0.00067$ and $p = 0.00025$, Wilcoxon rank-sum test).

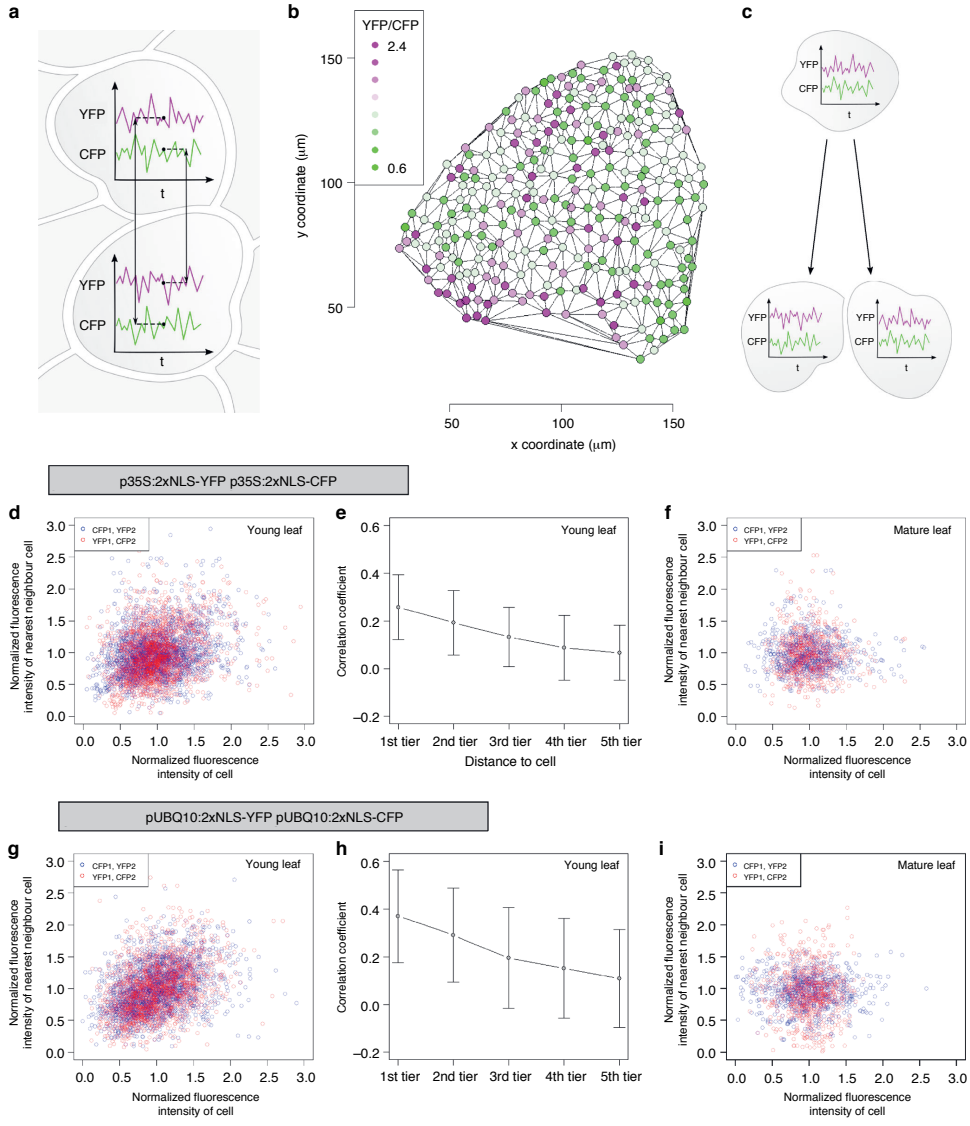


Figure 5.5: Figure caption on next page.

Figure 5.5: Nearest neighbour analysis of *p35S:2xNLS-YFP p35S:2xNLS-CFP* and $pUBQ10 : 2xNLS - YFP$ $pUBQ10 : 2xNLS - CFP$ plants. a, Schematic illustration of the experimental setup to determine the cofluctuation of neighbouring cells. b, Leaf area depicting the cell-to-cell variability of noise based on the YFP/CFP ratios in each cell. Colours show the YFP/CFP ratios as indicated in the legend. c, Schematic illustration of the effect of cell division on cofluctuation. d, Scatter plot of *p35S:2xNLS-YFP p35S:2xNLS-CFP* young leaves showing the normalised fluorescence intensities of cells plotted against the normalised fluorescence intensity of the nearest neighbour of the considered cells (neighbour cell with the lowest distance). Blue circles indicate the CFP fluorescence intensity of a cell (CFP1) plotted against the YFP fluorescence intensity of the nearest neighbouring cell (YFP2). Red circles show the YFP fluorescence intensity of a cell (YFP1) plotted against the CFP fluorescence intensity of the nearest neighbouring cell (CFP2) (n=2219 cells; $r = 0.341$; $p = 0.0002$, randomisation test). e, Mutual dependency of the distance to the neighbouring cell and the cofluctuation in young rosette leaves of *p35S:2xNLS-YFP p35S:2xNLS-CFP*. Neighbouring cells were grouped into five tiers according to their distance (cell diameters) to the considered cell. Mean values and standard deviations are shown (n=86, 541 neighbourhood analyses (2219 cells×39 cells)). f, Scatter plot of *p35S:2xNLS-YFP p35S:2xNLS-CFP* mature rosette leaves showing the normalised fluorescence intensities of cells plotted against the normalised fluorescence intensity of the nearest neighbour (n=757 cells; $r = 0.02$; $p = 0.433$, randomisation test). g, Scatter plot of $pUBQ10 : 2xNLS - YFP$ $pUBQ10 : 2xNLS - CFP$ young rosette leaves showing the normalised fluorescence intensities of cells plotted against the normalised fluorescence intensity of the nearest neighbour (n=2021 cells; $r = 0.413$; $p = 0.0003$, randomisation test). h, Mutual dependency of the distance to the neighbouring cell and the cofluctuation in young rosette leaves of $pUBQ10 : 2xNLS - YFP$ $pUBQ10 : 2xNLS - CFP$. Neighbouring cells were grouped into five tiers according to their distance (cell diameters) to the considered cell. Mean values and standard deviations are shown (n=78, 819 neighbourhood analyses (2021 cells×39 neighbouring cells)). i, Scatter plot of $pUBQ10 : 2xNLS - YFP$ $pUBQ10 : 2xNLS - CFP$ mature rosette leaves showing the normalised fluorescence intensities of cells plotted against the normalised fluorescence intensity of the nearest neighbour (n=775 cells; $r = -0.06$; $p = 0.681$, randomisation test).

Appendices

5.A Non-stationary auto-correlation of the two-stage gene expression model in presence of extrinsic noise

In the experiments we can estimate the auto-correlation of the gene expression in Arabidopsis. To be able to compare the estimation with the result from the two-stage gene expression model, we need to calculate the non-stationary auto-correlation between protein abundance in case of extrinsic noise. At time $t_0 = 0$ KikG is converted from green to red, *i.e.*, for $t > t_0$ one has a red and a green fluorescent protein population. The red protein population decays to zero due to protein degradation and the green population is building up from zero. We therefore consider a stochastic process with initial condition $n(t_0) = 0$. We denote the value at time t_2 conditioned on the value at time t_1 by: $n(t_2)|n(t_1)$ and the expectation value of n at time t_2 conditioned on the value $n(t_1)$ at time t_1 by $\langle n(t_2)|n(t_1)\rangle$. If we denote by $P(n_2, t_2|n_1, t_1)$ the non-stationary conditional probability density this reads explicitly:

$$\langle n(t_2)|n(t_1)\rangle \;=\; \sum_{n_2} n_2 P(n_2, t_2|n_1, t_1)$$

The correlation between $n(t_2)$ and $n(t_1)$ is given by:

$$\langle\langle (n(t_2)|n(t_1))(n(t_1)|n(t_0))\rangle\rangle \;=\; \sum_{n_2}\sum_{n_1} n_2 n_1 P(n_2, t_2|n_1, t_1) P(n_1, t_1|n_0, t_0)$$

$\langle\langle\cdot\rangle\rangle$ means averaging over all possible n_2 and over all possible n_1. We consider the two-stage protein expression model shown in Figure 1A of the main text. If the protein degradation rate d_1 is much lower than the mRNA degradation rate d_0 the

conditional probability reads in zero-order approximation in $\gamma = d_0/d_1$:[231]

$$P(n_2, t_2|n_1, t_1) = \sum_{m=0}^{n_1} \frac{n_1! f(t_2 - t_1)^{n_1 - m}(1 - f(t_2 - t_1))^m}{(n_1 - m)! m!} P_0(n_2 - m, t_2 - t_1) + \mathcal{O}(\gamma^{-1}).$$

The function P_0 denotes the probability distribution for $n_1 = 0$ initial proteins given by:

$$P_0(n, t) = \frac{(a)_n}{n!} \left(\frac{b}{1+b}\right)^n \left(\frac{1 + ne^{-d_1 t}}{1+b}\right)^a \sum_{k=0}^{n} \frac{(-n)_k (-a)_k}{(1 - a - n)_k k!} \left(\frac{(1+b)e^{-d_1 t}}{1 + be^{-d_1 t}}\right)^k + \mathcal{O}(\gamma^{-1}),$$

where $(a)_n$ is the Pochhammer symbol,[232] the parameters and the function f are defined as:

$$a = \frac{\nu_0}{d_1}, \quad b = \frac{\nu_1}{d_0}, \quad f(t) = 1 - e^{-d_1 t}. \tag{5.4}$$

In general the parameters will slightly vary from cell to cell and also over time (time-dependent extrinsic noise). For simplicity, we consider cell-to-cell variability in the transcription and translation rates ν_0 and ν_1 and keep the degradation rates constant (note that in the simulations we only vary the translation rate ν_1). This means the transcription and translation rates vary from cell to cell, but are otherwise constant. To calculate the auto-correlation, one has to average over the extrinsic noise:

$$c(t_2, t_1; t_0) \equiv \frac{\langle\langle\langle (n(t_2)|n(t_1))(n(t_1)|n(t_0))\rangle\rangle\rangle_e - \langle\langle n(t_2)|n(t_0)\rangle\rangle_e \langle\langle n(t_1)|n(t_0)\rangle\rangle_e}{\sqrt{\left[\langle\langle (n(t_2)|n(t_0))^2\rangle\rangle_e - \langle\langle n(t_2)|n(t_0)\rangle\rangle_e^2\right]\left[\langle\langle (n(t_1)|n(t_0))^2\rangle\rangle_e - \langle\langle n(t_1)|n(t_0)\rangle\rangle_e^2\right]}}$$

Note that the above expression depends on the initial time t_0. Using the above stated approximations we find after some algebra for the auto-correlation:

$$c(t_2, t_1; t_0) = \frac{V^2 f_{20} f_{10} + \langle ab\rangle_e f_{10}(1 - f_{21}) + \langle ab^2\rangle_e g_{10}(1 - f_{21})}{\sqrt{\left[V^2 f_{20}^2 + \langle ab\rangle_e f_{20} + \langle ab^2\rangle_e g_{20}\right]\left[V^2 f_{10}^2 + \langle ab\rangle_e f_{10} + \langle ab^2\rangle_e g_{10}\right]}} + \mathcal{O}(\gamma^{-1}) \tag{5.5}$$

with

$$\begin{aligned} V^2 &= \langle a^2 b^2 \rangle_e - \langle ab \rangle_e^2 \\ f_{pq} &= 1 - e^{-d_1(t_p - t_q)}, \quad g_{pq} = 1 - e^{-2d_1(t_p - t_q)} \end{aligned}$$

It is instructive to consider limiting cases:

- No extrinsic noise, $V^2 = 0$:

$$\lim_{V^2 \to 0} c(t_2, t_1; t_0) = \frac{f_{10}(1 - f_{21}) + (1 - f_{21}) b g_{10}}{\sqrt{[f_{20} + b g_{20}]\,[f_{10} + b g_{10}]}} + \mathcal{O}(\gamma^{-1})$$

- The limit $a \to \infty$, $b \to 0$ with ab finite corresponds to the one-stage gene expression model, simple production and degradation of a protein:

$$\lim_{\substack{ab=\text{const} \\ a \to \infty \\ b \to 0}} c(t_2, t_1; t_0) = \frac{V^2 f_{20} f_{10} + \langle ab \rangle_e f_{10}(1 - f_{21})}{\sqrt{\left[V^2 f_{20}^2 + \langle ab \rangle_e f_{20}\right]\left[V^2 f_{10}^2 + \langle ab \rangle_e f_{10}\right]}}$$

  This is the non-stationary autocorrelation for a death-birth process with extrinsic noise. It is exact, the correction of order $\mathcal{O}(\gamma^{-1})$ vanishes. To see why this is the case, write $a = \nu_0 / d_0 \gamma$ and $b = \nu_1 / d_1 \gamma^{-1}$ and take the limit $\gamma \to \infty$.

- Taking also the limit $V^2 \to 0$ yields:

$$\lim_{\substack{ab=\text{const} \\ a \to \infty \\ b \to 0}} \lim_{V^2 \to 0} c(t_2, t_1; t_0) = c_0(t_2, t_1; t_0) = \sqrt{\frac{f_{10}}{f_{20}}}(1 - f_{21})$$

  which is the non-stationary autocorrelation for a death-birth process with constant parameters. For $t_1 - t_0 \gg d_1^{-1}$ $(t_2 \geq t_1)$ we obtain:

$$c_0(t_2, t_1) = (1 - f_{21}) = e^{-d_1(t_2 - t_1)}.$$

Note that c_0 only depends on the protein degradation rate.

5.A.1 The birth-death process as a lower bound

In the following we prove that the auto-correlation of the birth-death process provides a lower bound for the auto-correlation of the two-stage model with extrinsic noise.

We start by rewriting Eq. (5.5):

$$\begin{aligned} c(t_2,t_1;t_0) &= c_0(t_2,t_1;t_0)\Gamma(t_2,t_1;t_0) \\ \Gamma(t_2,t_1;t_0) &= \frac{\frac{V^2}{(1-f_{21})\langle ab\rangle_e}f_{20}+1+\frac{\langle ab^2\rangle_e}{\langle ab\rangle_e}\frac{g_{10}}{f_{10}}}{\sqrt{\left[\frac{V^2}{\langle ab\rangle_e}f_{20}+1+\frac{\langle ab^2\rangle_e}{\langle ab\rangle_e}\frac{g_{20}}{f_{20}}\right]\left[\frac{V^2}{\langle ab\rangle_e}f_{10}+1+\frac{\langle ab^2\rangle_e}{\langle ab\rangle_e}\frac{g_{10}}{f_{10}}\right]}} \end{aligned}$$

It is sufficient to show that $\Gamma \geq 1$ for all possible parameter combinations. Because $1-f_{21} \leq 1$ we have:

$$\Gamma(t_2,t_1;t_0) \geq \frac{\frac{V^2}{\langle ab\rangle_e}f_{20}+1+\frac{\langle ab^2\rangle_e}{\langle ab\rangle_e}\frac{g_{10}}{f_{10}}}{\sqrt{\left[\frac{V^2}{\langle ab\rangle_e}f_{20}+1+\frac{\langle ab^2\rangle_e}{\langle ab\rangle_e}\frac{g_{20}}{f_{20}}\right]\left[\frac{V^2}{\langle ab\rangle_e}f_{10}+1+\frac{\langle ab^2\rangle_e}{\langle ab\rangle_e}\frac{g_{10}}{f_{10}}\right]}}$$

Further, because $f_{20} \geq f_{10}$ ($t_2 \geq t_1$) we find:

$$\Gamma(t_2,t_1;t_0) \geq \frac{\frac{V^2}{\langle ab\rangle_e}f_{20}+1+\frac{\langle ab^2\rangle_e}{\langle ab\rangle_e}\frac{g_{10}}{f_{10}}}{\sqrt{\left[\frac{V^2}{\langle ab\rangle_e}f_{20}+1+\frac{\langle ab^2\rangle_e}{\langle ab\rangle_e}\frac{g_{20}}{f_{20}}\right]\left[\frac{V^2}{\langle ab\rangle_e}f_{20}+1+\frac{\langle ab^2\rangle_e}{\langle ab\rangle_e}\frac{g_{10}}{f_{10}}\right]}}$$

We have

$$\begin{aligned} \frac{g_{20}}{f_{20}} &\leq \frac{g_{10}}{f_{10}} \Leftrightarrow g_{20}f_{10} \leq g_{10}f_{20} \Leftrightarrow (1-(1-f_{20})^2)f_{10} \leq (1-(1-f_{10})^2)f_{20} \\ \Leftrightarrow f_{10}f_{20}^2 &\geq f_{10}^2 f_{20} \Leftrightarrow f_{20} \geq f_{10} \end{aligned}$$

and with this

$$\Rightarrow \Gamma(t_2,t_1;t_0) \geq \frac{\frac{V^2}{\langle ab\rangle_e}f_{20}+1+\frac{\langle ab^2\rangle_e}{\langle ab\rangle_e}\frac{g_{10}}{f_{10}}}{\sqrt{\left[\frac{V^2}{\langle ab\rangle_e}f_{20}+1+\frac{\langle ab^2\rangle_e}{\langle ab\rangle_e}\frac{g_{10}}{f_{10}}\right]\left[\frac{V^2}{\langle ab\rangle_e}f_{20}+1+\frac{\langle ab^2\rangle_e}{\langle ab\rangle_e}\frac{g_{10}}{f_{10}}\right]}} = 1.$$

It follows that $1 \geq c(t_2,t_1;t_0) \geq c_0(t_2,t_1;t_0)$.

5.A.2 Estimation of the protein half-life

In order to obtain an estimate for c_0 we estimated the protein degradation rate by measuring the fluorescent of the remaining red fluorescence proteins after conversion. Note that we were only interested in obtaining a rough estimate of the lifetime of the protein. For simplicity we ignored bleaching effects; by this we may overestimate the degradation rate, by means of which our estimate is an upper bound of the true rate. To obtain the estimate we processed the data as follows. For each cell we

have three measurements: x_0 the fluorescence directly after the conversion ($t_0 = 0$), x_1 ($t_1 = 3\,\mathrm{h}$) and x_2 ($t_2 = 6\,\mathrm{h}$) after the conversion. The amount of red proteins at time $t \geq 0$ is given by $x(t) = x_0 e^{-d_1 t}$. We transformed the data logarithmically and calculated for each cell an optimal degradation value by minimising:

$$L = \left(\ln \left[\frac{x_1}{x_0} \right] + d_1 t_1 \right)^2 + \left(\ln \left[\frac{x_2}{x_0} \right] + d_1 t_2 \right)^2 \tag{5.6}$$

with respect to d_1. Because the logarithm is a strictly monotonic function this transformation preserves the optimum in absence of noise and measurements errors, but simplifies its calculation considerably. As long as the noise is small the optima (with and without transformation) are very similar. We incur for the optimal d_1:

$$d_1 = -\frac{t_1 \ln \left[\frac{x_1}{x_0} \right] + t_2 \ln \left[\frac{x_2}{x_0} \right]}{t_1^2 + t_2^2}.$$

We find for d_1: $(d_1)_{0.5} = 0.095\,\mathrm{h}^{-1}$ (median), $\bar{d_1} = 0.091\,\mathrm{h}^{-1}$ (mean), and $\sigma = 0.023\,\mathrm{h}^{-1}$ (SD). Using $d_1 = 0.09\,\mathrm{h}^{-1}$ together with $t_2 = 6$ h, $t_1 = 3$ h and $t_0 = 0$ h in the expression for the auto-correlation we find $c_0 \approx 0.6$. Thus the theoretical prediction for the auto-correlation is $1 \geq c \geq 0.6$.

5.A.3 Simulation of the KikGR experiment

In order to test our analytical results we simulated the KikGR experiment using the stochastic simulation algorithm (SSA).[55] For sake of simplicity we kept all parameters constant besides the translational rate ν_1. We simulated 10^5 trajectories of a single reporter according to:

1. Draw the translation rate ν_1 from a log-normal distribution with $\langle \nu_1 \rangle = 45\,\mathrm{h}^{-1}$ and $\mathrm{Var}(\nu_1) \in \{0\,\mathrm{h}^{-2}, 10^2\,\mathrm{h}^{-2}\}$.

2. Simulate the reporter until stationary state with zero protein initial condition and mRNA initial condition drawn from the Poisson distribution

$$P(m) = \frac{\left(\frac{\nu_0}{d_0}\right)^m}{m!} e^{-\frac{\nu_0}{d_0}}.$$

3. Compute the non-stationary auto-correlation.

In Fig. 5.2b we show example trajectories before and after the converting light pulse. After the bleach at time $t_0 = 0$ the red proteins decay and the green proteins build up until being in stationary state. In Fig. 5.2c, we compare c_0 to the simulation results for different CVs of the protein translation rate ν_1. In Supplementary

Fig. 5.24A we compare the zero-order approximation for the auto-correlation given in Eq. (5.5) with simulation results for finite γ. We also show c_0 as the lower bound of the auto-correlation. The dependency of the non-stationary auto-correlation on t_1 can be seen in Supplementary Fig. 5.24B ($\gamma = 12.5$) and Supplementary Fig. 5.24C ($\gamma = 1250$), respectively.

5.B Spatial correlation

5.B.1 Diffusion-like transport

Movement of molecules between cells can induce a spatial correlation. To see this, we briefly investigate a simplified system. We consider a protein which synthesized with rate σ_i in cell i, is degraded with rate λ and is transported across cell boundaries with rate μ. We assume that μ and λ are homogeneous and constant over the grid. The synthesis, however, are stochastic variables. For sake of simplicity, we consider them to be time independent, *i.e.*, the synthesis rates σ_i vary from cell to cell, but are otherwise constant. The concentration of the protein on the tissue can be described by:

$$\frac{dx}{dt} = (\mu C - \lambda \mathbb{1})x + b$$

where x is a vector of the concentrations of the protein on the cellular grid, C is the connectivity matrix of the grid and b is a vector with the production rates σ_i. The connectivity matrix is given by $C_{ii} = -|\Omega_i|$ and $C_{ij} = 1$ for $i \neq j$ if cell i is connected to cell j and $C_{ij} = 0$ otherwise. $|\Omega_i|$ denotes the number of connected neighbors of cell i. For low mobility the contribution to the correlation due protein mobility will be proportional to $\varepsilon = \mu/\lambda$. This can be seen as follows. In steady state the solution for x can be written formally:

$$x = (\mathbb{1} - \varepsilon C)^{-1}\tilde{b}$$

with $\tilde{b} = \lambda^{-1} b$. In case $\varepsilon \ll 1$ we can expand x in terms of ε:

$$x = \sum_{n=0}^{N} \varepsilon^n C^n \tilde{b},$$

(N is the number of cells on the grid) from which follows:

$$x_i \quad = \tilde{b}_i + \varepsilon \sum_{j}^{N} C_{ij} \tilde{b}_j + \mathcal{O}(\varepsilon^2).$$

For the covariance between two cells we find:

$$\mathrm{cov}(x_i, x_j) \quad = \quad \mathrm{cov}(\tilde{b}_i, \tilde{b}_j) + \varepsilon \left[2V^2 + \sum_{n \neq j}^{N} C_{in} \mathrm{cov}(\tilde{b}_j, \tilde{b}_n) + \sum_{m \neq i}^{N} C_{jm} \mathrm{cov}(\tilde{b}_i, \tilde{b}_m) \right] + \mathcal{O}(\varepsilon^2).$$

In case of no spatial correlation between the synthesis rates all $\mathrm{cov}(\tilde{b}_n, \tilde{b}_m)$ vanish; however, the covariance between two cells would be non-zero and proportional to ε ($V^2 = \langle b_i^2 \rangle - \langle b_i \rangle^2$, $\forall i \in \{1, ..., N\}$). Since we are interested in the covariance between the expression systems of two neighbouring cells (*i.e.*, $\mathrm{cov}(\tilde{b}_n, \tilde{b}_m)$), it is important to find an estimate for ε. To obtain this we consider the case in which we have only one cell producing the protein (the donor cell), all other cells are receptor cells, *i.e.*, $\tilde{b}_1 = \sigma/\lambda$, $\tilde{b}_i = 0$ for $i \neq 1$. In this case we have: $x_1 = \tilde{b}_1 - \varepsilon |\Omega_1| \tilde{b}_1$ and $x_i = \varepsilon \tilde{b}_1$, from which follows:

$$\frac{x_i}{x_1} = \varepsilon + \mathcal{O}(\varepsilon^2).$$

By building the ratio of the measured fluorescence of a receptor cell to the measured fluorescence of the donor cell, *i.e.*, by determining the fraction of fluorescence obtained by the receptor cell, one can find an approximation for ε.

5.B.2 Characterization of localization and movement of YFP tags

We aimed to choose CFP and YFP tags that minimize intercellular movement and that localize completely to the nucleus to enable accurate measurements. We compared three tags: YFP, 2xNLS-YFP and NLS-2xYFP that was published to be cell-autonomous. Arabidopsis leaves were transiently transformed by biolistic transformations. Free YFP was localized in the cytoplasm and the nucleus (Supplementary Fig. 5.1A). As compared to NLS-2xYFP, 2xNLS-YFP showed a stronger nuclear localization, while NLS-2xYFP was still prominently localized in the cytoplasm (Supplementary Fig. 5.1B,C). Unexpectedly, all YFP variants tested showed the ability to move to adjacent cells. To obtain an estimate for ε we have to build the ratio between the donor and the receptor cells. A precise value for ε would

be difficult to obtain, but for our purpose it is sufficient to obtain an upper bound. This simplifies the experimental analysis considerably. For each donor cell we only determine the ratio between the receptor cell with the highest fluorescence and the donor cell. This systematically overestimate the rate, but provides a sufficient upper bound. Free YFP showed the highest movement ($\varepsilon = 0.08\ (+0.019/-0.022)$, basic bootstrap confidence intervalls). 2xNLS-YFP ($\varepsilon = 0.0218\ (+0.0061/-0.0066)$) and NLS-2xYFP ($\varepsilon = 0.0047\ (+0.0016/-0.0018)$) both revealed much lower values (Supplementary Fig. 5.1D). We decided to use 2xNLS-YFP for our analysis of the spatial correlation as this fusion exhibits only slightly higher movement rates, but enabled us to precisely measure the fluorescence in individual cells due to its strong nuclear localization.

5.B.3 Pearson correlation between neighbouring cells

The half-life of the used reporter CFP and YFP (24 h)[227] is of the same order as the cell division time of epidermal cells on young leaves (33 h).[226] Due to this the stochastic gene expression system is never in stationary state on young growing leaves. This gives rise to an extra term when calculating the covariance from a single reporter. The daughter cells inherit mRNA and protein from their mother cell, by which the initial conditions are the same (or at least very similar). In order to analyse the consequence of this we write:

$$\mathrm{cov}(x_1, x_2) \;=\; \langle\langle\langle (x_1|N_1, z_1)(x_2|N_2, z_2)\rangle_N\rangle_e - \langle\langle\langle x_1|N_1, z_1\rangle\rangle_N\rangle_e\langle\langle\langle x_2|N_2, z_2\rangle\rangle_N\rangle_e$$

where N_1 and N_2 denote the protein amounts directly after cell division and z_1 and z_2 denote the extrinsic stochastic processes in cell 1 and cell 2, resp. $\langle\cdot\rangle_N$ means averaging over all possible initial conditions. We assume that the intrinsic processes in cell 1 and cell 2 are independent in a mechanistic sense, *i.e.*, they do not influence each other in any way (directly or indirectly through feedbacks to the environment), which implies they are statistically independent for a given realisation of sample extrinsic noise z_1 and z_2. Further, we assume the stochastic processes are the same in all cells, *i.e.*, $\langle\langle\langle x_1|N_1, z_1\rangle\rangle_N\rangle_e \equiv \langle\langle\langle x_2|N_2, z_2\rangle\rangle_N\rangle_e \equiv \langle\langle\langle x|N, z\rangle\rangle_N\rangle_e$. Therefore, we can rewrite the covariance:

$$\mathrm{cov}(x_1, x_2) \;=\; \langle\langle\langle x_1|N_1, z_1\rangle\langle x_2|N_2, z_2\rangle\rangle_N\rangle_e - \langle\langle\langle x|N, z\rangle\rangle_N\rangle_e^2$$

We investigate now the case where the stochastic extrinsic processes z_i and z_j are identical. If the initial conditions N_1 and N_2 are independent we find:

$$\mathrm{cov}(x_1, x_2) = \langle\langle\langle x|N, z\rangle\rangle_N^2\rangle_e - \langle\langle\langle x|N, z\rangle\rangle_N\rangle_e^2 = \sigma^2_{\mathrm{ext}}. \tag{5.7}$$

However, if the initial conditions are identical as it is in case of cell division, one

obtains (again for identical z_1 and z_2):

$$\mathrm{cov}(x_1, x_2) = \langle\langle\langle x|N,z\rangle^2\rangle_N\rangle_e - \langle\langle\langle x|N,z\rangle\rangle_N\rangle_e^2 \neq \sigma^2_{\mathrm{ext}}. \tag{5.8}$$

In a growing tissue there will be always next neighbour cells which are daughter cells and others which are not. Note that in stationary state the expressions given in Eq. (5.7) and (5.8) are identical, because all dependence on the initial conditions is decayed. Since on young developing leaves the gene expression systems are not in stationary state it is better to estimate the spatial correlation of the extrinsic noise by using the dual reporter system correlating CFP from cell 1 to YFP in cell 2:

$$\mathrm{cov}(c_1, y_2) \;=\; \langle\langle\langle c_1|N_1,z_1\rangle\langle y_2|N_2,z_2\rangle\rangle_N\rangle_e - \langle\langle\langle c|N,z\rangle\rangle_N\rangle_e\langle\langle\langle y|N,z\rangle\rangle_N\rangle_e$$

with c_i and y_i denoting the two channels of the dual reporter system of cell i. The initial mRNA and protein abundance of CFP are identical in the two daughter cells and the same is true for YFP, but the initial conditions of CFP are different and independent of YFP and vice versa. We arrive at:

$$\mathrm{cov}(c_1, y_2) \;=\; \langle\langle\langle c_1|N_1,z_1\rangle\rangle_N\langle\langle y_2|N_2,z_2\rangle\rangle_N\rangle_e - \langle\langle\langle c|N,z\rangle\rangle_N\rangle_e\langle\langle\langle y|N,z\rangle\rangle_N\rangle_e \tag{5.9}$$

which yields for identical stochastic extrinsic processes z_1 and z_2 Eq. (5.7). We therefore define Pearson's correlation coefficient between two neighbouring cells:

$$r = \frac{\mathrm{cov}(c_1, y_2)}{\sigma^2_{\mathrm{ext}}}. \tag{5.10}$$

To analyse the experimental data we define for each set of cells from a given leaf: $\hat{c} = \frac{1}{N}\sum_{i=1}^{N} x_i$, $\hat{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, $\hat{\sigma}^2_{\mathrm{ext}} = \frac{1}{N-1}\sum_{i=1}^{N}(c_i - \hat{c})(y_i - \hat{y})$, and

$$r \;=\; \frac{1}{2\hat{\sigma}^2_{\mathrm{ext}}\sum_{i=1}^{N}|\Omega_i|}\sum_{i=1}^{N}\sum_{j\in\Omega_i}(c_i - \hat{c})(y_j - \hat{y}) + (y_i - \hat{y})(c_j - \hat{c}) \tag{5.11}$$

where N are the number of cells on the particular leaf, the neighbours of cell i are given by the set Ω_i and the number of neighbours by $|\Omega_i|$. Note that r is also scale invariant, *i.e.*, a scaling factor between c and y does not change r.

5.C Measuring intrinsic and extrinsic noise

In presence of intrinsic and extrinsic noise the total variance of a stochastic component x can be decomposed:[12, 225] $\sigma^2 = \sigma^2_{\text{int}} + \sigma^2_{\text{ext}}$ with:

$$\begin{aligned}\sigma^2_{\text{int}} &= \langle\langle x^2|z\rangle\rangle_e - \langle\langle x|z\rangle^2\rangle_e \\ \sigma^2_{\text{ext}} &= \langle\langle x|z\rangle^2\rangle_e - \langle\langle x|z\rangle\rangle^2_e\end{aligned}$$

z denotes the extrinsic stochastic process and the outer brackets indicate averages over all states of z (extrinsic noise). Using the dual reporter system (with c and y denoting the reporter abundance) we can write:

$$\begin{aligned}\langle\langle(c-y)^2|z\rangle\rangle_e &= \langle\langle c^2|z\rangle\rangle_e + \langle\langle y^2|z\rangle\rangle_e - 2\langle\langle c|z\rangle\langle y|z\rangle\rangle_e \\ \langle\langle cy|z\rangle\rangle_e - \langle\langle c|z\rangle\rangle_e\langle\langle y|z\rangle\rangle_e &= \langle\langle c|z\rangle\langle y|z\rangle\rangle_e - \langle\langle c|z\rangle\rangle_e\langle\langle y|z\rangle\rangle_e\end{aligned}$$

If the reporters are identical and independent we have:

$$\begin{aligned}\langle\langle(c-y)^2|z\rangle\rangle_e &= 2\langle\langle c^2|z\rangle\rangle_e - 2\langle\langle c|z\rangle^2\rangle_e = 2\sigma^2_{\text{int}} \\ \langle\langle cy|z\rangle\rangle_e - \langle\langle c|z\rangle\rangle_e\langle\langle y|z\rangle\rangle_e &= \langle\langle c|z\rangle^2\rangle_e - \langle\langle c|z\rangle\rangle^2_e = \sigma^2_{\text{ext}}.\end{aligned}$$

Using the definition for the intrinsic and extrinsic noise:

$$\eta^2_{\text{int}} = \frac{\sigma^2_{\text{int}}}{\langle c\rangle^2}, \quad \eta^2_{\text{ext}} = \frac{\sigma^2_{\text{ext}}}{\langle c\rangle^2}$$

we arrive at the expressions:[7]

$$\eta^2_{\text{int}} = \frac{\langle(c-y)^2\rangle}{2\langle c\rangle\langle y\rangle}, \quad \eta^2_{\text{ext}} = \frac{\langle cy\rangle - \langle c\rangle\langle y\rangle}{\langle c\rangle\langle y\rangle}. \tag{5.12}$$

We used $\langle\cdot\rangle$ as a short hand notation for the average over extrinsic and intrinsic stochasticity.

In the experiment one cannot measure the abundance of the reporter directly, but rather quantifies the fluorescence. When marking the region of interest (ROI) one may mark pixels which do not belong to the true ROI or miss pixels which do. If we denote the average intensity of the true ROI with N pixels as c and the average intensity of the extra or missed n pixels as s (note that n can be positive or negative) we can write for the measured signal m:

$$m = p\frac{Nc + ns}{N + n} \tag{5.13}$$

where p is the proportionally factors between fluorescence and abundance. The

average intensity s of the extra or missed pixels is related to the average intensity c of the nucleus. We therefore write $s = c(1+\eta)$ where η is a stochastic number with $|\eta| < 1$, $\langle c\eta\rangle = \langle c\rangle\langle\eta\rangle$ and $\langle\eta\rangle = 0$ ($\Rightarrow \langle s\rangle = \langle c\rangle$). With this we rewrite Eq. (5.13):

$$m = pc\,(1+\varepsilon_m)\,, \tag{5.14}$$

with the definition:

$$\varepsilon_m = \frac{\frac{n}{N}}{1+\frac{n}{N}}\eta. \tag{5.15}$$

It is important to realise that n, the number of erroneously marked or missed pixels and s the average intensity of these pixels are uncorrelated stochastic numbers. Due to this we obtain $\langle\varepsilon_m\rangle = 0$ and $\langle m\rangle = p\langle c\rangle$. Altogether, we can write for the measured CFP and YFP signals:

$$m_c = p_c c(1+\varepsilon_m) + \varepsilon_c \tag{5.16}$$

$$m_y = p_y y(1+\varepsilon_m) + \varepsilon_y. \tag{5.17}$$

ε_c and ε_y denote the technical error due to stochasticity of the equipment (in general different for c and y). Because there is no bias in the experiment $\langle\varepsilon_c\rangle = \langle\varepsilon_y\rangle = 0$ holds. The proportionally factors between fluorescence and abundance p_c and p_y are in general different. This causes a problem, when one estimates the intrinsic noise from data. To see this we ignore the errors in Eqs. (5.16) and (5.17) and calculate the intrinsic noise using Eq. (5.12):

$$\frac{\langle(m_c-m_y)^2\rangle}{2\langle m_c\rangle\langle m_y\rangle} = \frac{p_c}{p_y}\frac{\left\langle\left(c-\frac{p_y}{p_c}y\right)^2\right\rangle}{2\langle c\rangle\langle y\rangle} \neq \frac{\langle(c-y)^2\rangle}{2\langle c\rangle\langle y\rangle}. \tag{5.18}$$

We therefore normalise the data with the mean and again (for the sake of the argument) ignore the errors:

$$\frac{\langle(m_c/\langle m_c\rangle - m_y/\langle m_y\rangle)^2\rangle}{2} = \frac{\langle(c/\langle c\rangle - y/\langle y\rangle)^2\rangle}{2} \overset{\langle c\rangle=\langle y\rangle}{=} \frac{\langle(c-y)^2\rangle}{2\langle c\rangle\langle y\rangle}$$

The last equality holds if $\langle c\rangle = \langle y\rangle$, which is true for the dual reporter system. In presence of technical and measurement errors the estimated (apparent) intrinsic noise is given by:

$$\eta^2_{\mathrm{int,app}} = \frac{\langle(m_c/\langle m_c\rangle - m_y/\langle m_y\rangle)^2\rangle}{2} = \eta^2_{\mathrm{int}}\left(1+\langle\varepsilon_m^2\rangle\right) + \mathcal{E}_t^2$$

with the total technical error:

$$\mathcal{E}_t^2 = \frac{\langle \varepsilon_c^2 \rangle}{2\langle m_c \rangle^2} + \frac{\langle \varepsilon_y^2 \rangle}{2\langle m_y \rangle^2}.$$

For the estimated (apparent) extrinsic noise we find:

$$\eta_{\text{ext,app}}^2 = \frac{\langle m_c m_y \rangle - \langle m_c \rangle \langle m_y \rangle}{\langle m_c \rangle \langle m_y \rangle} = \eta_{\text{ext}}^2 \left(1 + \langle \varepsilon_m^2 \rangle\right). \tag{5.19}$$

In both cases the apparent noise overestimates the true intrinsic and extrinsic noise, respectively. The factors (relative error) by which the intrinsic and extrinsic noise is overestimated are given by:

$$f_{\text{int}} = \frac{\eta_{\text{int,app}}}{\eta_{\text{int}}} - 1 = \sqrt{\frac{1 + \langle \varepsilon_m^2 \rangle}{1 - \frac{\mathcal{E}_t^2}{\eta_{\text{int,app}}^2}}} - 1, \quad f_{\text{ext}} = \frac{\eta_{\text{ext,app}}}{\eta_{\text{ext}}} - 1 = \sqrt{1 + \langle \varepsilon_m^2 \rangle} - 1 \tag{5.20}$$

In order to estimate the technical error generated by the used instruments ($\langle \varepsilon_c^2 \rangle$, $\langle \varepsilon_y^2 \rangle$), a yellow autofluorescent plastic slide (ChromaTechnologies, https://www.chroma.com/products/accessories/92001-autofluorescent-plastic-slides) was imaged with identical CFP and YFP settings and laser intensities as used for noise measurements. Because fluorochromes do not bleach and diffuse in this type of slides, only one region of interest was imaged 221 times to avoid variation in our measurements due to inhomogeneities. Because we do not have any additional errors besides the technical error we can write: $m_c^s = p_c^s s + \varepsilon_c$ and $m_y^s = p_y^s s + \varepsilon_y$, where s is the signal from the yellow autofluorescent plastic slide. We find:

$$\frac{\text{Var}(m_c^s)}{\langle m_c^s \rangle^2} = \frac{\text{Var}(s)}{\langle s \rangle^2} + \frac{\langle \varepsilon_c^2 \rangle}{\langle m_c^s \rangle^2}, \quad \frac{\text{Var}(m_y^s)}{\langle m_y^s \rangle^2} = \frac{\text{Var}(s)}{\langle s \rangle^2} + \frac{\langle \varepsilon_y^2 \rangle}{\langle m_y^s \rangle^2}$$

and

$$\langle (m_c^s / \langle m_c^s \rangle - m_y^s / \langle m_y^s \rangle)^2 \rangle = \frac{\langle \varepsilon_c^2 \rangle}{\langle m_c^s \rangle^2} + \frac{\langle \varepsilon_y^2 \rangle}{\langle m_y^s \rangle^2}.$$

Taken together we obtain estimates for the technical errors:

$$\frac{\text{Var}(s)}{\langle s \rangle^2} = 3.7 \times 10^{-6}, \quad \frac{\langle \varepsilon_c^2 \rangle}{\langle m_c^s \rangle^2} = 1.6 \times 10^{-5}, \quad \frac{\langle \varepsilon_y^2 \rangle}{\langle m_y^s \rangle^2} = 9.6 \times 10^{-7}. \tag{5.21}$$

In contrast to the technical errors coming from the microscopy equipment the measurement error is the same for both channels per analysed cell. To obtain an estimate for ε_m, we performed the following series of experiments. The ROI were

on three leaves marked three times. This means that the technical errors and the CFP and YFP amount are constant for the three subsequent measurements of the nuclei. The estimator for covariance between the measured signal for CFP and YFP reads:

$$\widehat{\mathrm{cov}}(m_c^n, m_y^n) = \frac{1}{2}\sum_{i=1}^{3}\left(m_c^{in} - \frac{1}{3}\sum_{i=1}^{3} m_c^{in}\right)\left(m_y^{in} - \frac{1}{3}\sum_{i=1}^{3} m_y^{in}\right). \tag{5.22}$$

$m_{c/y}^{in}$ are the signals obtained from the i^{th} measurements of the n^{th} nucleus. Using Eqs. (5.16) and (5.17) we find:

$$\widehat{\mathrm{cov}}(m_c^n, m_y^n) = \frac{p_c p_y c_n y_n}{2}\sum_{i=1}^{3}\left(\varepsilon_{m,in} - \frac{1}{3}\sum_{i=1}^{3}\varepsilon_{m,in}\right)\left(\varepsilon_{m,in} - \frac{1}{3}\sum_{i=1}^{3}\varepsilon_{m,in}\right) \tag{5.23}$$

It follows for the average over sufficiently many nuclei:

$$\frac{1}{N}\sum_{n}^{N}\widehat{\mathrm{cov}}(m_c^n, m_y^n) = \left\langle\widehat{\mathrm{cov}}(m_c^n, m_y^n)\right\rangle = p_c p_y \langle cy\rangle\langle\varepsilon_m^2\rangle. \tag{5.24}$$

We also find:

$$\langle m_c m_y\rangle = \frac{1}{3N}\sum_{n=1}^{N}\sum_{i=1}^{3} m_c^{in} m_y^{in} = p_c p_y \left\langle (c(1+\varepsilon_m))(y(1+\varepsilon_m))\right\rangle = p_c p_y \langle cy\rangle(1+\langle\varepsilon_m^2\rangle) \tag{5.25}$$

From Eqs. (5.24) and (5.25) one obtains:

$$\langle\varepsilon_m^2\rangle = \frac{\left\langle\widehat{\mathrm{cov}}(m_c^n, m_y^n)\right\rangle}{\langle m_c m_y\rangle - \left\langle\widehat{\mathrm{cov}}(m_c^n, m_y^n)\right\rangle} \tag{5.26}$$

We estimate $\langle\varepsilon_m^2\rangle$ by averaging over 100 nuclei on 3 different leaves (*i.e.*, 300 nuclei in total):

$$\langle\varepsilon_m^2\rangle = 4.0\times 10^{-3}. \tag{5.27}$$

The measurement noise is two orders of magnitude larger than the technical errors from the microscopy equipment and thus dominates the relative errors made by estimating the intrinsic and extrinsic noise. The error of the extrinsic noise is the same for all measurements while the error of the intrinsic noise depends on the actual measured apparent intrinsic noise (Eq. 5.20). We find $f_{\mathrm{ext}} = 2.0\times 10^{-3}$ and $f_{\mathrm{int}} < 2.1\times 10^{-3}$ for all measured values of the intrinsic noise. We conclude that in both cases the error introduced due to mistakes made by marking of the ROI and fluctuations in the technical equipment are negligible.

5.D Spatial correlation due to cell division

When cells divide the mRNA and protein content of the mother cell is inherited to the daughter cells. When calculating the CFP-YFP cross-correlation of the dual reporter system one avoids to introduce a correlation due to identical initial conditions. However, even though the initial conditions of CFP and YFP of the daughter cells are different they come from the same distribution, the underlying protein distribution of the mother cell. This induces a correlation which will decay with time. To show this we start with Eq. (5.9) and make the dependence on the extrinsic noise process z_0 of the mother cell explicit:

$$\begin{aligned}\mathrm{cov}(c_1, y_2) &= \langle\langle\langle\langle c_1|N_1, z_1, z_0\rangle\rangle_N\langle\langle y_2|N_2, z_2, z_0\rangle\rangle_N\rangle_{z_1}\rangle_{z_2}\rangle_{z_0} \\ &\quad -\langle\langle\langle\langle c|N, z, z_0\rangle\rangle_N\rangle_z\rangle_{z_0}\langle\langle\langle\langle y|N, z, z_0\rangle\rangle_N\rangle_z\rangle_{z_0}\end{aligned}$$

If we assume for simplicity that z_0, z_1 and z_2 are independent we find:

$$\mathrm{cov}(c_1, y_2) = \langle\langle\langle\langle x|N, z\rangle\rangle_N\rangle_z^2\rangle_{z_0} - \langle\langle\langle x|N, z\rangle\rangle_N\rangle_z\rangle_{z_0}^2.$$

At $t = 0$, right after cell division the expression above yields σ^2_{ext}, while for $t \to \infty$ the gene expression of the daughter cells are independent from the stochastic processes in the mother cell, therefore expect we $\mathrm{cov}(c_1, y_2) \to 0$ for large t. It follows that for the correlation (defined by Eq. 5.10) between daughter cells $r(t) \leq 1$ with $r(t = 0) = 1$ and $\lim_{t\to\infty} r(t) = 0$ holds. To determine $r(t)$ we calculate the non-stationary covariance between two daughter cells as well as σ^2_{ext} using the two-stage gene expression model. Because we only wish to estimate the contribution of cell division to the overall spatial correlation, we employ a simple cell division model. At time $t = 0$ an exact copy of the mother cell is produced. Thereby we avoid all complication introduced by cell growth. We obtain:

$$\sigma^2_{\mathrm{ext}}(t) = \left(\langle a^2b^2\rangle - \langle ab\rangle^2\right)\left(1 - 2e^{-d_1 t} + 2e^{-2d_1 t}\right) \tag{5.28}$$

$$\mathrm{cov}(c_1, y_2)(t) = \left(\langle a^2b^2\rangle - \langle ab\rangle^2\right)e^{-2d_1 t} \tag{5.29}$$

$$r(t) = \frac{e^{-2d_1 t}}{(1 - 2e^{-d_1 t} + 2e^{-2d_1 t})}. \tag{5.30}$$

5.D.1 Simulation of the stochastic gene expression in daughter cells

We calculate the correlation between two cells that originate from the same mother cell. The daughter cells inherit all rates from the mother cell, beside the translational rate ν_1. In order to estimate how much of the correlation between cells stems from mRNA and protein inheritance, we assume that the translation rates of the daughter

cells are in general different to each other and also differ from the translational rate of the mother cell. The dual reporter system is simulated as follows:

1. Draw the translation rate ν_1^0 of the mother cell from a log-normal distribution with $\langle \nu_1^0 \rangle = 500\,\mathrm{h}^{-1}$ and $\mathrm{Var}(\nu_1^0) = 1000\,\mathrm{h}^{-2}$.

2. mRNA and protein of CFP and YFP of the mother cell are drawn from the steady state distributions of the two-stage gene expression model.[231] We set $\gamma = 12.5$ and for this the deviations of the analytical protein steady state distribution and the true are small.[231]

3. Draw the translation rates ν_1^1 and ν_1^2 of the daughters cell from a log-normal distribution with $\langle \nu_1^1 \rangle = \langle \nu_1^2 \rangle = \nu_1^0$ and $\mathrm{Var}(\nu_1^1) = \mathrm{Var}(\nu_1^2) = 1000\,\mathrm{h}^{-2}$.

4. Simulate the dual reporter system until stationary state.

This process is repeated 10^6 times. At each interval of $\Delta t = 0.01$, we compute the Pearson correlation according to Eq. (5.11). Typical trajectories of CFP and YFP for mother and daughter cells are shown in Supplementary Fig. 5.25A. After cell division the trajectories for CFP and YFP become different because of the different translational rates (extrinsic noise) and the stochasticity of the gene expression systems itself (intrinsic noise). From the simulation we estimate the temporal correlation between the two daughter cells. The results are shown in Supplementary Fig 5.25B, which are in agreement with the theoretical predictions.

5.D.2 Estimating the contribution of cell division to the measured next neighbour correlation

As shown above, cell division introduces a correlation between neighbouring cells. This means that even though there are no further correlated processes, one may find a correlation due to inheritance of mRNA and protein. It is therefore important to estimate the value of this contribution. When we calculate the correlation between cells and their next neighbours we do not know which cells are daughter cells, but to make progress we reason that for any given cell two cells out of its neighbouring cells are progeny cells from the last two cell division events. Because we also do not know when after the cell division we observe the cells, we argue that all times within the intervals $[0, T]$ for the last cell division and $[T, 2T]$ from the next to last cell division are equally likely, where T is the inverse cell division rate. For young leaves we have $T = 33$ h.[226] We average $r(t)$ over these time intervals:

$$r_1 = \frac{1}{T}\int_0^T r(t)dt, \quad r_2 = \frac{1}{T}\int_T^{2T} r(t)dt$$

and find $r_1 = 0.69$ and $r_2 = 0.11$. We used $d_1 = 0.029$ h^{-1} as the degradation rate for CFP and YFP.[227] On average the cells have five neighbouring cells. Given that two from these five cells are correlated to the center cell through inheritance of mRNA and protein content and the others are uncorrelated we arrive at our final estimate for the contribution of cell division to the measure next-neighbour correlation: $r = (r_1 + r_2)/5 \approx 0.16$.

If the daughter cells inherit not only the mRNA and protein content but also all other relevant features from the mother cell, the three cells have the same rates. In this case the correlation between the cells is always one, given the rates are constant over time. Following the same argument as outlined above one would expect a total next neighbour correlation of $r = 0.4$.

Supplementary Table 1: List of constructs used in this study.

Main figures

	Construct	Tissue	Transformation
Figure 1	*p35S:NLS-KikGR* *pUBQ10:NLS-KikGR*	young leaf	stable
Figure 2	Simulation	-	-
Figure 3	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#1	young leaf mature leaf	stable
Figure 4	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#1	primary root tip hypocotyl stomata (mature leaf)	stable
Figure 5	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#1 *pUBQ10:2xNLS-YFP pUBQ10:2xNLS-CFP*	young leaf mature leaf young leaf mature leaf	stable

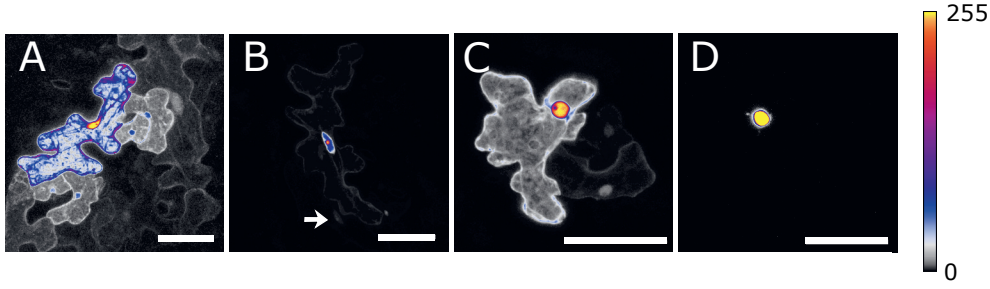
Supplementary figures

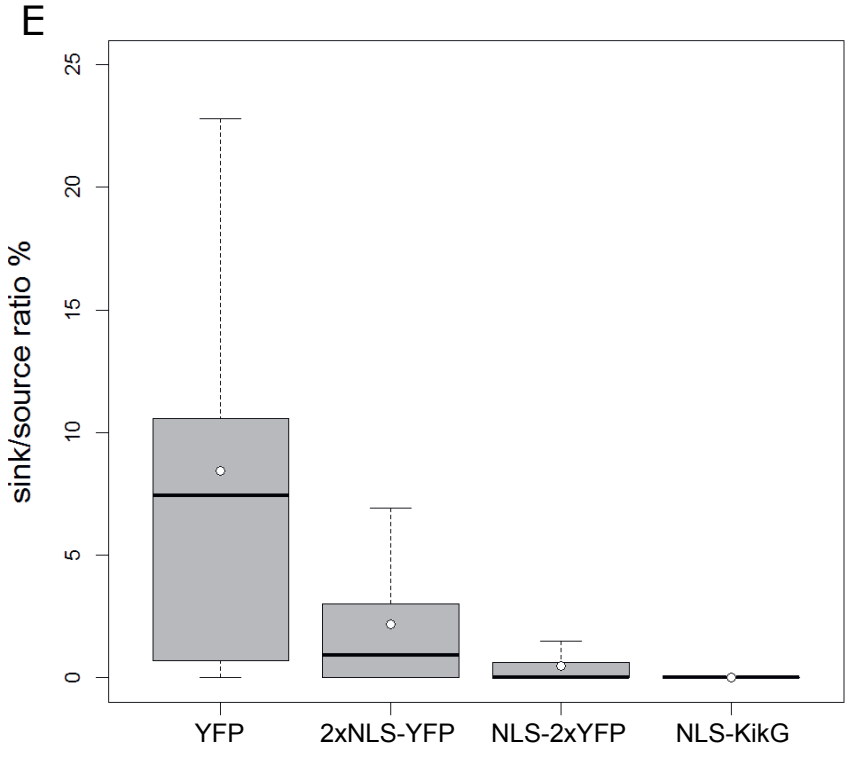
	Construct	Tissue	Transformation
Figure S1	*p35S:YFP* *p35S:2xNLS-YFP* *p35S:NLS-2xYFP* *p35S:NLS-KikGR*	mature leaf	transient
Figure S2	*p35S:NLS-KikGR*	young leaf	stable
Figure S3	*pUBQ10:NLS-KikGR*	young leaf	stable
Figure S4	*p35S:NLS-KikGR*	young leaf	stable
Figure S5	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#1	young leaf	stable
Figure S6	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#1	mature leaf	stable
Figure S7	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#2 *pUBQ10:2xNLS-YFP pUBQ10:2xNLS-CFP*	young leaf mature leaf young leaf mature leaf	stable
Figure S8	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#2	young leaf	stable
Figure S9	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#2	mature leaf	stable
Figure S10	*pUBQ10:2xNLS-YFP pUBQ10:2xNLS-CFP*	young leaf	stable
Figure S11	*pUBQ10:2xNLS-YFP pUBQ10:2xNLS-CFP*	mature leaf	stable
Figure S12	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#2 *p35S:2xNLS-YFP p35S:2xNLS-CFP* line#2 *p35S:2xNLS-YFP p35S:2xNLS-CFP* line#2	primary root tip hypocotyl stomata (mature leaf)	stable
Figure S13	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#1	primary root tip	stable
Figure S14	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#2	primary root tip	stable
Figure S15	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#1	hypocotyl	stable
Figure S16	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#2	hypocotyl	stable
Figure S17	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#1	stomata (mature leaf)	stable
Figure S18	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#2	stomata (mature leaf)	stable
Figure S19	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#1 *pUBQ10:2xNLS-YFP pUBQ10:2xNLS-CFP*	mature leaf	stable
Figure S20	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#2	mature leaf	stable
Figure S21	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#2	young leaf mature leaf	stable
Figure S22	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#1	primary root tip hypocotyl	stable
Figure S23	*p35S:2xNLS-YFP p35S:2xNLS-CFP* line#2	primary root tip hypocotyl	stable
Figure S24	Simulation	-	-

Supplementary Table 2: Acquisition settings for fluorescent proteins.

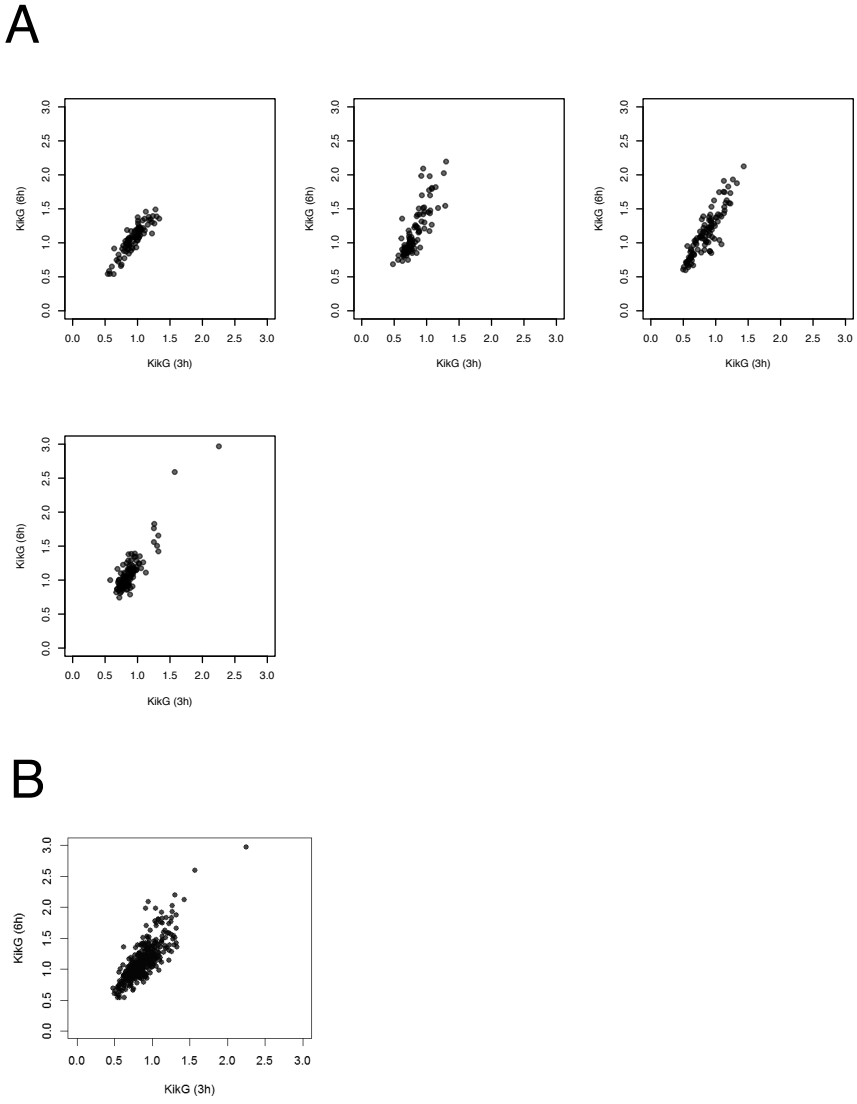
Fluorescent protein	Laser line [nm] Excitation	Detection range [nm] Emission
CFP	405	458-489
YFP	488	526-557
KikG	488	500-520
KikR	561	565-633

Supplementary Figures

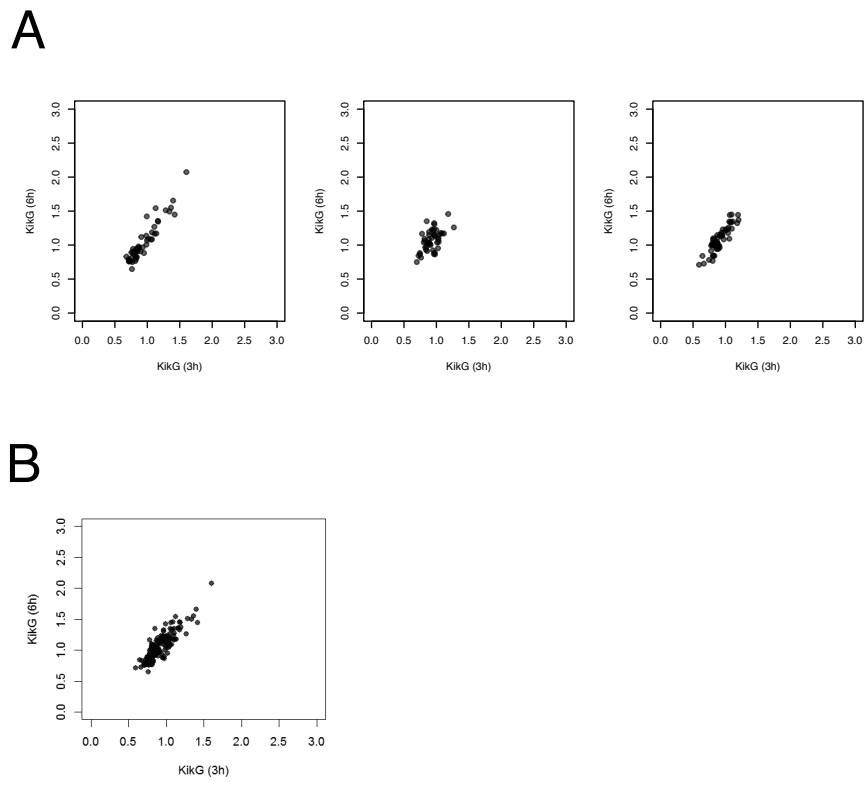




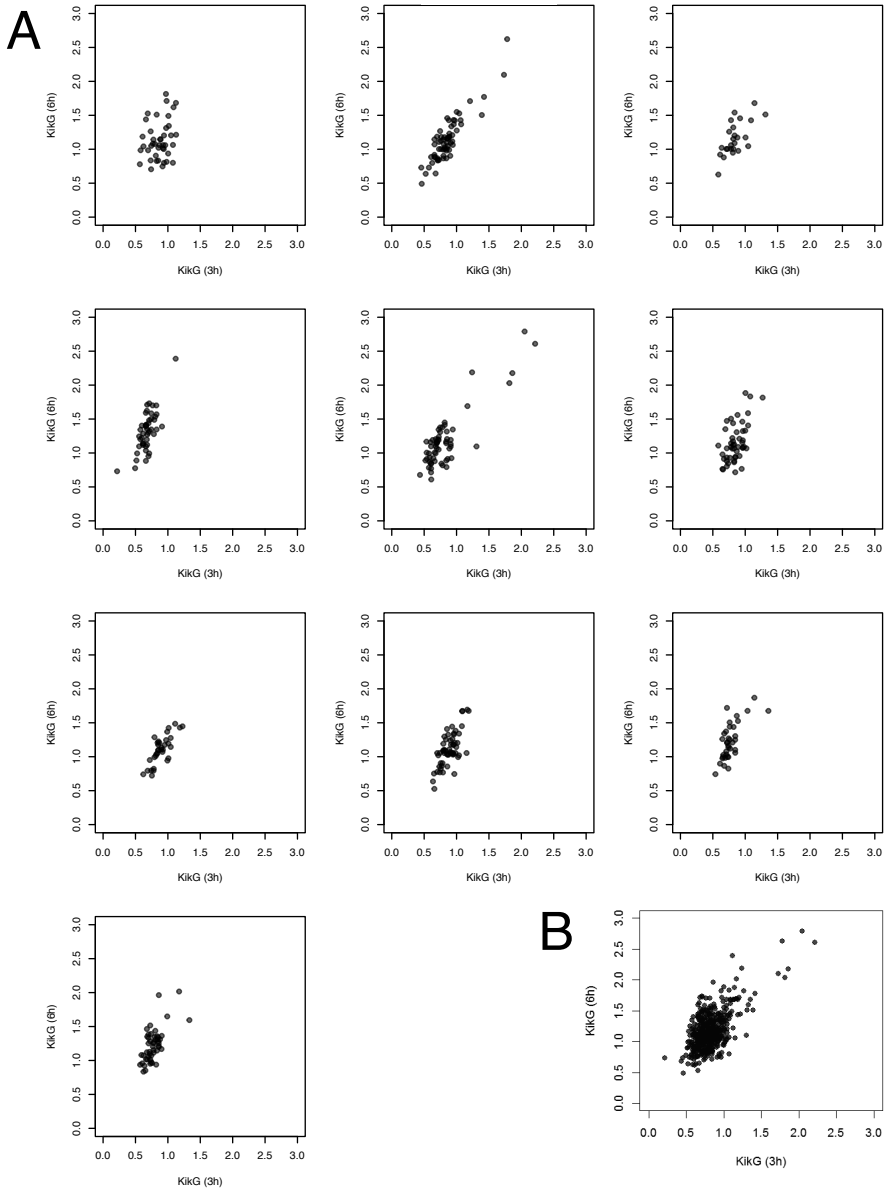
Supplementary Figure 5.1: Comparison of protein localisation and movement of YFP variants. (A-D) Representative CLSM images of transiently transformed single pavement cells in Arabidopsis expressing YFP-tagged and KikG marker proteins and marker movement into neighbouring cells. The colour code indicates the signal intensity in a scale between 0 and 255. (A) Free YFP. (B) 2xNLS-YFP. Arrow indicates fluorescence in the neighbouring cell. (C) NLS-2xYFP. (D) NLS-KikG. Colour code legend indicates grey scale values from 0 to 255 in A-D. (E) Box plot diagram of the ratio between sink and source cells in %. This corresponds to the mobility coefficient ϵ (see supplement). YFP ($n = 48$, $\epsilon = 0.0843$), 2xNLS-YFP ($n = 65$, $\epsilon = 0.0218$), NLS-2xYFP ($n = 61$, $\epsilon = 0.0046$) and NLS-KikG ($n = 20$, $\epsilon = 0.0$). Ratios for 2xNLS-YFP were significantly lower compared to YFP (Kolmogorov Smirnov test $p < 3 \times 10^{-6}$). Ratios for NLS-2xYFP was significantly reduced compared to 2xNLS-YFP and free YFP (Kolmogorov Smirnov test $p < 2 \times 10^{-4}$ and $p < 6 \times 10^{-12}$). Boxes show 25th, 75th quartiles and median. White dots show mean values.



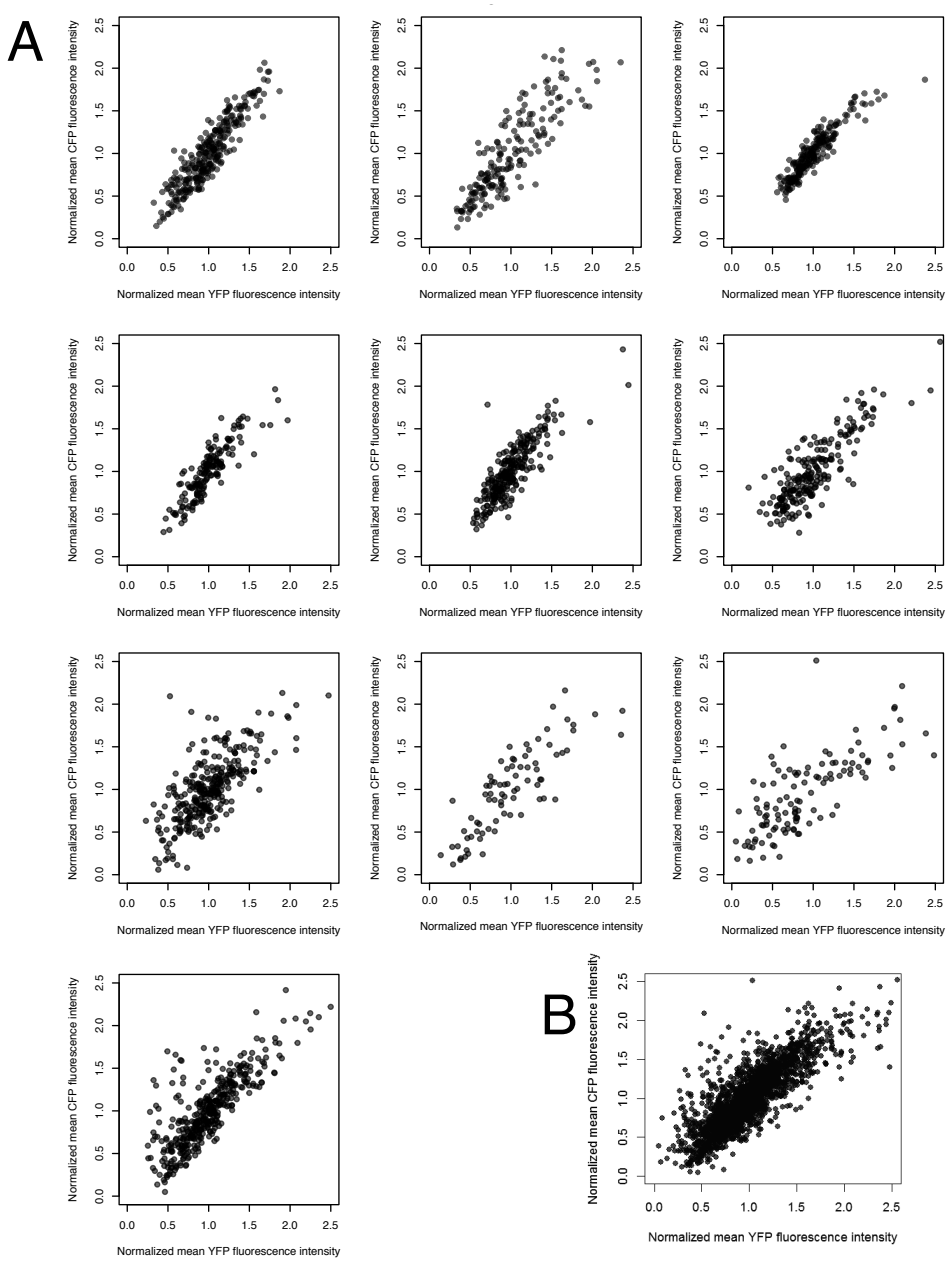
Supplementary Figure 5.2: Scatter plots of *p35S:NLS-KikG* expressing cells. (A) Scatter plot of *p35S:NLS-KikG* expressing cells obtained from four individual leaves. (B) Cumulative scatter plot of all leaves.



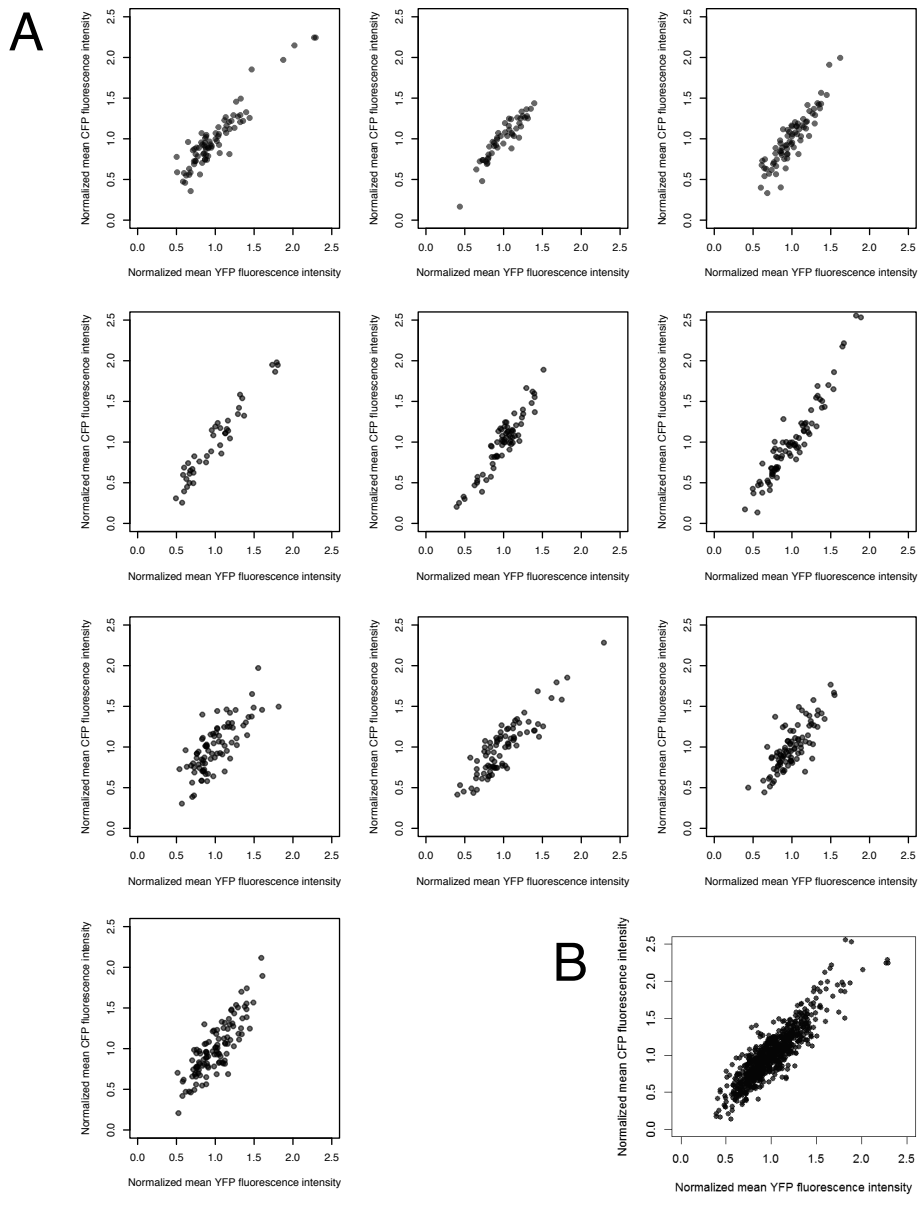
Supplementary Figure 5.3: Scatter plots of $pUBQ10 : NLS - KikG$ expressing cells. (A) Scatter plot of $pUBQ10 : NLS - KikG$ expressing cells obtained from three individual leaves. (B) Cumulative scatter plot of all leaves.



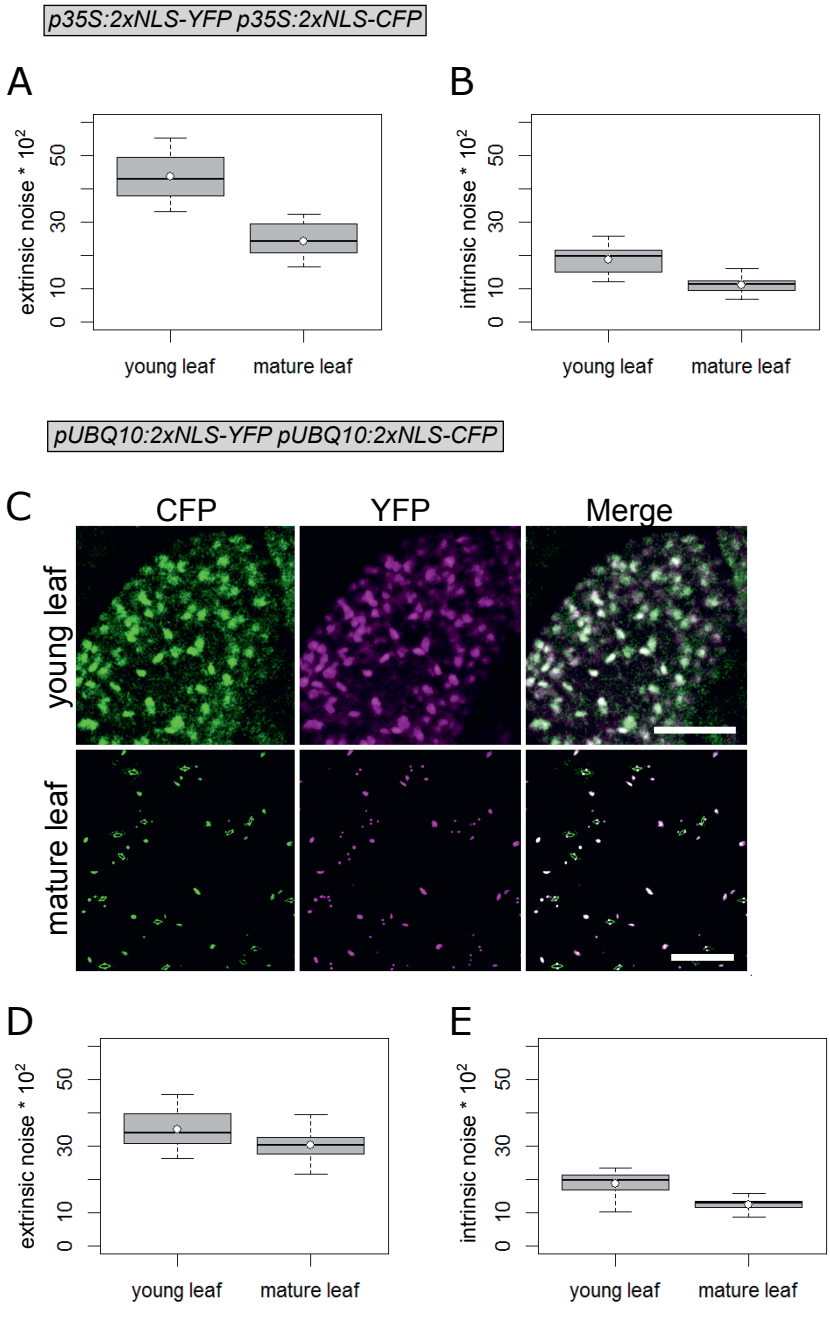
Supplementary Figure 5.4: Scatter plots of *p35S:NLS-KikG* expressing cells without dark treatment. (A) Scatter plot of *p35S:NLS-KikG* expressing cells obtained from ten individual leaves. (B) Cumulative scatter plot of all leaves.


A
B
Normalized mean CFP fluorescence intensity
Normalized mean YFP fluorescence intensity
0.0
0.5
1.0
1.5
2.0
2.5

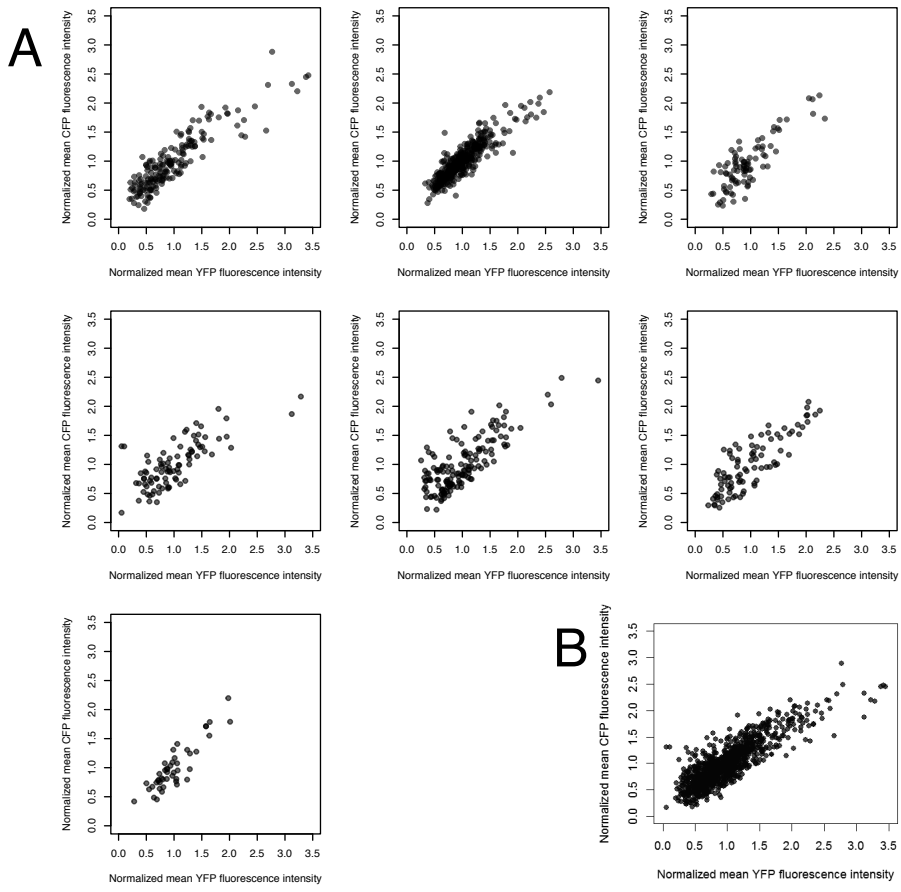
Supplementary Figure 5.5: Scatter plots of the CFP and YFP values in young leaves of *p35S:2xNLS-YFP p35S:2xNLS-CFP* plants (Transgenic line 1). (A) Scatter plots of the CFP and YFP values obtained from ten individual young leaves. A Kolmogorov Smirnov Test was used to test, whether the CFP and YFP value distribution significantly differed. This was not the case in all samples used in this study. (B) Cumulative scatter plot combining all samples.



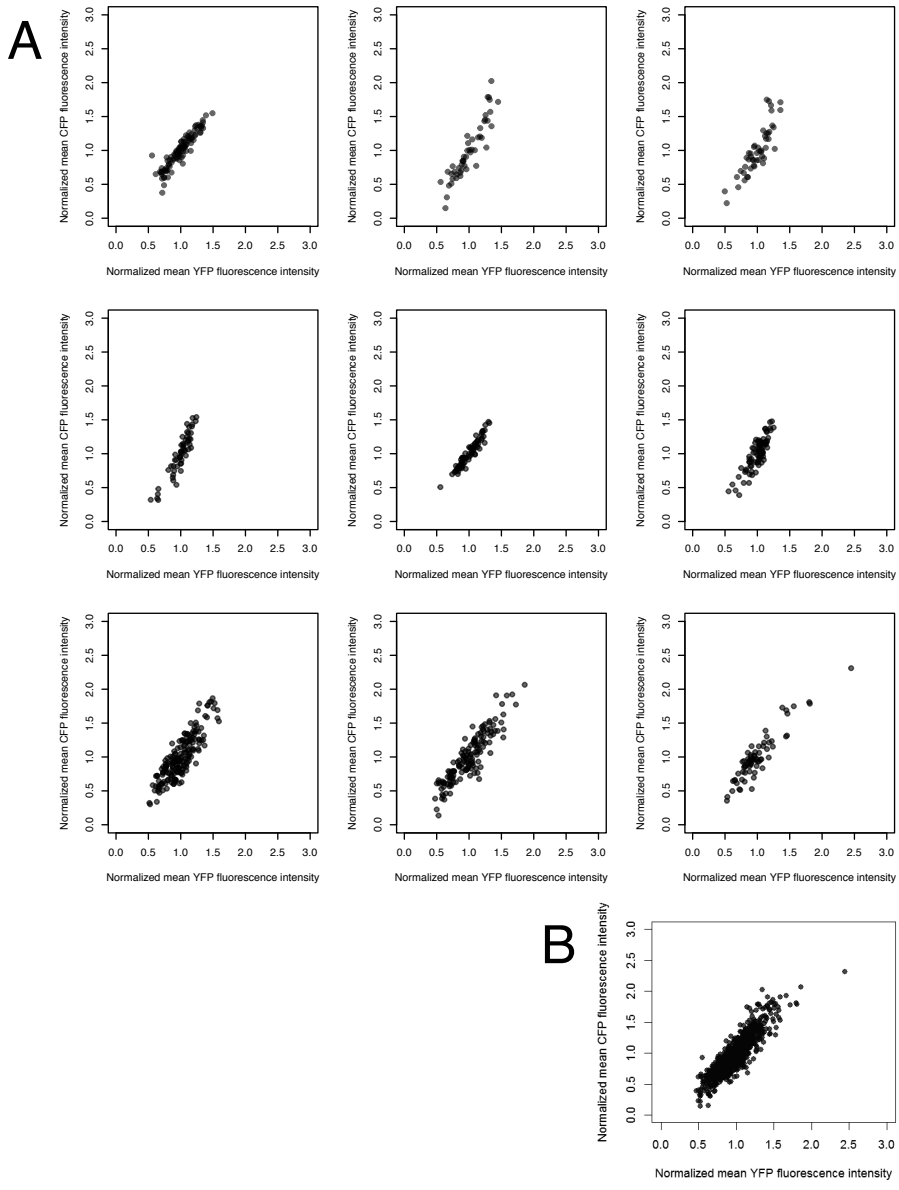
Supplementary Figure 5.6: Scatter plots of the CFP and YFP values in mature leaves of *p35S:2xNLS-YFP p35S:2xNLS-CFP* plants (Transgenic line 1). (A) Scatter plots of the CFP and YFP values obtained from ten individual mature leaves. A Kolmogorov Smirnov Test was used to test, whether the CFP and YFP value distribution significantly differed. This was not the case in all samples used in this study. (B) Cumulative scatter plot combining all samples.


p35S:2xNLS-YFP p35S:2xNLS-CFP
A
extrinsic noise * 10^2
0
10
30
50
young leaf
mature leaf
B
intrinsic noise * 10^2
0
10
30
50
young leaf
mature leaf
pUBQ10:2xNLS-YFP pUBQ10:2xNLS-CFP
C
CFP
YFP
Merge
young leaf
mature leaf
D
extrinsic noise * 10^2
0
10
30
50
young leaf
mature leaf
E
intrinsic noise * 10^2
0
10
30
50
young leaf
mature leaf

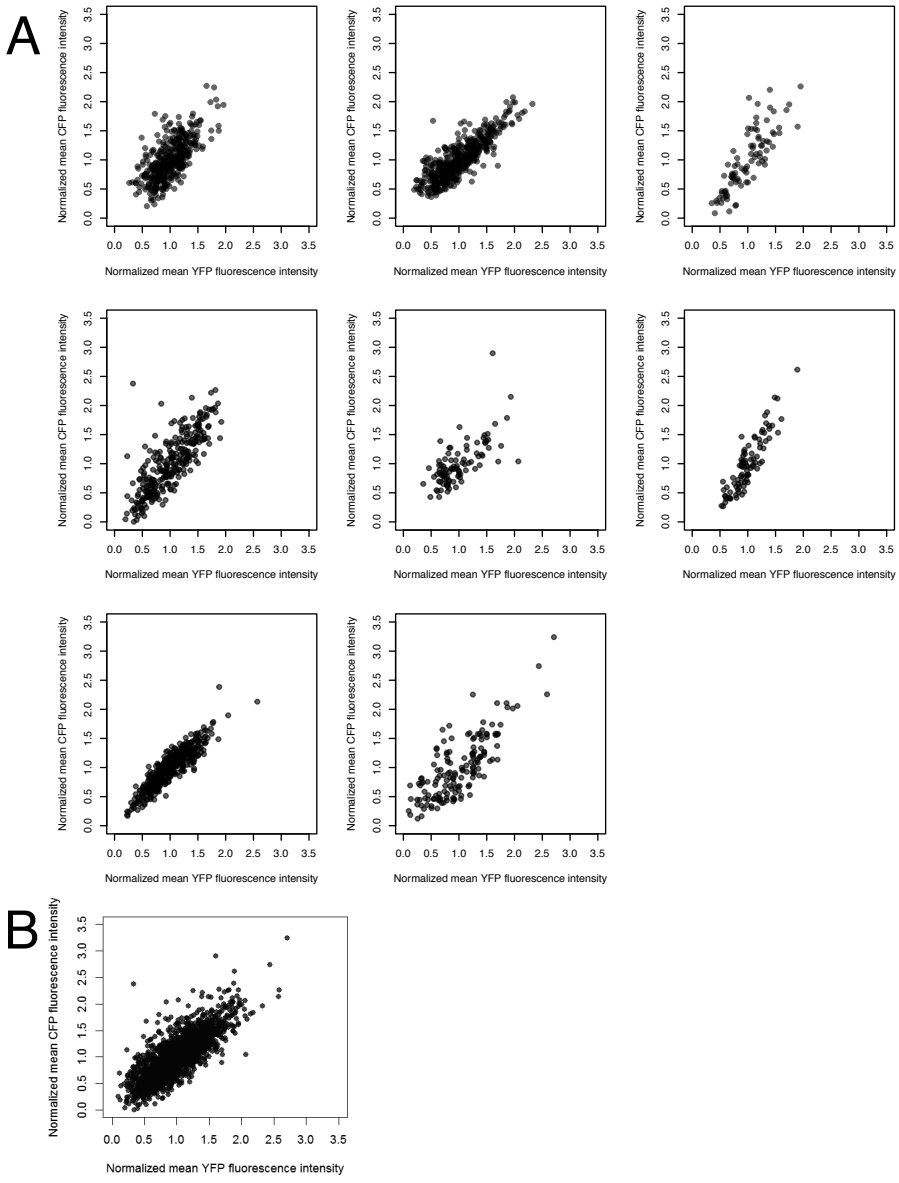
Supplementary Figure 5.7: Intrinsic and extrinsic noise in young and mature rosette leaves of a second independently transformed *p35S:2xNLS-YFP p35S:2xNLS-CFP* line (Transgenic line 2) and $pUBQ10:2xNLS-YFP$ $pUBQ10:2xNLS-CFP$ plants. (A) Box plot of extrinsic noise measurements of young leaves (7 leaves with a total number of 1020 cells, median=42.9) and mature leaves (9 leaves with a total number of 796 cells, median=24.0). The extrinsic noise in young leaves was significantly higher compared to mature leaves ($p = 0.0002$, Wilcoxon rank sum test). Boxes show 25th and 75th quartiles and median. White dots show mean values. (B) Box plot of intrinsic noise of young (7 leaves with a total number of 1020 cells, median=19.7) and mature leaves (9 leaves with a total number of 796 cells, median=11.3). The intrinsic noise in young leaves was significantly higher compared to mature leaves ($p = 0.005$, Wilcoxon rank sum test). The extrinsic noise in mature and young leaves was significantly higher than the intrinsic noise ($p = 4.1 \times 10^{-5}$ and $p = 0.0006$, Wilcoxon rank-sum test). (C) CLSM image of a young and a mature leaf of $pUBQ10:2xNLS-YFP$ $pUBQ10:2xNLS-CFP$. YFP is shown in magenta, CFP in green, overlapping fluorescence intensities are white. Note the auto-fluorescence in the CFP channel of stomata. Scale bar: 30 μm (upper row) and 100μm (lower row). (D) Box plot of extrinsic noise measurements of young (8 leaves with a total number of 2021 cells, median=33.9) and mature leaves (12 leaves with a total number of 775 cells, median=30.3). The extrinsic noise in young leaves was not significantly higher compared to mature leaves (p = 0.082, Wilcoxon rank sum test). (E) Box plot of intrinsic noise measurements of young (8 leaves with a total number of 2120 cells, median=19.6) and mature leaves (12 leaves with a total number of 775 cells, median=12.8). The intrinsic noise in young leaves was significantly higher compared to mature leaves (p = 0.003, Wilcoxon rank sum test). The extrinsic noise of young and mature leaves was significantly higher compared to the intrinsic noise ($p = 7.4 \times 10^{-7}$ and $p = 0.0002$, Wilcoxon signed-rank test).



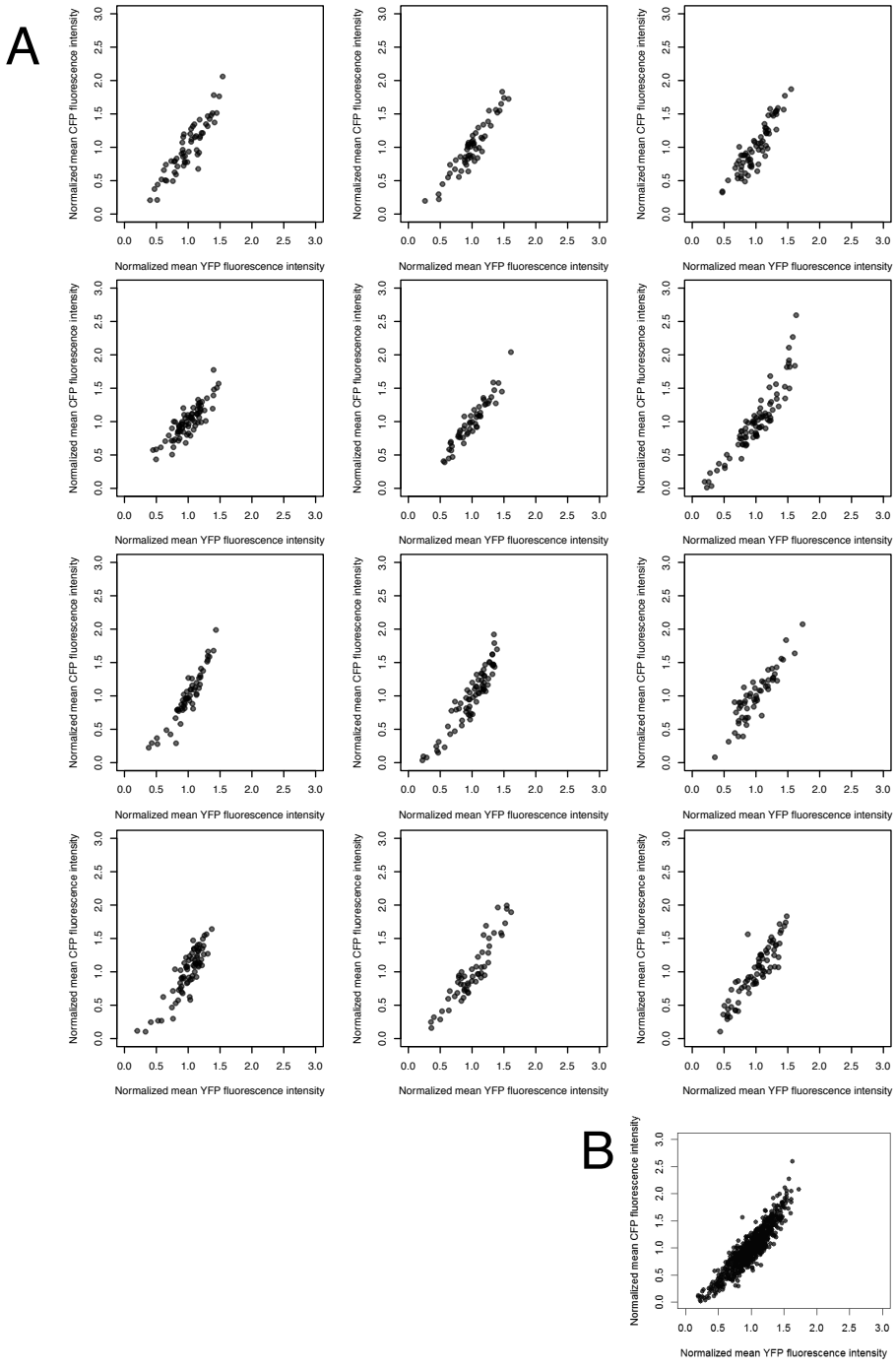
Supplementary Figure 5.8: Scatter plots of the CFP and YFP values in young leaves of *p35S:2xNLS-YFP p35S:2xNLS-CFP* plants (Transgenic line 2). (A) Scatter plots of the CFP and YFP values obtained from seven individual young leaves. A Kolmogorov Smirnov Test was used to test, whether the CFP and YFP value distribution significantly differed. This was not the case in all samples used in this study. (B) Cumulative scatter plot combining all samples.



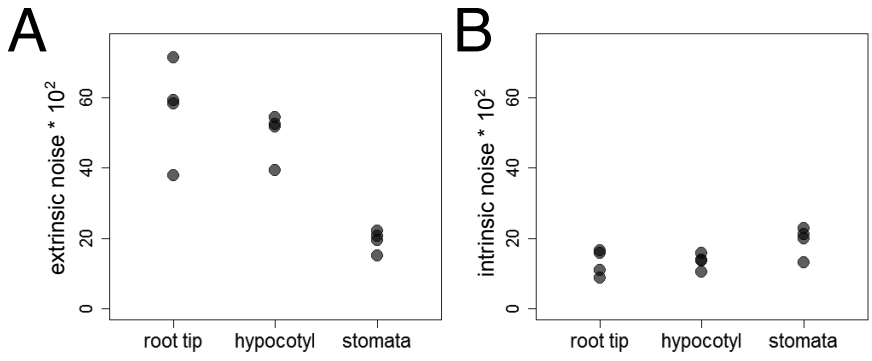
Supplementary Figure 5.9: Scatter plots of the CFP and YFP values in mature leaves of *p35S:2xNLS-YFP p35S:2xNLS-CFP* plants (Transgenic line 2). (A) Scatter plots of the CFP and YFP values obtained from nine individual mature leaves. A Kolmogorov Smirnov Test was used to test, whether the CFP and YFP value distribution significantly differed. This was not the case in all samples used in this study. (B) Cumulative scatter plot combining all samples.



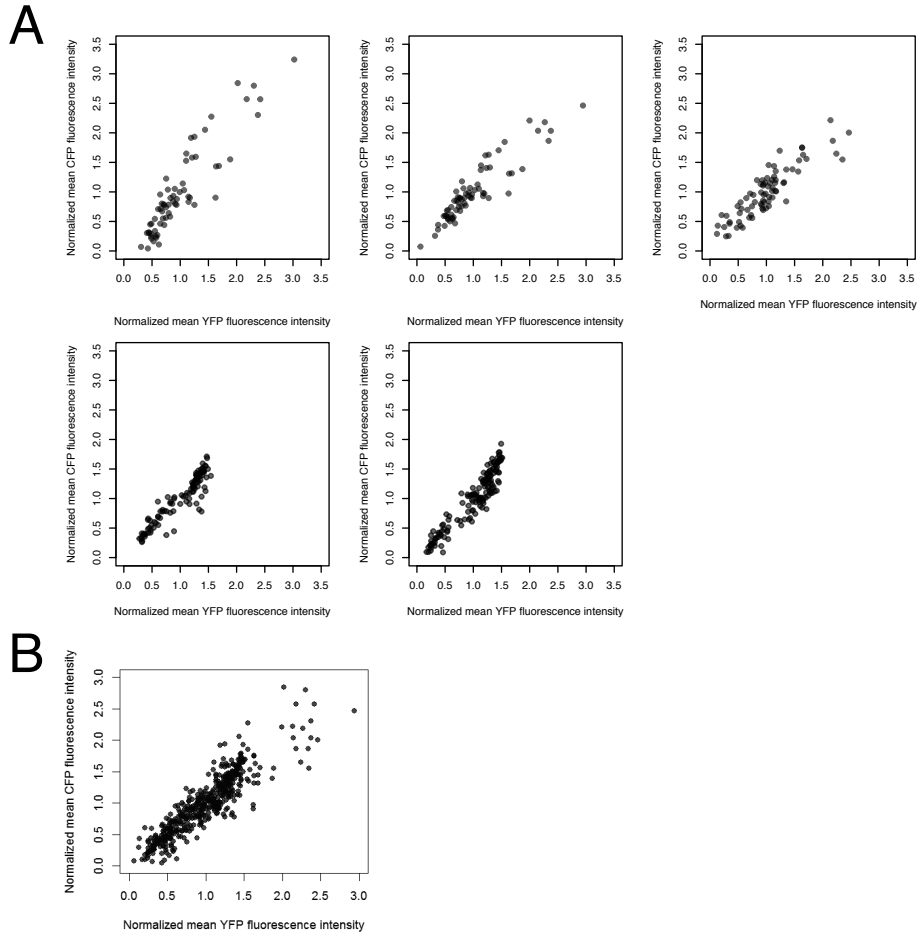
Supplementary Figure 5.10: Scatter plots of the CFP and YFP values in young leaves of $pUBQ10 : 2xNLS - YFP$ $pUBQ10 : 2xNLS - CFP$ plants. (A) Scatter plots of the CFP and YFP values obtained from eight individual young leaves. A Kolmogorov Smirnov Test was used to test, whether the CFP and YFP value distribution significantly differed. This was not the case in all samples used in this study. (B) Cumulative scatter plot combining all samples.


A
B
Normalized mean CFP fluorescence intensity
Normalized mean YFP fluorescence intensity
0.0
0.5
1.0
1.5
2.0
2.5
3.0

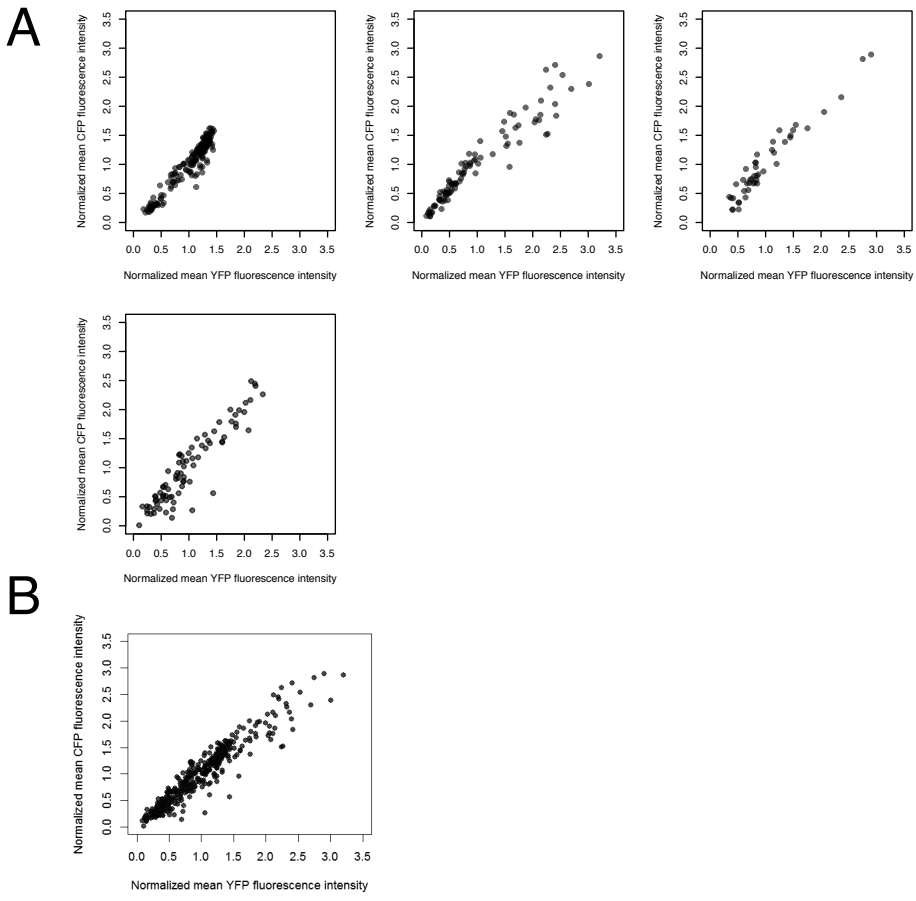
Supplementary Figure 5.11: Scatter plots of the CFP and YFP values in mature leaves of $pUBQ10:2xNLS-YFP$ $pUBQ10:2xNLS-CFP$ plants. (A) Scatter plots of the CFP and YFP values obtained from twelve individual mature leaves. A Kolmogorov Smirnov Test was used to test, whether the CFP and YFP value distribution significantly differed. This was not the case in all samples used in this study. (B) Cumulative scatter plot combining all samples.



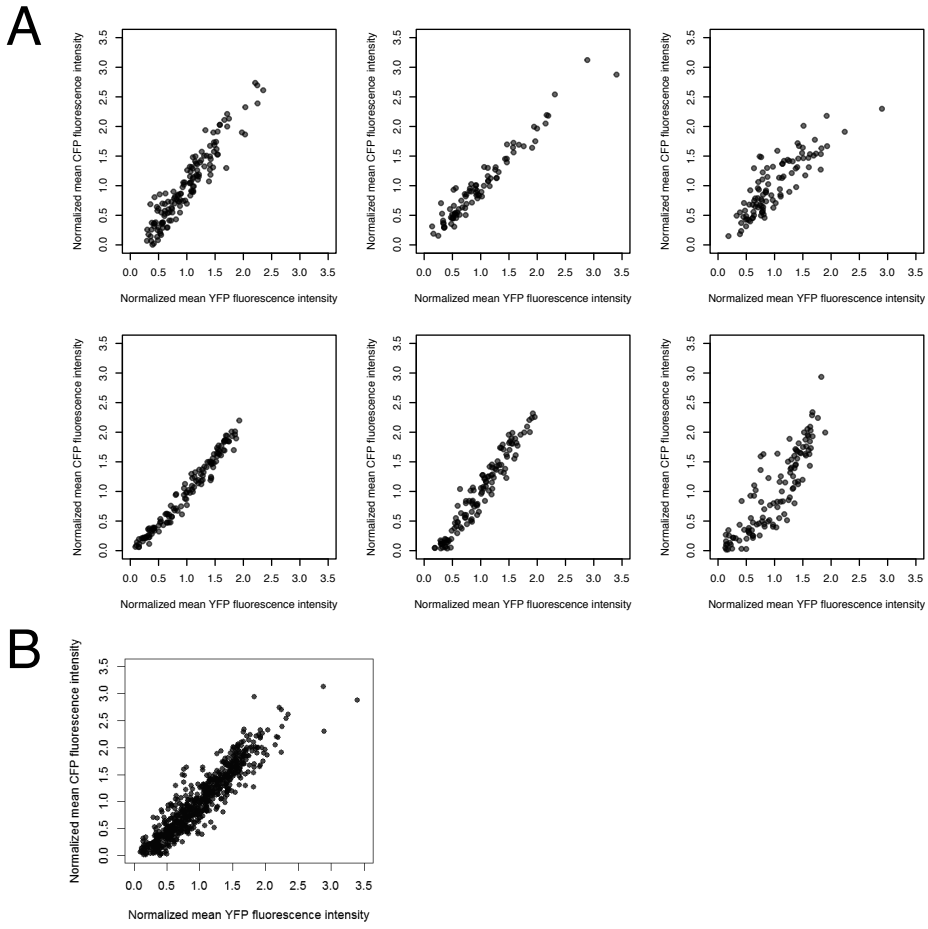
Supplementary Figure 5.12: Extrinsic and intrinsic noise in epidermal stomata cells, root tip cells and hypocotyls cells of a second independently transformed *p35S:2xNLS-YFP p35S:2xNLS-CFP* line (Transgenic line 2). (A) Plot of extrinsic noise of root tip cells (n=4 root tips with a total number of 385 cells, median=58.7), hypocotyls cells (n=4 hypocotyls with a total number of 481 cells, median=52.0) and stomata cells (n=4 mature leaves with a total number of 184 stomata cells, median=20.1). (B) Plot of intrinsic noise of root tip cells (n=4 root tips with a total number of 385 cells, median=13.4), hypocotyls cells (n=4 hypocotyls with a total number of 481 cells, median=13.7) and stomata cells (n=4 mature leaves with a total number of 184 stomata cells, median=20.6). The extrinsic noise in root tip cells and hypocotyls cells was significantly higher as in stomata cells ($p = 0.029$ and $p = 0.029$, Wilcoxon rank-sum test).



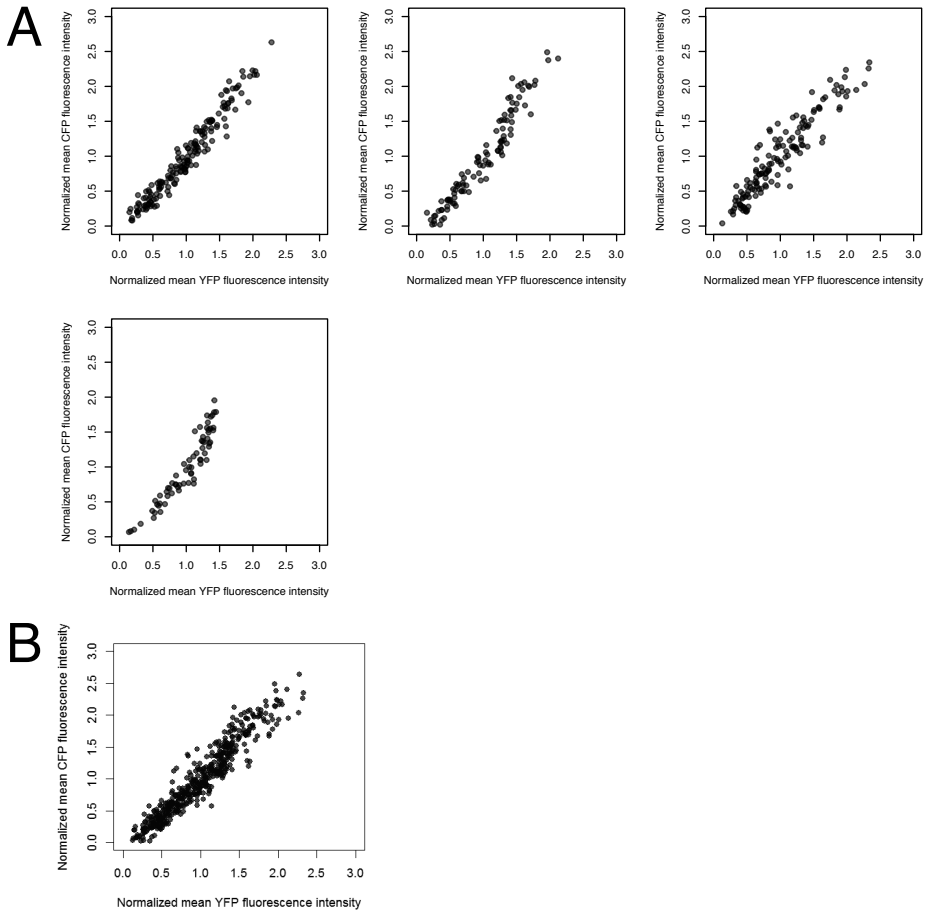
Supplementary Figure 5.13: Scatter plots of the CFP and YFP values in root tips of *p35S:2xNLS-YFP p35S:2xNLS-CFP* plants (Transgenic Line 1). (A) Scatter plots of the CFP and YFP values obtained from five individual root tips. A Kolmogorov Smirnov Test was used to test, whether the CFP and YFP value distribution significantly differed. This was not the case in all samples used in this study. (B) Cumulative scatter plot combining all samples.



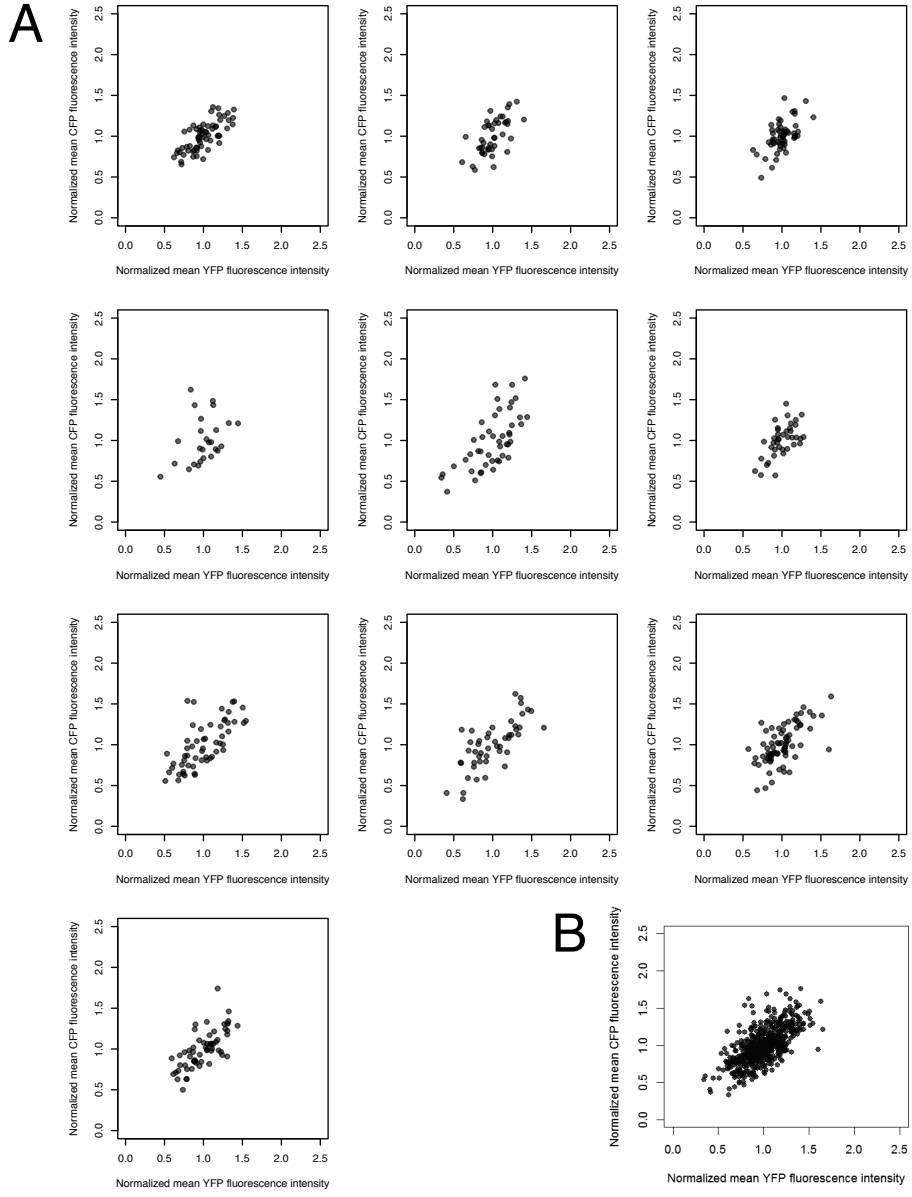
Supplementary Figure 5.14: Scatter plots of the CFP and YFP values in root tips of *p35S:2xNLS-YFP p35S:2xNLS-CFP* plants (Transgenic Line 2). (A) Scatter plots of the CFP and YFP values obtained from four individual root tips. A Kolmogorov Smirnov Test was used to test, whether the CFP and YFP value distribution significantly differed. This was not the case in all samples used in this study. (B) Cumulative scatter plot combining all samples.



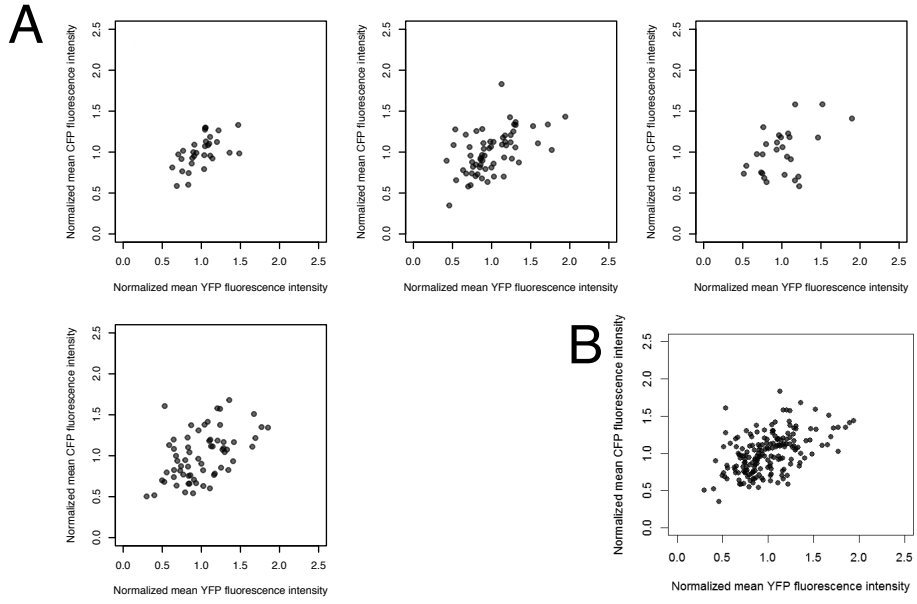
Supplementary Figure 5.15: Scatter plots of the CFP and YFP values in hypocotyl of *p35S:2xNLS-YFP p35S:2xNLS-CFP* plants (Transgenic Line 1). (A) Scatter plots of the CFP and YFP values obtained from six individual hypocotyls. A Kolmogorov Smirnov Test was used to test, whether the CFP and YFP value distribution significantly differed. This was not the case in all samples used in this study. (B) Cumulative scatter plot combining all samples.



Supplementary Figure 5.16: Scatter plots of the CFP and YFP values in hypocotyl cells of *p35S:2xNLS-YFP p35S:2xNLS-CFP* plants (Transgenic Line 2). (A) Scatter plots of the CFP and YFP values obtained from four individual hypocotyls. A Kolmogorov Smirnov Test was used to test, whether the CFP and YFP value distribution significantly differed. This was not the case in all samples used in this study. (B) Cumulative scatter plot combining all samples.

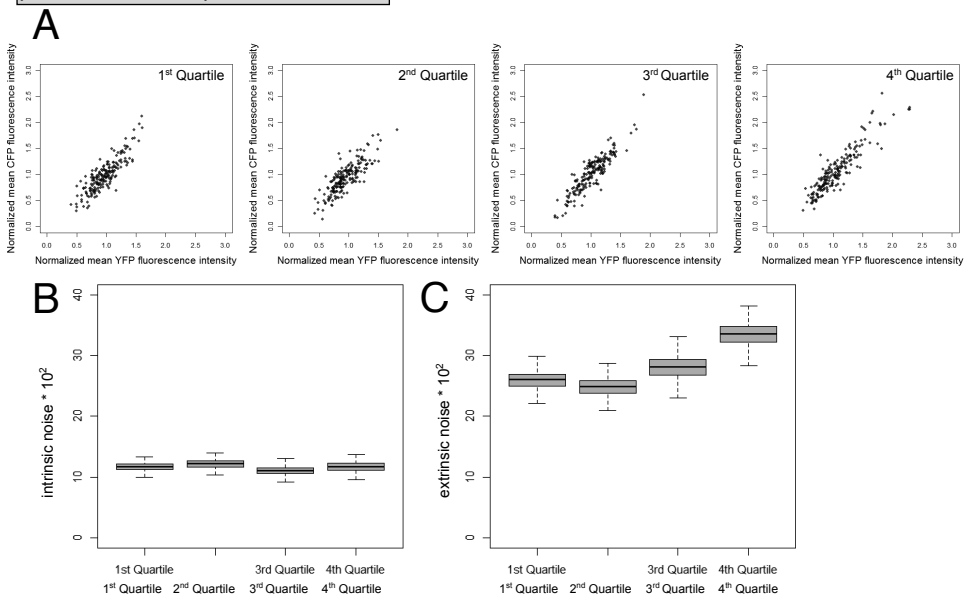


Supplementary Figure 5.17: Scatter plots of the CFP and YFP values in stomata of *p35S:2xNLS-YFP p35S:2xNLS-CFP* plants (Transgenic Line 1). (A) Scatter plots of the CFP and YFP values obtained from stomata on ten individual mature leaves. A Kolmogorov Smirnov Test was used to test, whether the CFP and YFP value distribution significantly differed. This was not the case in all samples used in this study. (B) Cumulative scatter plot combining all samples.

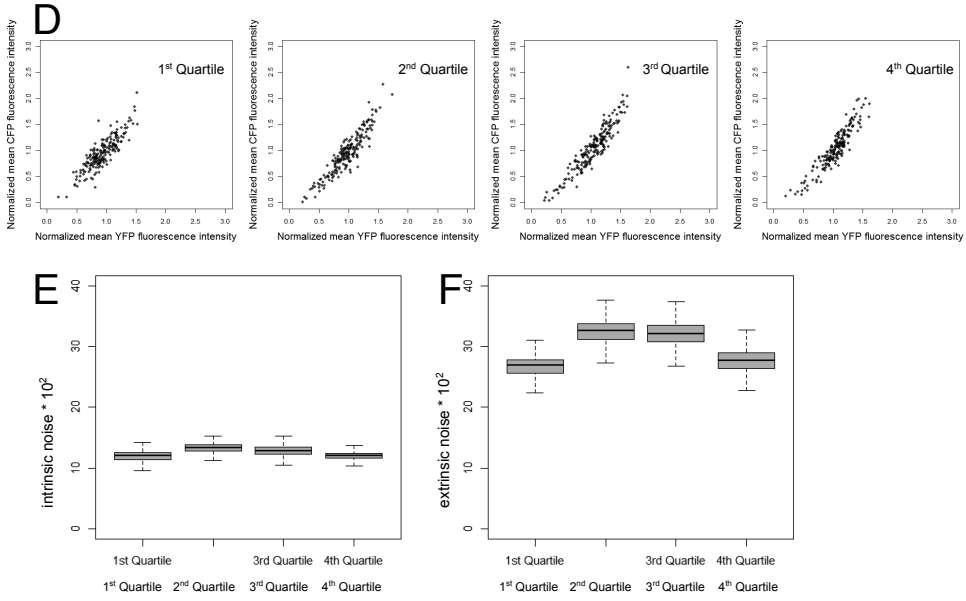


Supplementary Figure 5.18: Scatter plots of the CFP and YFP values in stomata of *p35S:2xNLS-YFP p35S:2xNLS-CFP* plants (Transgenic Line 2). (A) Scatter plots of the CFP and YFP values obtained from stomata on four individual mature leaves. A Kolmogorov Smirnov Test was used to test, whether the CFP and YFP value distribution significantly differed. This was not the case in all samples used in this study. (B) Cumulative scatter plot combining all samples.

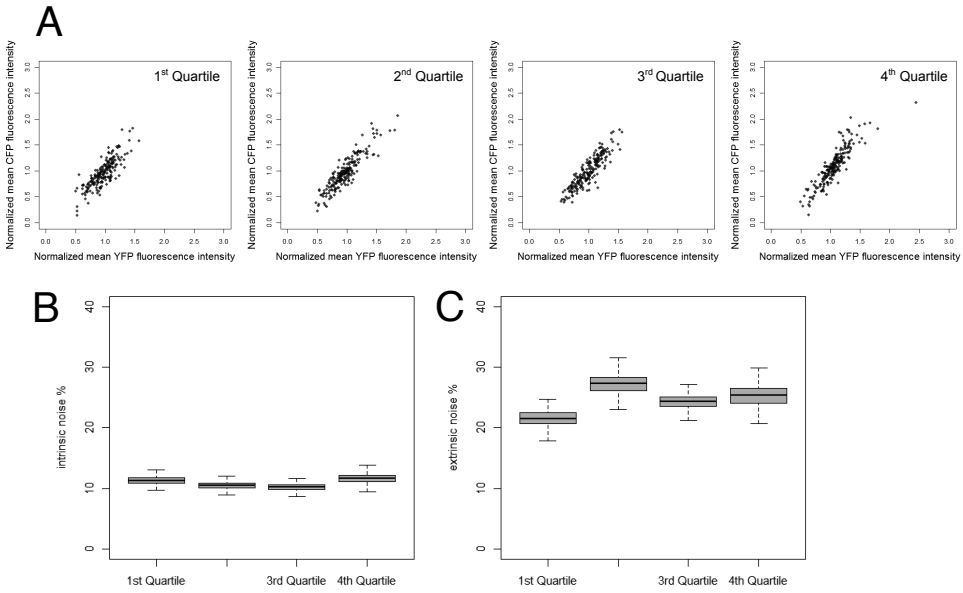
p35S:2xNLS-YFP p35S:2xNLS-CFP



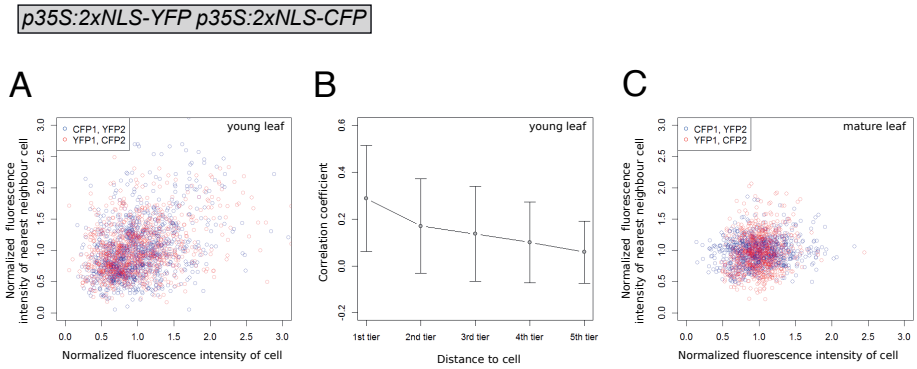
pUBQ10:2xNLS-YFP pUBQ10:2xNLS-CFP



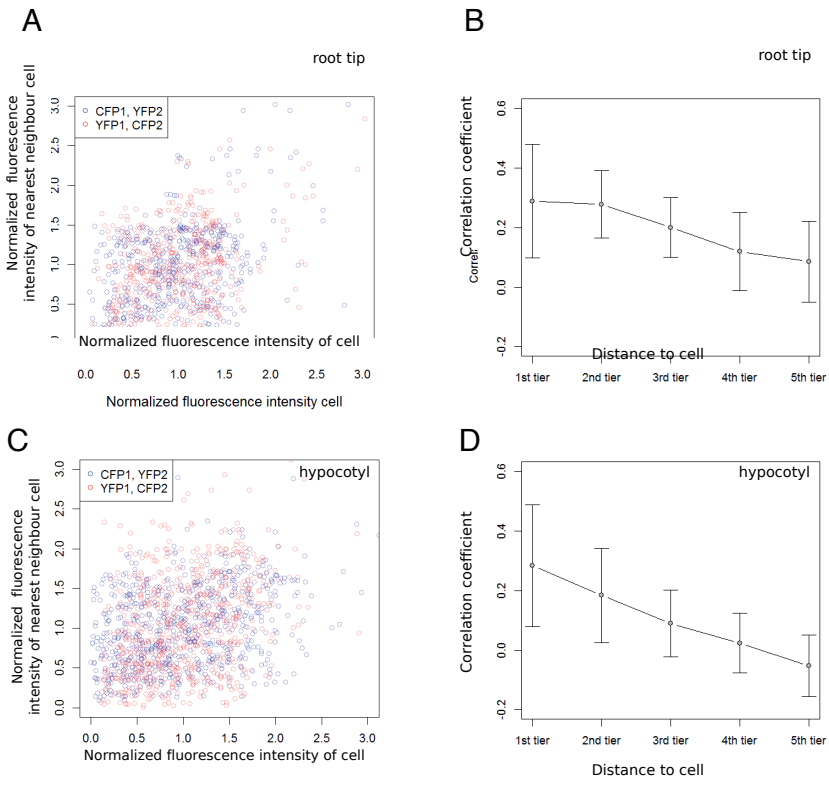
Supplementary Figure 5.19: Analysis of noise with respect to the nucleus size of pavement cells of *p35S:2xNLS-YFP p35S:2xNLS-CFP* and $pUBQ10 : 2xNLS-YFP\ pUBQ10 : 2xNLS-CFP$ plants. (A) Scatter plots of normalised mean CFP and YFP intensities of *p35S:2xNLS-YFP p35S:2x NLS-CFP* nuclei. Cells were separated into quartiles depended on their nucleus size and scatter plots are shown for each quartile. (B) Box plot analysis of intrinsic noise in *p35S:2xNLS-YFP p35S:2xNLS-CFP* plants (n=757×1000 cells) in four nuclear size quartiles. Intrinsic noise was similar in all four quartiles. (C) Box plot analysis of extrinsic noise in *p35S:2xNLS-YFP p35S:2xNLS-CFP* plants (n=757×1000) in four nuclear size quartiles. Extrinsic noise was higher in larger nuclei (3rd and 4th quartile) as compared to the 1st quartile (Bootstrap analysis; $p < 0.001$, Wilcoxon rank-sum test). (D) Scatter plot of normalized mean CFP and YFP intensities of $pUBQ10 : 2xNLS-YFP\ pUBQ10 : 2xNLS-CFP$ nuclei for each quartile. (E) Box plot analysis of intrinsic noise of $pUBQ10 : 2xNLS-YFP$ $pUBQ10 : 2xNLS-CFP$ plants (n=775 " 1000 cells) in four nuclear size quartiles. Intrinsic noise was similar in all four quartiles. (F) Box plot analysis of extrinsic noise of $pUBQ10 : 2xNLS-YFP\ pUBQ10 : 2xNLS-CFP$ plants (n=775×1000 cells) in four nuclear size quartiles. Extrinsic noise was higher in larger nuclei (2nd, 3rd and 4th quartile) as compared to the 1st quartile (Bootstrap analysis; $p < 0.001$, Wilcoxon rank-sum test).



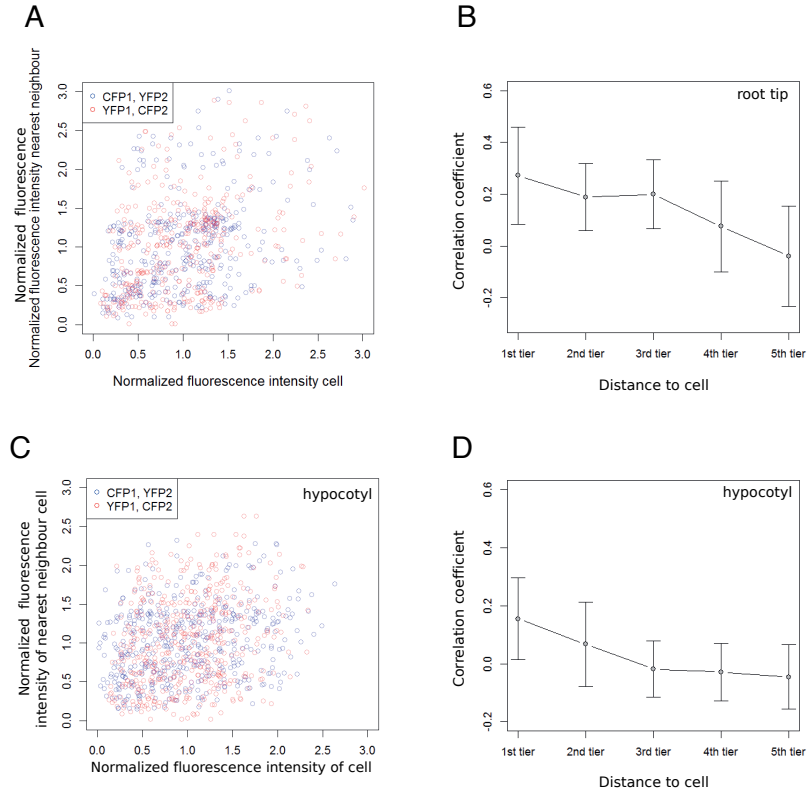
Supplementary Figure 5.20: Analysis of noise with respect to the nucleus size of pavement cells of a second independently transformed *p35S:2xNLS-YFP p35S:2xNLS-CFP* line (Transgenic line 2). (A) Scatter plot of normalized mean CFP and YFP intensities of *p35S:2xNLS-YFP p35S:2xNLS-CFP* nuclei for each quartile. (B) Box plot analysis of intrinsic noise of *p35S:2xNLS- YFP p35S:2xNLS-CFP* plants (n=796×1000 cells) in four nuclear size quartiles. Intrinsic noise was similar in all four quartiles. (C) Box plot analysis of extrinsic noise of *p35S:2xNLS-YFP p35S:2xNLS-CFP* plants (n=796×1000 cells) in four nuclear size quartiles. Extrinsic noise was higher in larger nuclei (2nd, 3rd and 4th quartile) as compared to the 1st quartile (Bootstrap analysis; $p < 0.001$, Wilcoxon rank-sum test)



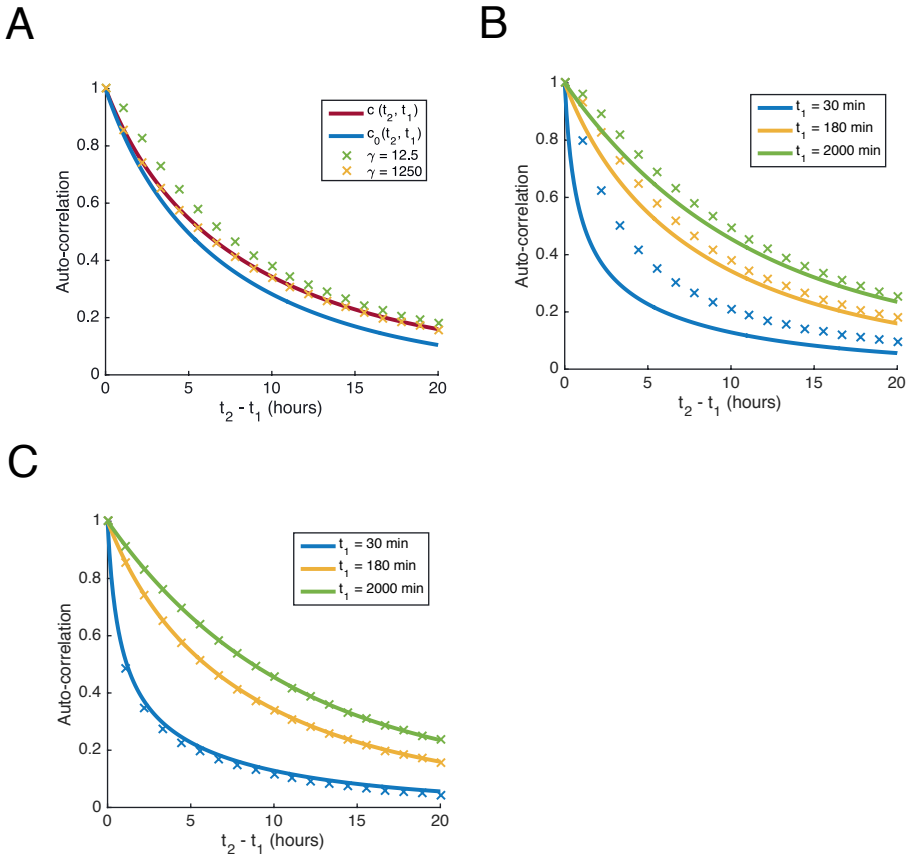
Supplementary Figure 5.21: Nearest neighbour analysis of a second independently transformed *p35S:2xNLS-YFP p35S:2xNLS-CFP* line (Transgenic line 2). (A) Scatter plot of *p35S:2xNLS-YFP p35S:2xNLS-CFP* young rosette leaves showing the normalized fluorescence intensities of cells plotted against the normalized fluorescence intensity of the nearest neighbour of the considered cells (neighbour cell with the lowest distance). Blue circles indicate the CFP fluorescence intensity of a cell (CFP1) plotted against the YFP fluorescence intensity of the nearest neighbouring cell (YFP2). Red circles show the YFP fluorescence intensity of a cell (YFP1) plotted against the CFP fluorescence intensity of the nearest neighbouring cell (CFP2) (n=1020 cells, $r = 0.423$, $p = 0.0014$, randomization test). (B) Dependency of the distance to the neighbouring cell and co-fluctuation in young rosette leaves of *p35S:2xNLS-YFP p35S:2xNLS-CFP*. Neighbouring cells were grouped into five tiers dependent on their distance (cell diameters) to the considered cell. Mean values and standard deviations are shown (n=39780 neighbourhood analyses). (C) Scatter plot of *p35S:2xNLS-YFP p35S:2xNLS-CFP* mature rosette leaves showing the normalized fluorescence intensities of cells plotted against the normalized fluorescence intensity of the nearest neighbour (n=796 cells, $r = 0.014$, $p = 0.451$, randomization test).



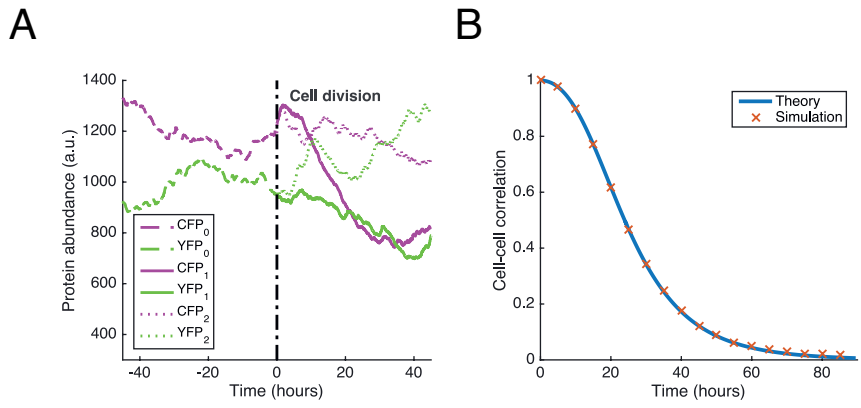
Supplementary Figure 5.22: Nearest neighbour analysis of root tips and hypocotyls of *p35S:2xNLS-YFP p35S:2xNLS-CFP* line.(A) Scatter plot of *p35S:2xNLS-YFP p35S:2xNLS-CFP* root tip cells showing the normalized fluorescence intensities of cells plotted against the normalized fluorescence intensity of the nearest neighbour of the considered cells (neighbour cell with the lowest distance). Blue circles indicate the CFP fluorescence intensity of a cell (CFP1) plotted against the YFP fluorescence intensity of the nearest neighbouring cell (YFP2). Red circles show the YFP fluorescence intensity of a cell (YFP1) plotted against the CFP fluorescence intensity of the nearest neighbouring cell (CFP2) (n=463 cells, r=0.447, p = 0.0008, randomization test). (B) Dependency of the distance to the neighbouring cell and co-fluctuation in root tip cells of *p35S:2xNLS-YFP* p35S:2xNLS-CFP. Neighbouring cells were grouped into five tiers dependent on their distance (cell diameters) to the considered cell. Mean values and standard deviations are shown (n=18057 (463 cells $\times$ 39 neighbouring cells) neighbourhood analyses). (C) Scatter plot of p35S:2xNLS- YFP *p35S:2xNLS-CFP* hypocotyl cells showing the normalized fluorescence intensities of cells plotted against the normalized fluorescence intensity of the nearest neighbour (n=690 cells; r = 0.379, p = 0.0, randomization test). (D) Dependency of the distance to the neighbouring cell and co-fluctuation in hypocotyl cells of *p35S:2xNLS-YFP* p35S:2xNLS-CFP. Neighbouring cells were grouped into five tiers dependent on their distance (cell diameters) to the considered cell. Mean values and standard deviations are shown (n=26910 (690 cells " 39 neighbouring cells) neighbourhood analyses).



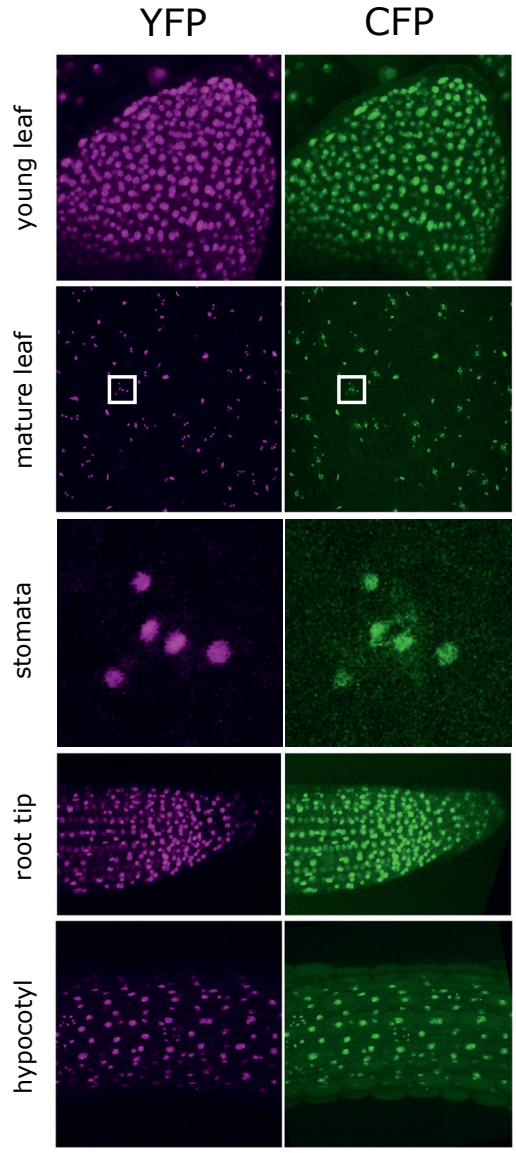
Supplementary Figure 5.23: Nearest neighbour analysis of root tips and hypocotyls of a second independently transformed *p35S:2xNLS-YFP p35S:2xNLS-CFP* line (Transgenic line 2).(A) Scatter plot of *p35S:2xNLS-YFP p35S:2xNLS-CFP* root tip cells showing the normalized fluorescence intensities of cells plotted against the normalized fluorescence intensity of the nearest neighbour of the considered cells (neighbour cell with the lowest distance). Blue circles indicate the CFP fluorescence intensity of a cell (CFP1) plotted against the YFP fluorescence intensity of the nearest neighbouring cell (YFP2). Red circles show the YFP fluorescence intensity of a cell (YFP1) plotted against the CFP fluorescence intensity of the nearest neighbouring cell (CFP2) (n=385 cells, $r = 0.375$, $p = 0.0018$, randomization test). (B) Dependency of the distance to the neighbouring cell and co-fluctuation in root tip cells of *p35S:2xNLS-YFP* p35S:2xNLS-CFP. Neighbouring cells were grouped into five tiers dependent on their distance (cell diameters) to the considered cell. Mean values and standard deviations are shown (n=15015 (385 cells $\times$ 39 neighbouring cells) neighbourhood analyses). (C) Scatter plot of *p35S:2xNLS-YFP p35S:2xNLS-CFP* hypocotyl cells showing the normalized fluorescence intensities of cells plotted against the normalized fluorescence intensity of the nearest neighbour (n=481 cells; $r = 0.241$, $p = 0.0135$, randomization test). (D) Dependency of the distance to the neighbouring cell and co-fluctuation in hypocotyl cells of *p35S:2xNLS-YFP* p35S:2xNLS-CFP. Neighbouring cells were grouped into five tiers dependent on their distance (cell diameters) to the considered cell. Mean values and standard deviations are shown (n=18759 (481 cells $\times$ 39 neighbouring cells) neighbourhood analyses).



Supplementary Figure 5.24: Simulation of the KikGR experiments. (A, B) Auto-correlation of the reporter system computed for different t_1. Solid lines denote theoretical predictions of the auto-correlation given by Eq. (2), crosses denote the stochastic simulation results, computed from 105 trajectories. (B) Auto-correlation with $\gamma = 12.5$. (C) Auto-correlation with $\gamma = 1250$. Parameters are for (A - C): $\nu_0 = 2.25$, $d_1 = 0.09$. For $\gamma = 12.5$: $\langle \nu_1 \rangle = 45$,$\mathrm{Var}(\nu_1) = 8.1$, $d_0 = 1.125$ and for $\gamma = 1250$: $\langle \nu_1 \rangle = 4500$, $\mathrm{Var}(\nu_1) = 8.1 \times 10^4$, $d_0 = 112.5$.



Supplementary Figure 5.25: Correlation due to inheritance. Cell division with dual reporter system. At $t = 0$ an identical copy of the mother cell is produced. All parameters of the two-stage model are inherited, besides the translational rates, which are in general different between mother and daughter cells. (a) Example trajectories of the situation before and after cell division are shown. The daughter cells inherit mRNA and protein content from the mother cell. All parameters of the two- stage gene expression model are as well inherited besides the translational rates, which are different between mother ($\nu_1 = 14.581$) and daughter cells ($\nu_1 = 10.449$ and $\nu_1 = 15.917$, respectively). The other parameters are: $\nu_0 = 2.25$, $d_0 = 1.125$, $d_1 = 0.029$. (b) Correlation calculated using Eq. 5.11 between daughter cells, computed from 10^5 simulated trajectories. The solid line denotes theoretical predictions of the autocorrelation given by Eq. 5.30, crosses denote the stochastic simulation results. Parameters are: $\nu_0 = 2.25$, $\rangle\nu_1\langle = 14.5$, $\mathrm{Var}(\nu_1) = 5$, $d_0 = 1.125$, $d_1 = 0.029$.



Supplementary Figure 5.26: Raw images of different cell types. Raw images are shown for the YFP and CFP channel for all cell types/tissues. The mean fluorescence intensities were in the following ranges (YFP, CFP). For Line 1: hypocotyl 110, 101, mature leaf 155, 118, root tip 112, 110, stomata 99, 106, young leaf 60, 83. For Line 2: hypocotyl 75, 90, mature leaf 114, 96, root tip 118, 116, stomata 134, 126, young leaf 39, 74. For the ubiquitin promoter line: mature leaf 141, 83, young leaf 58, 52.

Chapter 6

General discussion

In this thesis, I have investigated how environmental noise impacts phenotype and functioning for several biological systems through the integration of experimental approaches and mathematical modelling. These systems range from isolated gene regulatory networks (**Chapter 2**) and isogenic populations of prokaryotes (**Chapters 3–4**) to plant cells which are embedded in a tissue (**Chapter 5**). Due to limitations in how many features we can measure experimentally, we can not always incorporate every aspect that causes variability in our models. A question that may arise at this point is whether we can distill fundamental relationships between phenotypic variability and cellular functioning, or whether we need to account for all sources of variability to understand biology. There are distinct challenges associated with the analysis of each of the biological systems discussed in this thesis. These challenges present themselves both in the generation of experimental data and the correct mathematical representation and interpretation of gene expression noise. More often than not, experimental and theoretical challenges are intricately linked. In this chapter, I will discuss some of these challenges, how they have impacted the results presented in this work, and look ahead to opportunities for future progress.

6.1 Stochastic gene expression in dynamic environments: a changing definition of extrinsic noise?

Traditionally, the variability that is observed in the expression levels of genetically identical cells is categorised as having intrinsic and extrinsic origins. Extrinsic noise is defined as the variation that arises from the environment in which the network is embedded, and thus will depend on what details are included in the definition of the system. The intrinsic component comes from the random timing of individual reaction events, its magnitude dictated by the architecture of the gene regulatory network and the parameters associated with it. We have found in all the examples discussed in this work, extrinsic noise contributes substantially to the total variability in gene expression levels.

In the gene expression models studied with the approximate analytical framework described in **Chapter 2**, extrinsic noise enters the system through fluctuations in the reaction rates. These fluctuations, which are characterised by their slow time scale and non-Gaussian character, arise from unobserved upstream influences, such as changes in temperature or ribosome numbers. Depending on the reaction within the gene regulatory network that is affected by these external influences, the mean number of proteins in the system can either increase or decrease. In addition, the variability of protein levels over time and across populations can increase substantially as the amplitude of the extrinsic noise gets larger. We have also shown that noise properties such as the time scale and strength of the extrinsic fluctuations determine how signal transduction is affected. Our results imply that robustness to extrinsic noise should be seriously taken into account when designing

synthetic biological systems.

The conclusions from **Chapter 2** are dependent on the condition that the system is in a stationary state. However, for many biological systems this is not the case: for example, bacterial populations in the exponential growth phase continually grow and divide. Classical models of stochastic gene expression lack a cell size and growth description. Now that modern experimental techniques such as time-lapse microscopy allow for the dynamic monitoring of gene expression in individual cells over multiple generations, in recent years we have seen a changing definition of extrinsic noise. With this broader definition of extrinsic noise coming from the population dynamics, the old methods of measuring gene expression noise are no longer sufficient: due to the age structure of the population, the statistics of lineages and population snapshots are not equivalent.[17,23,233] In addition to the age structure of a population, cell cycle variations (division time variability, asymmetric partitioning, etc.) also contribute to gene expression noise, and the inheritance of protein and mRNA content from the mother cell affects the next generation of cells. However, as we have shown in **Chapter 5** in plants, this inheritance of mRNAs and proteins is often not enough to account for the variability in daughter cells. This implies that the inheritance of other cell features, such as cell size or cellular state (*e.g.* the abundances of polymerases and ribosomes), also propagate noise from one generation to the next causing reaction rates to fluctuate over time and between cells.

In **Chapters 3–4**, we have studied the CRISPR-Cas immune response in a population of *Escherichia coli* which continually grow and divide. Due to cell growth, factors like cellular growth rate, cell cycle length, and the inheritance of proteins become important contributions to gene expression noise at the single-cell level. As the molecular basis of the CRISPR-Cas immune system is too complex to gain mechanistic insight through analytical treatment, we have employed stochastic simulations of the system dynamics. In order to incorporate the population dynamics we have made use of an agent-based simulation framework, in which each cell is an agent for which we simulate the intracellular stochastic reaction dynamics individually. This approach enables a realistic comparison with the experimental single-cell lineage data, and allows us to relate single-cell fluctuations to heterogeneity at the population level. We have found that cell-to-cell differences in growth rate affect survival probabilities: on the one hand, fast growth enables bacteria to clear previously encountered invaders faster, whereas slow growth is associated with increased chances of adapting to new threats. Cellular growth rate was also found to influence the concentration of the CRISPR surveillance complex Cascade, which is responsible for locating invading elements and initiating the immune response. Due to the stability of this protein complex, cells with increased Cascade levels can pass this on to future generations, thus increasing their probability of successfully defending the cell against future invading elements. As the prolonged presence of plasmids, in contrast to other mobile genetic elements such as bacteriophages, does

not result in cell death, this allowed for long-term imaging of the bacterial population. However, future experiments with *e.g.* phages are necessary to corroborate the long-term effects of growth rate on population fitness and survival.

6.2 Limitations of the methodology

The stochasticity of reaction kinetics that arises from low copy numbers has historically been studied in detail using the mathematical framework of the chemical master equation. The CME assumes the reaction volume is both well-mixed and dilute, which allows one to ignore the spatial organisation of the system. However, these assumptions are not always a realistic portrayal of physiological *in vivo* conditions. Instead, the interiors of cells are crowded environments, containing many different molecular species which may be present in low numbers but together make up a substantial portion of the system volume ($20 - 30\%$ in general).[234,235] This affects reaction rates in opposing ways: on the one hand, the presence of macromolecules reduces the diffusion coefficients of molecules. On the other hand, the volume excluded by these macromolecules decreases the available reaction volume and thus enhances the association of interacting molecular species. The overall effect of crowding on reaction rates is thus complex and depends on multiple factors, such as the magnitude of excluded volume and the nature of the reaction in question.

Gillespie's stochastic simulation algorithm, which is used to simulate trajectories which are exact realisations of the underlying chemical reaction process modelled with the CME, suffers from the same limitation as the CME itself. In addition to this, a limitation of stochastic simulation algorithms with extrinsic processes, such as the *Extrande* extension used in this thesis, is that it assumes that the external inputs influence the system of interest but the latter does not influence the inputs. In other words, the method requires that the inputs can be pre-simulated.[184] This limits the nature of the external stimuli that affect the system: it allows us to investigate how light and temperature affect the stochastic dynamics, but the method is less generally applicable to chemical stimuli. In the case of a fluctuating chemical input, the method can accurately describe the effects on the system dynamics if the system does not in turn affect the original stimulus. Alternatively, if a regulatory mechanism between the system and the extrinsic process exists, it is necessary to incorporate these interactions in the system definition.

6.3 Applicability

Systems biology aims to uncover how the dynamic system behaviour of complex biological systems emerges from the interactions of its many components. With *omics* approaches becoming cheaper and more accurate, nowadays we are able

to amass large amounts of biological data. These allow a top-down descriptive understanding of the system, and we can infer how components within a system are connected. However, the statistical analysis of high-throughput measurements does not automatically lead to a mechanistic understanding of how cells, tissues, and organisms function. Often, statistical analysis returns information at the level of gene-phenotype correlations. These correlations can be used in a top-down approach and then lead to statistical models that may yield accurate predictions within the conditions of the original data. However, it does not allow for generalisation, and the high dimensionality of data makes it difficult to distinguish between correlations and causal relationships.[236]

On the other side of the spectrum are mechanistic bottom-up modelling approaches, which describe system interaction at the level of individual molecular components and their local environment. These bottom-up approaches require at least some level of *a priori* knowledge of the underlying gene regulatory interactions and reaction mechanism. In **Chapter 2**, we have seen that the use of analytical methods such as those based on the linear noise approximation is limited to a subset of simple network motifs. While at this high level of mechanistic detail, it is computationally expensive to simulate the stochastic system dynamics at the level of a cell or an organism, these small models do allow us to make experimental predictions and provide support for biological hypotheses. In future years, challenges lie in bridging the gap between these detailed stochastic models and more efficient phenomenological approaches.[237]

In order to study biological systems using stochastic models, knowledge of the kinetic rate parameters that govern the biochemical reactions is required. As bulk measurements provide only information about the average behaviour of intracellular dynamics, parameters obtained from ensemble averages may not result in good predictions of the behaviour of individual cells. In some cases, fluctuating single-cell measurements in fact contain more information about model parameters than ensemble averages.[238] However, the reliable inference of reaction rate constants and other model parameters from heterogeneous single-cell data is challenging, especially in the presence of extrinsic noise.[239] Existing approaches can be roughly classified into those focussing on measurements from population snapshot[238,240–242] or time-lapse microscopy approaches.[239,243] The former does not result in measurements that contain temporal correlations on a single-cell level, whereas the latter allows cells to be tracked throughout an experiment, resulting in multiple measurements of the same cell. Unlike snapshot studies, in time-lapse imaging the measurements are not statistically independent of each other. Failing to take this statistical dependence into account ignores important information contained in the ancestry of cells, and can lead to biased results.[243] In order to maximise the information we can gain from measurements about the model parameters, there is a need for a theory of single-cell experimental design. While becoming increasingly important, assessing structural identifiability when performing parameter inference for stochastic models

in systems biology is still in its infancy.[199]

6.4 Challenges associated with experimental single-cell approaches

6.4.1 Using fluorescent biosensors to illuminate cell-to-cell variability

Fluorescent reporters have become widely utilised to study phenotypic heterogeneity, and have become instrumental in uncovering the molecular mechanisms behind cellular processes. In **Chapters 3–5** of this thesis, we have made use of fluorescent protein (FP) reporters to quantify gene expression noise. Fluorescent biosensors have the advantage that cells can be observed non-invasively and directly show the variability of processes. However, the number of molecular species that can be tracked simultaneously is limited due to a limited number of fluorescent proteins and spectral overlap, and so most processes remain unobserved.

Many single-cell studies are based on snapshot analysis, where fluorescence levels of all cells in the population are only recorded for a single time point. While snapshot studies can capture the heterogeneity of cell populations, the resulting estimate of the phenotypic heterogeneity can only give information on long-term behaviour of single cells if these cells behave ergodically. In these systems, the information obtained from averaging over the history of one cell is equivalent to the information obtained from averaging over a whole ensemble of cells at one moment in time. In practice, this is not always the case, and as a result statistics obtained from cross-sectional (snapshot) and longitudinal studies are not equivalent. Furthermore, data obtained from snapshot studies limits how much insight can be gained about the sources of this variability. This is because measurements that are not time-resolved do not allow for reliable identification of the molecular and cellular dynamics underlying the heterogeneity. For example, we cannot isolate factors that contribute to this heterogeneity such as the cell cycle phase or cell age. Additionally, population snapshots give no information on the lifetime of fluctuating or periodic influences: two stochastic processes that fluctuate on different time scales can result in the same copy number distributions.[244] Instead, the autocorrelation function of time series can give information about the time scale on which molecular components fluctuate, and thus the structural properties of the noise.

Calculating the autocorrelation function thus requires time-series data of the processes under investigation. However, long-term monitoring of cell traits requires keeping the extracellular environment as constant as possible. In **Chapter 5**, we used plasmid-based fluorescent reporters and protein-FP fusions to study gene expression in different cell types of *Arabidopsis thaliana*. Due to the use of excised leaves, it is a challenge to keep cells and tissues healthy for prolonged periods of

time under a microscope. For this reason, we had access to a limited number of data points to quantify the temporal fluctuations of the fluorescent protein concentrations. The consequence of this is that we cannot detect fluctuations which happen on a timescale faster than the inter-measurement time. In **Chapter 3** and **Chapter 4**, we used time-lapse microscopy to follow the expression of a plasmid-based fluorescent reporter protein and Cascade-FP fusion protein expression levels for a prolonged period (over 36 hours) and were able to obtain measurements every few minutes. During time-lapse imaging, it is important that a constant cellular environment is maintained: for example, temperature gradients can affect cell growth and result in artefacts in the data. For this reason, the microscope set-up used in **Chapters 3–4** was placed in a temperature-controlled box. However, induction of the CRISPR-Cas response at the start of the experiment causes an increase in the metabolic burden, which results in a non-stationary population growth rate on the timescale of the experiment thus complicating the statistical analysis of the experimental data.

While providing very detailed temporal information on cell growth and gene expression dynamics, long-term imaging requires extensive processing of the acquired images: correcting the automatically generated segmentation of cells in each frame and tracking of cells between frames. We have used a 'traditional' computational segmentation program to automatically annotate pixels to detect cell boundaries (custom software based on the *Schnitzcells* package[150]). In the near future, approaches based on machine-learning algorithms might result in faster, more reliable segmentation and tracking results, although calibration on a per-experiment basis is likely still required.

An alternative to fluorescence methods can be found in single-cell *omics* approaches such as single-cell RNA sequencing or single-cell proteomics. These methods are cheap and high-throughput, as well as enable the quantification of a large number of different molecules. However, due to their destructive nature, we can obtain only snapshot data from the population, which means we lose temporal information about the fluctuations. An overview of the current state of and challenges associated with single-cell data science can be found in the recent review by Lähnemann *et al.*[245]

6.4.2 Is decomposing gene expression noise into intrinsic and extrinsic components informative?

The introduction of the dual reporter system by Elowitz *et al.* allowed for the quantification of intrinsic and extrinsic contributions to the total gene expression noise in a way that is easy to implement and measure experimentally. The magnitude of the extrinsic noise is defined as the normalised covariance between the protein levels of two statistically independent twin reporters, leaving the remainder of the total noise as intrinsic.[7] However, it has been pointed out that the decomposition of

gene expression noise as measured by the dual reporter method does not generally give a valid description of intrinsic versus extrinsic noise.[12] The correlations only give valid interpretations for static environmental heterogeneity (which varies across systems but is constant in time), and not for a dynamically changing cellular environment. In those cases, it is necessary to know the history of the environment.[12,225] Furthermore, when only the intrinsic and extrinsic contributions to the variance are measured by means of twin reporter proteins, it is generally not possible to infer what the mechanisms are that cause the variability.[246] In order to distinguish between different mechanisms, higher order moments and temporal correlations are required.[246]

In growing cell populations, the variability in cell division times causes heterogeneity of cell ages, thus complicating the quantification of the extrinsic component to the variance of molecule numbers. In order to separate the variability that results from this periodic cell growth effect from that due to other stochastic components, it is necessary to condition the covariance of dual reporter expression levels on the cell age.[17] This means that in order to estimate intrinsic and extrinsic noise in single-cell lineages, we need time-lapse studies of dual reporter systems.[247] If the cell age is unknown there are some experimental approaches that can be used to make the extrinsic contribution to the total variability accessible from snapshot data.[12,17,248]

6.5 Stochastic effects: connecting the microscopic and the macroscopic scale

6.5.1 The microscopic scale: is there noise beyond unobserved information?

Biochemical reactions are often said to be 'inherently stochastic', but where exactly does this non-deterministic behaviour come from? Biochemical reactions are a consequence of erratic Brownian motion: the random movement of particles in a fluid due to their collisions with other atoms or molecules. In association reactions, involving two molecular species, this causes the two molecules to collide with enough energy to spark a reaction. Unimolecular reactions, such as isomerisation, dissociation, or degradation reactions, are able to take place because these collisions with other particles result in the required energy for activation of the reaction.

Extrinsic noise comes from unobserved processes that affect the system dynamics. In the event where one is able to measure the abundances of all molecular components which affect the system of interest at a detailed temporal and spatial level, this will result in accurate but very large and possibly intractable models. Furthermore, many extrinsic noise processes have both stochastic and deterministic influences, and it is not straightforward to delineate between their relative contri-

butions to the overall variability.[249] For example, while the cell cycle is periodic there is cell-to-cell variability in the duration of the different cell cycle phases which results in a non-uniform cell cycle duration. The sources of this variability are largely unknown.[250] There have been instances where stochastic extrinsic noise has been found in fact to have deterministic origins and causes correlations over multiple generations.[251,252]

6.5.2 The macroscopic scale: does stochastic gene expression affect higher organisms?

Throughout this thesis, we have moved along the biological 'ladder' from gene regulatory networks, bacterial populations, to cells embedded within a tissue. In recent years it has become more clear what the origins and implications of gene expression noise are at the (sub)cellular level, and how the resulting cell-to-cell phenotypic variability can benefit bacterial populations. It might seem natural that isogenic populations of bacteria exploit this heterogeneity: after all, they have limited alternative options to defend themselves against a changing environment. However, for many larger organisms diversity can arise from other factors, such as sexual reproduction and (learned) behaviour. Thus, on the macroscopic scale, it is usually assumed that gene expression noise is negligible. However, it is still largely unknown if phenotypic variation at the level of plant and animal cells averages out in favour of robust development, or if gene expression noise could actually affect these higher-level organisms. While the study of cell-to-cell variability in higher eukaryotes is still in its infancy, there are some examples from developmental biology that show that stochastic gene expression is in fact functional and can trigger cell-fate decision.[253–255] Currently, the analysis of gene expression noise is still limited to a small subset of genes, whereas in higher organisms there are many different processes going on at the same time that need to be tightly coordinated. These processes usually evolve over the course of multiple cell cycles. However, the *in vivo* monitoring of stochastic gene expression in most mammalian tissues at a high spatiotemporal resolution with live imaging approaches is extremely challenging, except for experiments with limited durations.[256]

In this thesis, we have among other things considered extrinsic noise coming from the cell cycle dynamics. Beyond the inheritance of the mRNA and protein content at cell division, recently it was shown in bacterial and mammalian cells that other cellular features such as cell cycle duration can be passed on from a mother cell to its daughters, which leads to correlations on a timescale longer than the cell cycle itself.[15,251] It has been suggested that some of these long-term correlations are caused by epigenetic mechanisms (*e.g.* DNA methylation), which may be inherited under certain conditions and can thus modify gene expression on the scale of multiple generations.[254,255] These epigenetic mechanisms will add another layer of complexity to the study of how environmental factors affect biological functioning

and long-term population fitness.

6.6 Outlook

"Reality is not a point; it is a cloud of possibilities" – Amos Tversky

In modern-day science, 'nature versus nurture' is often interpreted as the relative contributions of genetic and environmental factors to phenotypic variability. Since around 1900, we are aware that genetic diversity causes differences between and within species. This variation in the gene pool of populations allows for natural selection, thus enabling a population to adapt to a changing environment. One gradually came to the insight that an organism's genome does not code for a specific phenotype, but rather acts as a rule book for development. These rules drive self-organisation, and so determine the range of possible outcomes. In this landscape of possibilities, environmental influences shape the phenotype of an organism. But genetically identical bacteria still display phenotypic heterogeneity, despite being embedded in controlled laboratory conditions. The reason for this is the unavoidable stochasticity inherent to biochemical reaction events. In this view, 'nature versus nurture' should perhaps be replaced by 'nature versus nurture versus noise', which refers to the genetic, environmental, and intrinsic contributions to the overall phenotypic variability respectively.

While it has been well established that genetic diversity plays an important role in adaptability and survival of species, more recently it has become clear that both prokaryotes and eukaryotes exploit non-genetic variability. Cells appear to have evolved gene regulatory network architectures in order to be better equipped against changing circumstances on a population level, either by exploiting or suppressing this noise.[32] Bet-hedging strategies have been found in bacteria, yeast, and mammals.[257–261] In turn, there is evidence that natural selection has lead to minimised expression noise of genes in yeast for which under- or overexpression is harmful.[262] These examples show that an interplay between nature, nurture, and noise on cellular functioning can exist. However, it is not always straightforward to describe how cell-to-cell variability is driven by each of these categories.

While computational and experimental advances in molecular biology have moved our understanding of variability in biological systems forward, many challenges remain in the experimental quantification, statistical analysis, and computational modelling of gene expression noise. In this thesis, we have discussed some of these computational and experimental difficulties. However, the biggest challenge in mathematical biology yet might be the adoption of models as predictive tools. Although currently mathematical models are often used to explain experimental results, hopefully we can move towards a future where models are used to generate hypotheses which can be supported by experimental studies. With the increasing

availability of high-throughput data, dynamic mechanistic models remain necessary to connect cell-to-cell variability to biological function.

Bibliography

[1] JM Schleiden. Beiträge zur phytogenesis. *Archiv für Anatomie, Physiologie und Wissenschaftliche Medicin, 4: 137*, 170, 1838.

[2] Theodor Schwann. Mikroskopische untersuchungen über die uebereinstimmung in der struktur und dem wachstum der thiere und pflanzen berlin. *Microscopical researches into the accordance in the structure and growth of animals and plants*, page 141, 1839.

[3] M Delbrück. The burst size distribution in the growth of bacterial viruses (bacteriophages). *Journal of bacteriology*, 50(2):131–135, 1945.

[4] Harley H McAdams and Adam Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, 94(3):814–819, 1997.

[5] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

[6] Adam Arkin, John Ross, and Harley H McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected escherichia coli cells. *Genetics*, 149(4):1633–1648, 1998.

[7] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[8] J. M. Raser and E. K. O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304, 2004.

[9] Juan M Pedraza and Alexander van Oudenaarden. Noise propagation in gene networks. *Science*, 307(5717):1965–1969, 2005.

[10] Peter S Swain, Michael B Elowitz, and Eric D Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800, 2002.

[11] William J Blake, Mads Kærn, Charles R Cantor, and James J Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637, 2003.

[12] Andreas Hilfinger and Johan Paulsson. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences*, 108(29):12167–12172, 2011.

[13] Berend Snijder and Lucas Pelkmans. Origins of regulated cell-to-cell variability. *Nature reviews Molecular cell biology*, 12(2):119–125, 2011.

[14] Frank J Bruggeman, Jaap Schouten, Daan H de Groot, and Robert Planqué. Cell fate determination by lamarckian molecule-inheritance and chance. *bioRxiv*, page 618199, 2019.

[15] Bram Cerulus, Aaron M New, Ksenia Pougach, and Kevin J Verstrepen. Noise and epigenetic inheritance of single-cell division times influence population fitness. *Current Biology*, 26(9):1138–1147, 2016.

[16] Johan H Van Heerden, Hermannus Kempe, Anne Doerr, Timo Maarleveld, Niclas Nordholt, and Frank J Bruggeman. Statistics and simulation of growth of single bacterial cells: illustrations with b. subtilis and e. coli. *Scientific reports*, 7(1):1–11, 2017.

[17] Philipp Thomas. Intrinsic and extrinsic noise of gene expression in lineage trees. *Scientific reports*, 9(1):1–16, 2019.

[18] Dmitri Volfson, Jennifer Marciniak, William J Blake, Natalie Ostroff, Lev S Tsimring, and Jeff Hasty. Origins of extrinsic variability in eukaryotic gene expression. *Nature*, 439(7078):861–864, 2006.

[19] Anne Schwabe and Frank J Bruggeman. Contributions of cell growth and biochemical reactions to nongenetic variability of cells. *Biophysical journal*, 107(2):301–313, 2014.

[20] Jakub Jędrak and Anna Ochab-Marcinek. Contributions to the 'noise floor' in gene expression in a population of dividing cells. *Scientific Reports*, 10(1):1–13, 2020.

[21] Ruben Perez-Carrasco, Casper Beentjes, and Ramon Grima. Effects of cell cycle variability on lineage and population measurements of messenger rna abundance. *Journal of the Royal Society Interface*, 17(168):20200360, 2020.

[22] Lucy Ham, Marcel Jackson, and Michael Stumpf. Pathway dynamics can delineate the sources of transcriptional noise in gene expression. *eLife*, 10:e69324, oct 2021.

[23] Philipp Thomas. Making sense of snapshot data: ergodic principle for clonal cell populations. *Journal of The Royal Society Interface*, 14(136):20170467, 2017.

[24] Philipp Thomas and Vahid Shahrezaei. Coordination of gene expression noise with cell size: analytical results for agent-based models of growing cell populations. *Journal of the Royal Society Interface*, 18(178):20210274, 2021.

[25] Ines SC Baptista and Andre S Ribeiro. Stochastic models coupling gene expression and partitioning in cell division in escherichia coli. *Biosystems*, 193:104154, 2020.

[26] I. Golding, J. Paulsson, S.M. Zawilski, and E.C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123:1025–1036, 2005.

[27] Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337, 2011.

[28] David M Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028):472–474, 2011.

[29] Guilhem Chalancon, Charles NJ Ravarani, S Balaji, Alfonso Martinez-Arias, L Aravind, Raja Jothi, and M Madan Babu. Interplay between gene expression noise and regulatory network architecture. *Trends in genetics*, 28(5):221–232, 2012.

[30] A. Becksei and L. Serrano. Engineering stability in gene networks by autoregulation. *Nature*, 405:590–593, 2000.

[31] Mads Kaern, Timothy C Elston, William J Blake, and James J Collins. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6(6):451, 2005.

[32] Y. Dublanche, K. Michalodimitrakis, N. Kümmerer, M. Foglierini, and L. Serrano. Noise in transcription negative feedback loops: simulation and experimental analysis. *Molecular Systems Biology*, 2(41), 2006.

[33] Wiet Hendrik de Ronde, Filipe Tostevin, and Pieter Rein Ten Wolde. Effect of feedback on the fidelity of information transmission of time-varying signals. *Physical Review E*, 82(3):031914, 2010.

[34] Attila Becskei, Bertrand Séraphin, and Luis Serrano. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *The EMBO journal*, 20(10):2528–2535, 2001.

[35] William J Blake, Gábor Balázsi, Michael A Kohanski, Farren J Isaacs, Kevin F Murphy, Yina Kuang, Charles R Cantor, David R Walt, and James J Collins. Phenotypic consequences of promoter-mediated transcriptional noise. *Molecular cell*, 24(6):853–865, 2006.

[36] Daniel A Charlebois, Nezar Abdennur, and Mads Kaern. Gene expression noise facilitates adaptation and drug resistance independently of mutation. *Physical review letters*, 107(21):218101, 2011.

[37] Stephen Smith and Ramon Grima. Single-cell variability in multicellular organisms. *Nature communications*, 9(1):1–8, 2018.

[38] Andriy Marusyk, Vanessa Almendro, and Kornelia Polyak. Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, 12(5):323–334, 2012.

[39] Corey E Hayford, Darren R Tyson, C Jack Robbins III, Peter L Frick, Vito Quaranta, and Leonard A Harris. An in vitro model of tumor heterogeneity resolves genetic, epigenetic, and stochastic sources of cell state variability. *PLoS biology*, 19(6):e3000797, 2021.

[40] David Sprinzak, Amit Lakhanpal, Lauren LeBon, Leah A Santat, Michelle E Fontes, Graham A Anderson, Jordi Garcia-Ojalvo, and Michael B Elowitz. Cis-interactions between notch and delta generate mutually exclusive signalling states. *Nature*, 465(7294):86–90, 2010.

[41] N.G. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, Amsterdam, 3 edition, 2007.

[42] R. Grima, D.R. Schmidt, and T.J. Newman. Steady-state fluctuations of a genetic feedback loop: An exact solution. *Journal of Chemical Physics*, 137:035104, 2012.

[43] Niraj Kumar, Thierry Platini, and Rahul V Kulkarni. Exact distributions for stochastic gene expression models with bursting and feedback. *Physical review letters*, 113(26):268105, 2014.

[44] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434, 1976.

[45] Vahid Shahrezaei and Peter S Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–17261, 2008.

[46] David F Anderson. A modified next reaction method for simulating chemical systems with time dependent propensities and delays. *The Journal of chemical physics*, 127(21):214107, 2007.

[47] Margaritis Voliotis, Philipp Thomas, Ramon Grima, and Clive G Bowsher. Stochastic simulation of biomolecular networks in dynamic environments. *PLoS computational biology*, 12(6):e1004923, 2016.

[48] N.G. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, 1997.

[49] E.W. Wallace, D.T. Gillespie, K.R. Sanft, and L.R. Petzold. Linear noise approximation is valid over limited times for any chemical system that is sufficiently large. *IET Systems Biology*, 6(4):102–115, 2012.

[50] Ramon Grima. Linear-noise approximation and the chemical master equation agree up to second-order moments for a class of chemical systems. *Physical Review E*, 92(4):042124, 2015.

[51] Michał Komorowski, Bärbel Finkenstädt, and Claire V Harper. Bayesian inference of biochemical kinetic parameters using the linear noise approximation. 2009.

[52] Bärbel Finkenstädt, Dan J Woodcock, Michal Komorowski, Claire V Harper, Julian RE Davis, Mike RH White, and David A Rand. Quantifying intrinsic and extrinsic noise in gene transcription using the linear noise approximation: An application to single cell data. *The Annals of Applied Statistics*, pages 1960–1982, 2013.

[53] Silvia Calderazzo, Marco Brancaccio, and Bärbel Finkenstädt. Filtering and inference for stochastic oscillators with distributed delays. *Bioinformatics*, 35(8):1380–1387, 2019.

[54] Elco Bakker and Peter S Swain. Estimating numbers of intracellular molecules through analysing fluctuations in photobleaching. *Scientific reports*, 9(1):1–13, 2019.

[55] D.T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81:2340–2361, 1977.

[56] Jonathan M Raser and Erin K O'shea. Noise in gene expression: origins, consequences, and control. *Science*, 309(5743):2010–2013, 2005.

[57] C.V. Rao, D.M. Wolf, and A.P. Arkin. Control, exploitation and tolerance of intracellular noise. *Nature*, 420:231–237, 2002.

[58] Raymond Cheong, Alex Rhee, Chiaochun Joanne Wang, Ilya Nemenman, and Andre Levchenko. Information transduction capacity of noisy biochemical signaling networks. *Science*, 334(6054):354–358, 2011.

[59] M.B. Elowitz, A.J. Levine, E.D. Siggia, and P.S. Swain. Stochastic gene expression in a single cell. *Science*, 297:1183–1186, 2002.

[60] J.M. Raser and E.K. O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304:1811–1814, 2004.

[61] Ilka Schultheiß Araújo, Jessica Magdalena Pietsch, Emma Mathilde Keizer, Bettina Greese, Rachappa Balkunde, Christian Fleck, and Martin Hülskamp. Stochastic gene expression in arabidopsis thaliana. *Nature communications*, 8(1):2132, 2017.

[62] V. Shahrezaei, J.F. Ollivier, and P.S. Swain. Colored extrinsic fluctuations and stochastic gene expression. *Molecular Systems Biology*, 4:196, 2008.

[63] David Schnoerr, Guido Sanguinetti, and Ramon Grima. Approximation and inference methods for stochastic biochemical kinetics: a tutorial review. *Journal of Physics A: Mathematical and Theoretical*, 50(9):093001, 2017.

[64] M. Scott, B. Ingalls, and M. Kaern. Estimations of intrinsic and extrinsic noise in models of nonlinear genetic networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 16(2):026107, 2006.

[65] T. Toni and B. Tidor. Combined model of intrinsic and extrinsic variability for computational network design with application to synthetic biology. *PLoS Computational Biology*, 9(3):e1002960, 2013.

[66] Elijah Roberts, Shay Be'er, Chris Bohrer, Rati Sharma, and Michael Assaf. Dynamics of simple gene-network motifs subject to extrinsic fluctuations. *Physical Review E*, 92(6):062717, 2015.

[67] N. Rosenfeld, J.W. Young, P.S. Swain, and M.B. Elowitz. Gene regulation at the single-cell level. *Science*, 307(5717):1962–1965, 2005.

[68] Zach Hensel, Haidong Feng, Bo Han, Christine Hatem, Jin Wang, and Jie Xiao. Stochastic expression dynamics of a transcription factor revealed by single-molecule noise analysis. *Nature Structural and Molecular Biology*, 19(8):797, 2012.

[69] Chikara Furusawa, Takao Suzuki, Akiko Kashiwagi, Tetsuya Yomo, and Kunihiko Kaneko. Ubiquity of log-normal distributions in intra-cellular reaction dynamics. *Biophysics*, 1:25–31, 2005.

[70] Martin Bengtsson, Anders Ståhlberg, Patrik Rorsman, and Mikael Kubista. Gene expression profiling in single cells from the pancreatic islets of langerhans reveals lognormal distribution of mrna levels. *Genome research*, 15(10):1388–1392, 2005.

[71] Peng Lu, Christine Vogel, Rong Wang, Xin Yao, and Edward M Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature biotechnology*, 25(1):117, 2007.

[72] N. G. van Kampen. A power series expansion of the master equation. *Canadian Journal of Physics*, 39(4):551–567, 1961.

[73] C.W. Gardiner. *Handbook of stochastic methods for physics, chemistry and natural sciences*. Springer-Verlag, Berlin; Heidelberg [etc.], 2 edition, 1990.

[74] A.N. Malakhov. *Cumulant Analysis of Non-Gaussian Random Processes and Their Transformations [in Russian]*. Sovetskoe Radio, Moscow, 1978.

[75] D.T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58(1):35–55, 2007.

[76] Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, January 2017.

[77] B. Bastian. ext_noise_expansion – Linear noise approximation with slow extrinsic fluctuations. `https://github.com/basbjo/ext_noise_expansion`, 2013.

[78] B. Bastian. Spatial and temporal patterns in biology. Master's thesis, University of Freiburg, Germany, May 2013.

[79] Ramon Grima. Linear-noise approximation and the chemical master equation agree up to second-order moments for a class of chemical systems. *Physical Review E*, 92(4):042124, 2015.

[80] Lubomír Brančík and Edita Kolářová. Simulation of multiconductor transmission lines with random parameters via stochastic differential equations approach. *Simulation*, 92(6):521–533, 2016.

[81] Jiajun Zhang, Zhanjiang Yuan, and Tianshou Zhou. Physical limits of feedback noise-suppression in biological networks. *Physical Biology*, 6(4):046009, 2009.

[82] Ioannis Lestas, Glenn Vinnicombe, and Johan Paulsson. Fundamental limits on the suppression of molecular fluctuations. *Nature*, 467(7312):174, 2010.

[83] Andreas Grönlund, Per Lötstedt, and Johan Elf. Transcription factor binding kinetics constrain noise suppression via negative feedback. *Nature communications*, 4:1864, 2013.

[84] Diego A Oyarzún, Jean-Baptiste Lugagne, and Guy-Bart V Stan. Noise propagation in synthetic gene circuits for metabolic control. *ACS Synthetic Biology*, 4(2):116–125, 2014.

[85] Zhixing Cao and Ramon Grima. Linear mapping approximation of gene regulatory networks with stochastic dynamics. *Nature communications*, 9(1):3305, 2018.

[86] Johan Paulsson. Models of stochastic gene expression. *Physics of life reviews*, 2(2):157–175, 2005.

[87] F Lang, M Ritter, E Wöll, H Weiss, D Häussinger, J Hoflacher, K Maly, and H Grunicke. Altered cell volume regulation in ras oncogene expressing nih fibroblasts. *Pflügers Archiv*, 420(5-6):424–427, 1992.

[88] Arren Bar-Even, Elad Noor, Yonatan Savir, Wolfram Liebermeister, Dan Davidi, Dan S Tawfik, and Ron Milo. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry*, 50(21):4402–4410, 2011.

[89] T.T. Marquez-Lago and J. Stelling. Counter-intuitive stochastic behaviour of simple gene circuits with negative feedback. *Biophysical Journal*, 98:1742–1750, 2010.

[90] Margaritis Voliotis and Clive G Bowsher. The magnitude and colour of noise in genetic negative feedback systems. *Nucleic acids research*, 40(15):7084–7095, 2012.

[91] Abhyudai Singh and Joao P Hespanha. Optimal feedback strength for noise suppression in autoregulatory gene networks. *Biophysical journal*, 96(10):4013–4023, 2009.

[92] Iain G Johnston, Bernadett Gaal, Ricardo Pires das Neves, Tariq Enver, Francisco J Iborra, and Nick S Jones. Mitochondrial variability as a source of extrinsic cellular noise. *PLoS computational biology*, 8(3):e1002416, 2012.

[93] Nitzan Rosenfeld, Michael B Elowitz, and Uri Alon. Negative autoregulation speeds the response times of transcription networks. *Journal of molecular biology*, 323(5):785–793, 2002.

[94] Filipe Tostevin and Pieter Rein Ten Wolde. Mutual information between input and output trajectories of biochemical networks. *Physical Review Letters*, 102(21):218101, 2009.

[95] Davide Bernardi and Benjamin Lindner. A frequency-resolved mutual information rate and its application to neural systems. *Journal of Neurophysiology*, 113(5):1342–1357, March 2015.

[96] F Gabbiani. Coding of time-varying signals in spike trains of linear and half-wave rectifying neurons. *Network: Computation in Neural Systems*, 1996.

[97] Shmoolik Mangan and Uri Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980–11985, 2003.

[98] Erez Dekel, Shmoolik Mangan, and Uri Alon. Environmental selection of the feed-forward loop circuit in gene-regulation networks. *Physical biology*, 2(2):81, 2005.

[99] Javier Macía, Stefanie Widder, and Ricard Solé. Specialized or flexible feed-forward loop motifs: a question of topology. *BMC systems biology*, 3(1):84, 2009.

[100] Fabian Fröhlich, Philipp Thomas, Atefeh Kazeroonian, Fabian J Theis, Ramon Grima, and Jan Hasenauer. Inference for stochastic chemical kinetics using moment equations and system size expansion. *PLoS computational biology*, 12(7):e1005030, 2016.

[101] Ramon Grima. An effective rate equation approach to reaction kinetics in small volumes: Theory and application to biochemical reactions in nonequilibrium steady-state conditions. *The Journal of Chemical Physics*, 133(3):035101, 2010.

[102] Ramon Grima, Philipp Thomas, and Arthur V Straube. How accurate are the nonlinear chemical fokker-planck and chemical langevin equations? *The Journal of chemical physics*, 135(8):084103, 2011.

[103] Ramon Grima. A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. *The Journal of chemical physics*, 136(15):04B616, 2012.

[104] Philipp Thomas, Arthur V Straube, Jens Timmer, Christian Fleck, and Ramon Grima. Signatures of nonlinearity in single cell noise-induced oscillations. *Journal of Theoretical Biology*, July 2013.

[105] Philipp Thomas, Christian Fleck, Ramon Grima, and Nikola Popović. System size expansion using feynman rules and diagrams. *Journal of Physics A: Mathematical and Theoretical*, 47(45):455007, 2014.

[106] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed April 21, 2022].

[107] Michel Le Bellac. *Quantum and Statistical Field Theory*. Clarendon Press, Oxford, 1991.

[108] Kirill A Datsenko, Ksenia Pougach, Anton Tikhonov, Barry L. Wanner, Konstantin Severinov, and Ekaterina Semenova. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nature Communications*, 3(1):945, jan 2012.

[109] Tim Künne, Sebastian N. Kieper, Jasper W Bannenberg, Anne I.M. Vogel, Willem R. Miellet, Misha Klein, Martin Depken, Maria Suarez-Diez, and S. J. J. Brouns. Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. *Molecular Cell*, 63(5):852–864, sep 2016.

[110] Jochem N.A. Vink, Koen J.A. Martens, Marnix Vlot, Rebecca E. McKenzie, Cristóbal Almendros, Boris Estrada Bonilla, Daan J.W. Brocken, Johannes Hohlbein, and S. J. J. Brouns. Direct Visualization of Native CRISPR Target Search in Live Bacteria Reveals Cascade DNA Surveillance Mechanism. *Molecular Cell*, 77(1):39–50.e10, jan 2020.

[111] Daan C. Swarts, Cas Mosterd, Mark W. J. van Passel, and S. J. J. Brouns. CRISPR Interference Directs Strand Specific Spacer Acquisition. *PLoS ONE*, 7(4):e35888, apr 2012.

[112] Olga Musharova, Vasily Sitnik, Marnix Vlot, Ekaterina Savitskaya, Kirill A Datsenko, Andrey Krivoy, Ivan Fedorov, Ekaterina Semenova, S. J. J. Brouns, and Konstantin Severinov. Systematic analysis of Type I-E Escherichia coli CRISPR-Cas PAM sequences ability to promote interference and primed adaptation. *Molecular Microbiology*, 111(6):1558–1570, jun 2019.

[113] Sungchul Kim, Luuk Loeff, Sabina Colombo, Slobodan Jergic, S. J. J. Brouns, and Chirlmin Joo. Selective loading and processing of prespacers for precise CRISPR adaptation. *Nature*, 579(7797):141–145, mar 2020.

[114] Chaoyou Xue and Dipali G. Sashital. Mechanisms of Type I-E and I-F CRISPR-Cas Systems in Enterobacteriaceae. *EcoSal Plus*, 8(2), jun 2019.

[115] Luuk Loeff, S. J. J. Brouns, and Chirlmin Joo. Repetitive DNA Reeling by the Cascade-Cas3 Complex in Nucleotide Unwinding Steps. *Molecular cell*, 70(3):385–394.e3, may 2018.

[116] Kaylee E. Dillard, Maxwell W. Brown, Nicole V. Johnson, Yibei Xiao, Adam Dolan, Erik Hernandez, Samuel D. Dahlhauser, Yoori Kim, Logan R. Myler, Eric V. Anslyn, Ailong Ke, and Ilya J. Finkelstein. Assembly and Translocation of a CRISPR-Cas Primed Acquisition Complex. *Cell*, 175(4):934–946.e15, nov 2018.

[117] James K. Nuñez, Lucas B. Harrington, Philip J. Kranzusch, Alan N. Engelman, and Jennifer A. Doudna. Foreign DNA capture during CRISPR-Cas adaptive immunity. *Nature*, 527(7579):535–8, nov 2015.

[118] Rodolphe Barrangou, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A. Romero, and Philippe Horvath. CRISPR provides acquired resistance against viruses in prokaryotes. *Science (New York, N.Y.)*, 315(5819):1709–12, mar 2007.

[119] Alexander Bolotin, Benoit Quinquis, Alexei Sorokin, and S Dusko Ehrlich. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, 151(8):2551–2561, aug 2005.

[120] S. J. J. Brouns, Matthijs M. Jore, Magnus Lundgren, Edze R. Westra, Rik J H Slijkhuis, Ambrosius P L Snijders, Mark J Dickman, Kira S Makarova, Eugene V. Koonin, and John van der Oost. Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science*, 321(5891):960–964, aug 2008.

[121] Ryan N. Jackson, Sarah M. Golden, Paul B.G. van Erp, Joshua Carter, Edze R. Westra, S. J. J. Brouns, John van der Oost, Thomas C. Terwilliger, Randy J. Read, and Blake Wiedenheft. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. *Science*, 345(6203):1473–1479, sep 2014.

[122] Hélène Deveau, Rodolphe Barrangou, Josiane E Garneau, Jessica Labonte, Christophe Fremaux, Patrick Boyaval, Dennis A. Romero, Philippe Horvath, and Sylvain Moineau. Phage Response to CRISPR-Encoded Resistance in Streptococcus thermophilus. *Journal of Bacteriology*, 190(4):1390–1400, feb 2008.

[123] Francisco J M Mojica, C Díez-Villaseñor, J García-Martínez, and C Almendros. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, 155(3):733–740, mar 2009.

[124] Edze R. Westra, Paul B.G. van Erp, Tim Künne, Shi Pey Wong, Raymond H. J. Staals, Christel L.C. Seegers, Sander Bollen, Matthijs M. Jore, Ekaterina Semenova, Konstantin Severinov, Willem M. de Vos, Remus T. Dame, Renko de Vries, S. J. J. Brouns, and John van der Oost. CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. *Molecular Cell*, 2012.

[125] Josiane E Garneau, Marie-Ève Dupuis, Manuela Villion, Dennis A. Romero, Rodolphe Barrangou, Patrick Boyaval, Christophe Fremaux, Philippe Horvath, Alfonso H. Magadán, and Sylvain Moineau. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, 468(7320):67–71, nov 2010.

[126] Ryan T. Leenay, Kenneth R. Maksimchuk, Rebecca A. Slotkowski, Roma N. Agrawal, Ahmed A. Gomaa, Alexandra E. Briner, Rodolphe Barrangou, and Chase L. Beisel. Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. *Molecular Cell*, 62(1):137–147, apr 2016.

[127] Ekaterina Semenova, Matthijs M. Jore, Kirill A Datsenko, Anna Semenova, Edze R. Westra, Barry L. Wanner, John van der Oost, S. J. J. Brouns, and Konstantin Severinov. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 108(25):10098–103, jun 2011.

[128] Peter C. Fineran, Matthias J H Gerritzen, María Suárez-Diez, Tim Künne, Jos Boekhorst, Sacha A F T van Hijum, Raymond H. J. Staals, and S. J. J. Brouns. Degenerate target sites mediate rapid primed CRISPR adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, 111(16):E1629–38, apr 2014.

[129] Anne M. Stringer, Lauren A. Cooper, Sujatha Kadaba, Shailab Shrestha, and Joseph T. Wade. Characterization of primed adaptation in the escherichia coli type I-E CRISPR-cas system, 2020.

[130] Ekaterina Semenova, Ekaterina Savitskaya, Olga Musharova, Alexandra Strotskaya, Daria Vorontsova, Kirill A Datsenko, Maria D. Logacheva, and Konstantin Severinov. Highly efficient primed spacer acquisition from targets destroyed by the Escherichia coli type I-E CRISPR-Cas interfering complex. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27):7626–31, jul 2016.

[131] Simon A. Jackson, Rebecca E. McKenzie, Robert D. Fagerlund, Sebastian N. Kieper, Peter C. Fineran, and S. J. J. Brouns. CRISPR-Cas: Adapting to change. *Science*, 356(6333):eaal5056, apr 2017.

[132] Addison V. Wright, Jun-Jie Liu, Gavin J. Knott, Kevin W. Doxzen, Eva Nogales, and Jennifer A. Doudna. Structures of the CRISPR genome integration complex. *Science*, 2017.

[133] Elizabeth Kutter, Daniel Bryan, Georgia Ray, Erin Brewster, Bob Blasdel, and Burton Guttman. From Host to Phage Metabolism: Hot Tales of Phage T4's Takeover of E. coli. *Viruses*, 10(7):387, jul 2018.

[134] John Davison. Pre-early functions of bacteriophage T5 and its relatives. *Bacteriophage*, 5(4):e1086500, oct 2015.

[135] Hannah G. Hampton, Bridget N.J. Watson, and Peter C. Fineran. The arms race between bacteria and their phage foes. *Nature*, 577(7790):327–336, jan 2020.

[136] Qiuyan Shao, Alexander Hawkins, and Lanying Zeng. Phage DNA Dynamics in Cells with Different Fates. *Biophysical Journal*, 108(8):2048–2060, apr 2015.

[137] Raymond H. J. Staals, Simon A. Jackson, Ambarish Biswas, S. J. J. Brouns, Chris M. Brown, and Peter C. Fineran. Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. *Nature communications*, 7(1):12853, nov 2016.

[138] Chaoyou Xue, Arun S. Seetharam, Olga Musharova, Konstantin Severinov, S. J. J. Brouns, Andrew J. Severin, and Dipali G. Sashital. CRISPR interference and priming varies with individual spacer sequences. *Nucleic Acids Research*, 43(22):10831–10847, dec 2015.

[139] Simon A. Jackson, Nils Birkholz, Lucía M. Malone, and Peter C. Fineran. Imprecise Spacer Acquisition Generates CRISPR-Cas Immune Diversity through Primed Adaptation. *Cell Host and Microbe*, 25(2):250–260.e4, feb 2019.

[140] John L. Spudich and D. E. Koshland. Non-genetic individuality: chance in the single cell. *Nature*, 262(5568):467–471, aug 1976.

[141] Stineke van Houte, Alice K. E. Ekroth, Jenny M. Broniewski, Hélène Chabas, Ben Ashby, Joseph Bondy-Denomy, Sylvain Gandon, Mike Boots, Steve Paterson, Angus Buckling, and Edze R. Westra. The diversity-generating benefits of a prokaryotic adaptive immune system. *Nature*, 532(7599):385–8, apr 2016.

[142] Corinna Richter, Ron L Dy, Rebecca E. McKenzie, Bridget N.J. Watson, Corinda Taylor, James T Chang, Matthew B McNeil, Raymond H. J. Staals,

and Peter C. Fineran. Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. *Nucleic acids research*, 42(13):8516–26, jul 2014.

[143] Lina Amlinger, Mirthe Hoekzema, E. Gerhart H. Wagner, Sanna Koskiniemi, and Magnus Lundgren. Fluorescent CRISPR Adaptation Reporter for rapid quantification of spacer acquisition. *Scientific Reports*, 7(1):10392, dec 2017.

[144] Nina Molin Høyland-Kroghsbo, Katrina Arcelia Muñoz, and Bonnie L Bassler. Temperature, by Controlling Growth Rate, Regulates CRISPR-Cas Activity in Pseudomonas aeruginosa. *mBio*, 9(6), nov 2018.

[145] C Díez-Villaseñor, Noemí M Guzmán, Cristóbal Almendros, J García-Martínez, and Francisco J M Mojica. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli. *RNA Biology*, 10(5):792–802, may 2013.

[146] Viktor Mamontov, Alexander Martynov, Natalia Morozova, Anton Bukatin, Dmitry B Staroverov, Konstantin A Lukyanov, Yaroslav Ispolatov, Ekaterina Semenova, and Konstantin Severinov. Long-term persistence of plasmids targeted by crispr interference in bacterial populations. *bioRxiv*, 2021.

[147] Jingwen Guan, Xu Shi, Roberto Burgos, and Lanying Zeng. Visualization of phage dna degradation by a type i crispr-cas system at the single-cell level. *Quantitative biology (Beijing, China)*, 5(1):67, 2017.

[148] Gert-Jan Kremers, Joachim Goedhart, Erik B. van Munster, and Theodorus W J Gadella. Cyan and Yellow Super Fluorescent Proteins with Improved Brightness, Protein Folding, and FRET Förster Radius † , ‡. *Biochemistry*, 45(21):6570–6580, may 2006.

[149] Daniel J. Kiviet, Philippe Nghe, Noreen Walker, Sarah Boulineau, Vanda Sunderlikova, and Sander J. Tans. Stochasticity of metabolism and growth at the single-cell level. *Nature*, 514(7522):376–379, sep 2014.

[150] Jonathan W. Young, James C W Locke, Alphan Altinok, Nitzan Rosenfeld, Tigran Bacarian, Peter S. Swain, Eric Mjolsness, and Michael B. Elowitz. Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy. *Nature Protocols*, 7(1):80–88, jan 2012.

[151] Martijn Wehrens, Dmitry Ershov, Rutger Rozendaal, Noreen Walker, Daniel Schultz, Roy Kishony, Petra Anne Levin, and Sander J. Tans. Size Laws and Division Ring Dynamics in Filamentous Escherichia coli cells. *Current Biology*, 28(6):972–979.e5, mar 2018.

[152] Joe H. Levine, Michelle E. Fontes, Jonathan Dworkin, and Michael B. Elowitz. Pulsed Feedback Defers Cellular Differentiation. *PLoS Biology*, 10(1):e1001252, jan 2012.

[153] Ido Yosef, Moran G Goren, and Udi Qimron. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Research*, 40(12):5569–5576, jul 2012.

[154] Alexander Martynov, Konstantin Severinov, and Iaroslav Ispolatov. Optimal number of spacers in CRISPR arrays. *PLOS Computational Biology*, 13(12):e1005891, dec 2017.

[155] Jaime Iranzo, Alexander E. Lobkovsky, Yuri I. Wolf, and Eugene V. Koonin. Evolutionary Dynamics of the Prokaryotic Adaptive Immunity System CRISPR-Cas in an Explicit Ecological Context. *Journal of Bacteriology*, 195(17):3834–3844, sep 2013.

[156] Serena Bradde, Marija Vucelja, Tiberiu Teşileanu, and Vijay Balasubramanian. Dynamics of adaptive immunity against phage in bacterial populations. *PLOS Computational Biology*, 13(4):e1005486, apr 2017.

[157] Lauren A. Cooper, Anne M. Stringer, and Joseph T. Wade. Determining the Specificity of Cascade Binding, Interference, and Primed Adaptation In Vivo in the Escherichia coli Type I-E CRISPR-Cas System. *mBio*, 9(2):e02100–17, apr 2018.

[158] Cheulhee Jung, John A. Hawkins, Stephen K. Jones, Yibei Xiao, James R. Rybarski, Kaylee E. Dillard, Jeffrey Hussmann, Fatema A. Saifuddin, Cagri A. Savran, Andrew D. Ellington, Ailong Ke, William H. Press, and Ilya J. Finkelstein. Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes on Next Generation Sequencing Chips. *Cell*, 170(1):35–47.e13, jun 2017.

[159] Stefan Klumpp. Growth-Rate Dependence Reveals Design Principles of Plasmid Copy Number Control. *PLoS ONE*, 6(5):e20403, may 2011.

[160] S. Lin-Chao and H. Bremer. Effect of the bacterial growth rate on replication control of plasmid pBR322 in Escherichia coli. *Molecular & general genetics : MGG*, 203(1):143–9, apr 1986.

[161] Hanne Ingmer, Christine Miller, and Stanley N. Cohen. The RepA protein of plasmid pSC101 controls Escherichia coli cell division through the SOS response. *Molecular Microbiology*, 42(2):519–526, oct 2001.

[162] Katarzyna H Maslowska, Karolina Makiela-Dzbenska, and Iwona J Fijalkowska. The sos system: a complex and tightly regulated response to dna damage. *Environmental and molecular mutagenesis*, 60(4):368–384, 2019.

[163] Tatiana Dimitriu, Elena Kurilovich, Urszula Lapinska, Konstantin Severinov, Stefano Pagliara, Mark D Szczelkun, and Edze R Westra. Bacteriostatic antibiotics promote the evolution of crispr-cas immunity. *bioRxiv*, 2021.

[164] Alexander P. Hynes, Manuela Villion, and Sylvain Moineau. Adaptation in bacterial CRISPR-Cas immunity can be driven by defective phages. *Nature communications*, 5(1):4399, jul 2014.

[165] Konstantin Severinov, Iaroslav Ispolatov, and Ekaterina Semenova. The Influence of Copy-Number of Targeted Extrachromosomal Genetic Elements on the Outcome of CRISPR-Cas Defense. *Frontiers in Molecular Biosciences*, 3, aug 2016.

[166] Thomas J. Nicholson, Simon A. Jackson, Bradley I. Croft, Raymond H. J. Staals, Peter C. Fineran, and Chris M. Brown. Bioinformatic evidence of widespread priming in type I and II CRISPR-Cas systems. *RNA Biology*, 16(4):566–576, apr 2019.

[167] Franklin L. Nobrega, Hielke Walinga, Bas E. Dutilh, and S. J. J. Brouns. Prophages are associated with extensive CRISPR-Cas auto-immunity. *Nucleic Acids Research*, 2020.

[168] Nathalie Q. Balaban, Jack Merrin, Remy Chait, Lukasz Kowalik, and Stanislas Leibler. Bacterial persistence as a phenotypic switch. *Science (New York, N.Y.)*, 305(5690):1622–5, sep 2004.

[169] Stephan Uphoff, Nathan D. Lord, Burak Okumus, Laurent Potvin-Trottier, David J. Sherratt, and Johan Paulsson. Stochastic activation of a DNA damage response causes cell-to-cell mutation rate variation. *Science*, 351(6277):1094–1097, mar 2016.

[170] Derek E. Moormeier, Jeffrey L. Bose, Alexander R. Horswill, and Kenneth W. Bayles. Temporal and Stochastic Control of Staphylococcus aureus Biofilm Development. *mBio*, 5(5), oct 2014.

[171] Murat Acar, Jerome T. Mettetal, and Alexander van Oudenaarden. Stochastic switching as a survival strategy in fluctuating environments. *Nature Genetics*, 40(4):471–475, apr 2008.

[172] Edze R. Westra, Stineke van Houte, Sam Oyesiku-Blakemore, Ben Makin, Jenny M. Broniewski, Alex Best, Joseph Bondy-Denomy, Alan Davidson, Mike Boots, and Angus Buckling. Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. *Current Biology*, 25(8):1043–1049, apr 2015.

[173] Alexandra Strotskaya, Ekaterina Savitskaya, Anastasia Metlitskaya, Natalia Morozova, Kirill A Datsenko, Ekaterina Semenova, and Konstantin Severinov. The action of Escherichia coli CRISPR-Cas system on lytic bacteriophages with different lifestyles and development strategies. *Nucleic acids research*, 2017.

[174] Bridget N.J. Watson, Reuben B. Vercoe, George P.C. Salmond, Edze R. Westra, Raymond H. J. Staals, and Peter C. Fineran. Type I-F CRISPR-Cas resistance against virulent phages results in abortive infection and provides population-level immunity. *Nature Communications*, 2019.

[175] Anna Lopatina, Nitzan Tal, and Rotem Sorek. Abortive Infection: Bacterial Suicide as an Antiviral Immune Strategy, 2020.

[176] Richard Moxon and Edo Kussell. The impact of bottlenecks on microbial survival, adaptation, and phenotypic switching in host–pathogen interactions. *Evolution*, 2017.

[177] Adrian G. Patterson, Simon A. Jackson, Corinda Taylor, Gary B. Evans, George P.C. Salmond, Rita Przybilski, Raymond H. J. Staals, and Peter C. Fineran. Quorum Sensing Controls Adaptive Immunity through the Regulation of Multiple CRISPR-Cas Systems. *Molecular Cell*, 2016.

[178] Luciano A Marraffini and Erik J Sontheimer. CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA. *Science*, 322(5909):1843–1845, dec 2008.

[179] Jen Nguyen, Juanita Lara-Gutiérrez, and Roman Stocker. Environmental fluctuations and their effects on microbial communities, populations and individuals. *FEMS Microbiology Reviews*, dec 2020.

[180] A. M. Spormann. Physiology of Microbes in Biofilms. In *Current Topics in Microbiology and Immunology*, pages 17–36. 2008.

[181] Kirill A Datsenko and Barry L. Wanner. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proceedings of the National Academy of Sciences*, 97(12):6640–6645, jun 2000.

[182] Stephan Uphoff, Rodrigo Reyes-Lamothe, Federico Garza De Leon, David J. Sherratt, and Achillefs N. Kapanidis. Single-molecule DNA repair in live bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 2013.

[183] Martin Ovesný, Pavel Křížek, Josef Borkovec, Zdeněk Švindrych, and Guy M. Hagen. ThunderSTORM: a comprehensive ImageJ plug-in for PALM

and STORM data analysis and super-resolution imaging. *Bioinformatics*, 30(16):2389–2390, aug 2014.

[184] Margaritis Voliotis, Philipp Thomas, Ramon Grima, and Clive G. Bowsher. Stochastic Simulation of Biomolecular Networks in Dynamic Environments. *PLOS Computational Biology*, 12(6):e1004923, jun 2016.

[185] Mitchell G Thompson, Nima Sedaghatian, Jesus F Barajas, Maren Wehrs, Constance B Bailey, Nurgul Kaplan, Nathan J Hillson, Aindrila Mukhopadhyay, and Jay D Keasling. Isolation and characterization of novel mutations in the psc101 origin that increase copy number. *Scientific reports*, 8(1):1–11, 2018.

[186] Andrey Krivoy, Marius Rutkauskas, Konstantin Kuznedelov, Olga Musharova, Christophe Rouillon, Konstantin Severinov, and Ralf Seidel. Primed crispr adaptation in escherichia coli cells does not depend on conformational changes in the cascade effector complex detected in vitro. *Nucleic acids research*, 46(8):4087–4098, 2018.

[187] Peter A Meacock and Stanley N Cohen. Partitioning of bacterial plasmids during cell division: a cis-acting locus that accomplishes stable plasmid inheritance. *Cell*, 20(2):529–542, 1980.

[188] Bin Shao, Jayan Rammohan, Daniel A Anderson, Nina Alperovich, David Ross, and Christopher A Voigt. Single-cell measurement of plasmid copy number and promoter activity. *Nature communications*, 12(1):1–9, 2021.

[189] I. Golding, J. Paulsson, S.M. Zawilski, and E.C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123:1025–1036, 2005.

[190] Sattar Taheri-Araghi, Serena Bradde, John T Sauls, Norbert S Hill, Petra Anne Levin, Johan Paulsson, Massimo Vergassola, and Suckjoon Jun. Cell-size control and homeostasis in bacteria. *Current biology*, 25(3):385–391, 2015.

[191] Marko Djordjevic, Magdalena Djordjevic, and Konstantin Severinov. Crispr transcript processing: a mechanism for generating a large number of small interfering rnas. *Biology direct*, 7(1):1–11, 2012.

[192] Sy Redding, Samuel H Sternberg, Myles Marshall, Bryan Gibb, Prashant Bhat, Chantal K Guegler, Blake Wiedenheft, Jennifer A Doudna, and Eric C Greene. Surveillance and processing of foreign dna by the escherichia coli crispr-cas system. *Cell*, 163(4):854–865, 2015.

[193] Olga Musharova, Sofia Medvedeva, Evgeny Klimuk, Noemi Marco Guzman, Daria Titova, Victor Zgoda, Anna Shiriaeva, Ekaterina Semenova, Konstantin Severinov, and Ekaterina Savitskaya. Prespacers formed during primed

adaptation associate with the cas1–cas2 adaptation complex and the cas3 interference nuclease–helicase. *Proceedings of the National Academy of Sciences*, 118(22), 2021.

[194] Rebecca E McKenzie, Emma M Keizer, Jochem NA Vink, Jasper van Lopik, Ferhat Büke, Vera Kalkman, Christian Fleck, Sander J Tans, and Stan JJ Brouns. Single cell variability of crispr-cas interference and adaptation. *bioRxiv*, 2021.

[195] Chaoyou Xue, Natalie R Whitis, and Dipali G Sashital. Conformational control of cascade interference and priming activities in crispr immunity. *Molecular cell*, 64(4):826–834, 2016.

[196] Robert D Fagerlund, Max E Wilkinson, Oleg Klykov, Arjan Barendregt, F Grant Pearce, Sebastian N Kieper, Howard WR Maxwell, Angela Capolupo, Albert JR Heck, Kurt L Krause, et al. Spacer capture and integration by a type if cas1–cas2-3 crispr adaptation complex. *Proceedings of the National Academy of Sciences*, 114(26):E5122–E5128, 2017.

[197] César Díez-Villaseñor, Noemí M Guzmán, Cristóbal Almendros, Jesús García-Martínez, and Francisco JM Mojica. Crispr-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among crispr-cas ie variants of escherichia coli. *RNA biology*, 10(5):792–802, 2013.

[198] Atsushi Miyawaki, Asako Sawano, Takako Kogure, et al. Lighting up cells: labelling proteins with fluorophores. *Nat Cell Biol*, 2003.

[199] Alexander P Browning, David J Warne, Kevin Burrage, Ruth E Baker, and Matthew J Simpson. Identifiability analysis for stochastic differential equation models in systems biology. *Journal of the Royal Society Interface*, 17(173):20200652, 2020.

[200] H. M. Meyer and A. H. Roeder. Stochasticity in plant cellular growth and patterning. *Front. Plant Sci.*, 5, 2014.

[201] A. Raj and A. Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135, 2008.

[202] L. Hong. Variable cell growth yields reproducible organdevelopment through spatiotemporal averaging. *Dev. Cell*, 38, 2016.

[203] J. Elsner, M. Michalski, and D. Kwiatkowska. Spatiotemporal variation of leaf epidermal cell growth: a quantitative analysis of arabidopsis thaliana wild-type and triple cyclind3 mutant plants. *Ann. Bot.*, 109, 2012.

[204] Adrienne HK Roeder, Vijay Chickarmane, Alexandre Cunha, Boguslaw Obara, BS Manjunath, and Elliot M Meyerowitz. Variability in the control of cell division underlies sepal epidermal patterning in arabidopsis thaliana. *PLoS biology*, 8(5):e1000367, 2010.

[205] Heather M Meyer, José Teles, Pau Formosa-Jordan, Yassin Refahi, Rita San-Bento, Gwyneth Ingram, Henrik Jönsson, James CW Locke, and Adrienne HK Roeder. Fluctuations of the transcription factor atml1 generate the pattern of giant cells in the arabidopsis sepal. *Elife*, 6:e19131, 2017.

[206] R. D. Dar. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc. Natl Acad. Sci. USA*, 109, 2012.

[207] S. Itzkovitz. Single-molecule transcript counting of stem-cell markers in the mouse intestine. *Nat. Cell Biol.*, 14, 2012.

[208] J. R. Newman. Single-cell proteomic analysis of s. cerevisiae reveals the architecture of biological noise. *Nature*, 441, 2006.

[209] A. Pare. Visualization of individual scr mrnas during drosophila embryogenesis yields evidence for transcriptional bursting. *Curr. Biol.*, 19, 2009.

[210] A. Sanchez and I. Golding. Genetic determinants and cellular constraints in noisy gene expression. *Science*, 342, 2013.

[211] Y. Taniguchi. Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329, 2010.

[212] R. Ietswaart, S. Rosa, Z. Wu, C. Dean, and M. Howard. Cell-size-dependent transcription of flc and its antisense long non-coding rna coolair explain cell-to-cell expression variation. *Cell Syst.*, 4, 2017.

[213] J. O. Brunkard, A. M. Runkel, and P. C. Zambryski. The cytosol must flow: intercellular transport through plasmodesmata. *Curr. Opin. Cell Biol.*, 35, 2015.

[214] A. Sigal. Variability and memory of protein levels in human cells. *Nature*, 444, 2006.

[215] N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz. Gene regulation at the single-cell level. *Science*, 307, 2005.

[216] D. W. Austin. Gene network shaping of inherent noise spectra. *Nature*, 439, 2006.

[217] J. M. Raser and E. K. O'Shea. Noise in gene expression: origins, consequences, and control. *Science*, 309, 2005.

[218] H. Tsutsui, S. Karasawa, H. Shimizu, N. Nukina, and A. Miyawaki. Semi-rational engineering of a coral fluorescent protein into an efficient highlighter. *EMBO Rep.*, 6, 2005.

[219] K. M. Leung. Asymmetrical beta-actin mrna translation in growth cones mediates attractive turning to netrin-1. *Nat. Neurosci.*, 9, 2006.

[220] B. Munsky, G. Neuert, and A. Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336, 2012.

[221] K. M. Crawford and P. C. Zambryski. Subcellular localization determines the availability of non-targeted proteins to plasmodesmatal transport. *Curr. Biol.*, 10, 2000.

[222] N. Maheshri and E. K. O'Shea. Living with noisy genes: how cells function reliably with inherent variability in gene expression. *Annu. Rev. Biophys. Biomol. Struct.*, 36, 2007.

[223] J. E. Melaragno, B. Mehrotra, and A. W. Coleman. Relationship between endopolyploidy and cell size in epidermal tissue of arabidopsis. *Plant Cell*, 5, 1993.

[224] G. Jovtchev, V. Schubert, A. Meister, M. Barow, and I. Schubert. Nuclear dna content and nuclear and cell volume are positively correlated in angiosperms. *Cytogenet. Genome Res.*, 114, 2006.

[225] C. G. Bowsher and P. S. Swain. Identifying sources of variation and the flow of information in biochemical networks. *Proc. Natl Acad. Sci. USA*, 109, 2012.

[226] S. Kalve, J. Fotschki, T. Beeckman, K. Vissenberg, and G. T. Beemster. Three-dimensional patterns of cell division and expansion throughout the development of arabidopsis thaliana leaves. *J. Exp. Bot.*, 65, 2014.

[227] Vladislav V Verkhusha, Irina M Kuznetsova, Olesia V Stepanenko, Andrey G Zaraisky, Michail M Shavlovsky, Konstantin K Turoverov, and Vladimir N Uversky. High stability of discosoma dsred as compared to aequorea egfp. *Biochemistry*, 42(26):7879–7884, 2003.

[228] I. Kim, F. D. Hempel, K. Sha, J. Pfluger, and P. C. Zambryski. Identification of a developmental transition in plasmodesmatal function during embryogenesis in arabidopsis thaliana. *Development*, 129, 2002.

[229] M. Pesch, I. Schultheiß, S. Digiuni, J. F. Uhrig, and M. Hülskamp. Mutual control of intracellular localisation of the patterning proteins atmyc1, gl1 and try/cpc in arabidopsis. *Development*, 140, 2013.

[230] A. M. Davis, A. Hall, A. J. Millar, C. Darrah, and S. J. Davis. Protocol: Streamlined sub-protocols for floral-dip transformation and selection of transformants in arabidopsis thaliana. *Plant Methods*, 5, 2009.

[231] Vahid Shahrezaei and Peter S Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–17261, 2008.

[232] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1964.

[233] Anjan Roy and Stefan Klumpp. Simulating genetic circuits in bacterial populations with growth heterogeneity. *Biophysical journal*, 114(2):484–492, 2018.

[234] Alice B Fulton. How crowded is the cytoplasm? *Cell*, 30(2):345–347, 1982.

[235] R John Ellis. Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Current opinion in structural biology*, 11(1):114–119, 2001.

[236] Szilvia Barsi and Bence Szalai. Modeling in systems biology: Causal understanding before prediction? *Patterns*, 2(6):100280, 2021.

[237] Mark K Transtrum and Peng Qiu. Bridging mechanistic and phenomenological models of complex biological systems. *PLoS computational biology*, 12(5):e1004915, 2016.

[238] Brian Munsky, Brooke Trinh, and Mustafa Khammash. Listening to the noise: random fluctuations reveal gene network parameters. *Molecular systems biology*, 5(1):318, 2009.

[239] Christoph Zechner, Michael Unger, Serge Pelet, Matthias Peter, and Heinz Koeppl. Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nature methods*, 11(2):197–202, 2014.

[240] Christoph Zechner, Jakob Ruess, Peter Krenn, Serge Pelet, Matthias Peter, John Lygeros, and Heinz Koeppl. Moment-based inference predicts bimodality in transient gene expression. *Proceedings of the National Academy of Sciences*, 109(21):8340–8345, 2012.

[241] Zhixing Cao and Ramon Grima. Accuracy of parameter estimation for auto-regulatory transcriptional feedback loops from noisy data. *Journal of The Royal Society Interface*, 16(153):20180967, 2019.

[242] Artémis Llamosi, Andres M Gonzalez-Vargas, Cristian Versari, Eugenio Cinquemani, Giancarlo Ferrari-Trecate, Pascal Hersen, and Gregory Batt. What population reveals about individual cell identity: single-cell parameter estimation of models of gene expression in yeast. *PLoS computational biology*, 12(2):e1004706, 2016.

[243] Irena Kuzmanovska, Andreas Milias-Argeitis, Jan Mikelson, Christoph Zechner, and Mustafa Khammash. Parameter inference for stochastic single-cell dynamics from lineage tree data. *BMC systems biology*, 11(1):1–13, 2017.

[244] Georgeos Hardo and Somenath Bakshi. Challenges of analysing stochastic gene expression in bacteria using single-cell time-lapse experiments. *Essays in Biochemistry*, 65(1):67–79, 2021.

[245] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

[246] Andreas Hilfinger, Mark Chen, and Johan Paulsson. Using temporal correlations and full distributions to separate intrinsic and extrinsic fluctuations in biological systems. *Physical review letters*, 109(24):248104, 2012.

[247] Jie Lin and Ariel Amir. Disentangling intrinsic and extrinsic gene expression noise in growing cells. *Physical Review Letters*, 126(7):078101, 2021.

[248] Euan Joly-Smith, Zitong Jerry Wang, and Andreas Hilfinger. Inferring gene regulation dynamics from static snapshots of gene expression variability. *Physical Review E*, 104(4):044406, 2021.

[249] Nils Eling, Michael D Morgan, and John C Marioni. Challenges in measuring and understanding biological noise. *Nature Reviews Genetics*, 20(9):536–548, 2019.

[250] CJ Zopf, Katie Quinn, Joshua Zeidman, and Narendra Maheshri. Cell-cycle dependence of transcription dominates noise in gene expression. *PLoS computational biology*, 9(7):e1003161, 2013.

[251] Oded Sandler, Sivan Pearl Mizrahi, Noga Weiss, Oded Agam, Itamar Simon, and Nathalie Q Balaban. Lineage correlations of single cell division time as a probe of cell-cycle dynamics. *Nature*, 519(7544):468–471, 2015.

[252] Shaon Chakrabarti, Andrew L Paek, Jose Reyes, Kathleen A Lasick, Galit Lahav, and Franziska Michor. Hidden heterogeneity and circadian-controlled cell fate inferred from single cell lineages. *Nature communications*, 9(1):1–13, 2018.

[253] Melanie L Bell, James B Earl, and Steven G Britt. Two types of drosophila r7 photoreceptor cells are arranged randomly: A model for stochastic cell-fate determination. *Journal of Comparative Neurology*, 502(1):75–85, 2007.

[254] Thomas M Norman, Nathan D Lord, Johan Paulsson, and Richard Losick. Stochastic switching of cell fate in microbes. *Annual review of microbiology*, 69:381–403, 2015.

[255] Guenter Vogt. Stochastic developmental variation, an epigenetic source of phenotypic diversity with far-reaching biological consequences. *Journal of Biosciences*, 40(1):159–204, 2015.

[256] Stavroula Skylaki, Oliver Hilsenbeck, and Timm Schroeder. Challenges in long-term imaging and quantification of single-cell dynamics. *Nature Biotechnology*, 34(11):1137–1144, 2016.

[257] Sasha F Levy, Naomi Ziv, and Mark L Siegal. Bet hedging in yeast by heterogeneous, age-correlated expression of a stress protectant. *PLoS biology*, 10(5):e1001325, 2012.

[258] Aaron M New, Bram Cerulus, Sander K Govers, Gemma Perez-Samper, Bo Zhu, Sarah Boogmans, Joao B Xavier, and Kevin J Verstrepen. Different levels of catabolite repression optimize growth in stable and variable environments. *PLoS biology*, 12(1):e1001764, 2014.

[259] Hubertus JE Beaumont, Jenna Gallie, Christian Kost, Gayle C Ferguson, and Paul B Rainey. Experimental evolution of bet hedging. *Nature*, 462(7269):90–93, 2009.

[260] M Olivia Casanueva, Alejandro Burga, and Ben Lehner. Fitness trade-offs and environmentally induced mutation buffering in isogenic c. elegans. *Science*, 335(6064):82–85, 2012.

[261] Ofer Fridman, Amir Goldberg, Irine Ronin, Noam Shoresh, and Nathalie Q Balaban. Optimization of lag time underlies antibiotic tolerance in evolved bacterial populations. *Nature*, 513(7518):418–421, 2014.

[262] Ben Lehner. Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Molecular systems biology*, 4(1):170, 2008.

Summary

The study of genetically identical cells frequently reveals that substantial variation exists between the expression levels of molecules, such as mRNAs and proteins, with respect to time in individual cells and across a population. This is due to intrinsic noise arising from the random timing of biochemical reactions in the gene regulatory network. An additional source of noise in biological systems comes from the interaction of unknown molecular components with the network. These interaction come from the cell's changing environment, from upstream processes, or from fluctuations associated with cell growth and division, and are collectively termed extrinsic noise. Mathematical modelling has been used as a means to understand the factors that contribute to the resulting stochasticity in the gene expression dynamics of cells. As many molecular species are present in small numbers, models based on ordinary differential equations that describe the time evolution of the mean values of these species do not accurately capture the system behaviour. For this reason, we require a stochastic description of these biochemical systems. **Chapter 1** of this thesis gives a general introduction into the sources of gene expression noise and non-genetic cell-to-cell variability. To study how biological processes affect the fluctuations of molecule numbers in cells, the chemical master equation (CME) and the stochastic simulation algorithm (SSA) are widely used to model the reaction kinetics inside cells. Throughout **Chapters 2–5** of this thesis, we investigate cell-to-cell variability in a wide range of biological systems through the integration of experimental approaches and mathematical modelling.

In **Chapter 2**, we derive a novel approximate method to obtain closed-form expressions for the means, variances, and power spectra of the molecular species within a chemical reaction network with intrinsic and slowly changing extrinsic noise. We do this by extending the conventional linear-noise approximation (LNA) to include systems where extrinsic noise manifests itself as fluctuations in the reaction rates. These extrinsic fluctuations are assumed to have a longer timescale, *e.g.* corresponding to the cell cycle period, compared to the typically fast intrinsic reaction processes. We verify the accuracy of the theory by its application to different models of gene regulatory networks, and comparing the analytical predictions to the more computationally costly results produced through stochastic simulations. Our results show that the effect of extrinsic noise on the means and variances of

molecule number fluctuations is dependent on the affected reaction rates, and that negative feedback control can suppress gene expression noise in the presence of environmental perturbations. Furthermore, we demonstrate how information flow between components in a gene regulatory network can be affected by extrinsic noise. Finally, we give an example of how the framework can be applied to aid in the design of robust synthetic circuits.

In **Chapter 3**, we study the bacterial adaptive immune system CRISPR-Cas in a growing population of *Escherichia coli*. In order to survive infection, bacteria are forced to obtain an immunological memory of past infections through a process termed priming, in order to enable recognition of previously encountered threats. In addition, the cell contains an operon of CRISPR-associated (Cas) proteins which are responsible for finding and eliminating these invaders through a process called interference. We use time-lapse microscopy in combination with microfluidics to obtain single-cell lineage data throughout the entire duration of CRISPR defence. For the first time, we quantify the variation that exists between cells in how fast they are able to respond to foreign mobile genetic elements. Clearance of previously encountered invaders through CRISPR interference is fast with a narrow distribution. However, invaders can accumulate mutations in the PAM region which allows them to escape direct interference, resulting in large cell-to-cell variability of clearance times. Further analysis of the experimental data, together with a specially developed agent-based stochastic framework which simulates the behaviour of the bacterial population, allow us to identify the acquisition of a new immunological memory (adaptation) as the source of the increased variation in priming. Statistical analysis of cell lineage features reveals that faster growth and cell division, as well as higher levels of the CRISPR surveillance complex Cascade, increase the probability of plasmid clearance by interference. In contrast, slower growth is associated with a higher rate of adaptation. Through mathematical modelling we estimate the influence of target and Cascade copy numbers, as well as Cascade binding affinity of the rate of priming. Our results show that the ability to adapt to an invading threat by primed CRISPR adaptation is highly stochastic, implying that only subpopulations of bacteria are able to respond to foreign invaders in a timely manner.

It has been shown that mutations in the protospacer adjacent motif (PAM), which flanks the DNA sequence targeted by the surveillance complex, affect target recognition by Cascade. In **Chapter 3**, we found that this reduced binding affinity of Cascade to targets with escaping PAM mutations can explain the observed wide distribution of plasmid loss times. However, the exact mechanism by which Cascade initiates primed adaptation, the acquisition of a new immunological memory, is largely unknown. In **Chapter 4**, we investigate the primed adaptation mechanism further by characterising the dynamics of CRISPR-mediated target clearance for three different PAM variants. We use the same microfluidics set-up as in the previous chapter, and compare the experimental single-cell lineages to simulated trajectories from two mechanisms of primed adaptation proposed in the literature

by adapting the reaction mechanism of the agent-based framework developed in the previous chapter. We show that features of the data are consistent with the interference-independent model for adaptation, in which a primed adaptation complex is responsible for the acquisition of new immunological memories. Our results show that the CRISPR-response of *E. coli* depends strongly on the PAM variant, which suggests these bacteria might employ a strategy to balance the relative rates of interference and the acquisition of new memories. This would lead to a diversified response to invaders, thus increasing the population's chance of survival.

While for bacteria the notion of cell-to-cell variability to enable bet-hedging strategies is sensible, for multicellular organisms reproducible and coordinated development might seem more important. Plants have evolved regulatory mechanisms to achieve specialised cell types and robust tissue growth. Although plant development is highly reproducible, some developmental stochasticity exists. In **Chapter 5** we quantify the noisiness of stochastic gene expression in *Arabidopsis thaliana* at the cellular level. To this end, we employ a combination of experimental and modelling approaches. First, we use the photoconvertible KikGR marker to show that the protein expressions of individual cells fluctuate over time. A dual reporter system is then used to study extrinsic and intrinsic noise. This reveals that extrinsic noise is the main source of protein variability in both young and old rosette leaves, and that extrinsic noise in stomata is clearly lower in comparison to several other cell types. Finally, through spatial analysis we show that cells are coupled with respect to stochastic protein expression in young leaves, hypocotyls and roots but not in mature leaves. Through theoretical analysis we find that the observed spatial correlation between cells can only partially be explained by the inheritance of mRNA and protein from a shared ancestor, which suggests other extrinsic noise sources affect the cellular dynamics.

The results from **Chapters 2–5** are discussed in a broader context in **Chapter 6**. In this general discussion, experimental and computational challenges relating to the study of gene expression noise are reviewed. Furthermore, I discuss possible implications of stochastic gene expression for phenotypic variability.

Acknowledgements

"One day I will find the right words, and they will be simple."

Until then, I will use many words in an attempt to express my gratitude to those who have supported me during my PhD. This thesis deals with how organisms manage to navigate and sometimes even adapt to challenging and selective circumstances. Throughout my PhD, I feel lucky to be able to say that the people around me have provided me with a stabilising environment at times when robust functioning was required.

I would like to thank my promotors for giving me the opportunity to conduct the research described in this thesis. **Vitor**, I've greatly enjoyed my time at SSB, where I feel I could develop myself as a researcher in a stimulating environment. **Jaap**, bedankt voor het vertrouwen waarmee je mij en Anna bij Biometris hebt verwelkomd, en de aanmoedigingen en goede adviezen die je gaf bij elk van onze meetings. Deze dingen hebben alle verschil gemaakt in zowel de overgangsperiode als bij de afronding van het promotietraject. Thanks also to my co-promotor and supervisor, **Christian**, for your guidance and the many helpful discussions we had. I recall you said to me at some point: 'Planning is everything, a plan is nothing'. It is certainly true that not many things during my PhD went according to plan, but I am grateful you were able to share your view on research with me over the years.

I want to thank all my colleagues from SSB, past and present. So many of you have been generous with your time and your support. Even though you probably only understood about 10% of my presentations I felt this support in many other ways, whether it was in the form of peer-pressured coffee breaks, outside lunches, after work drinks, dinners, and parties which provided very welcome distractions. I have especially fond memories of my time at the Dreijen with the old core of the 'super sexy beasts'/MIB, and the PhD trip. **Rob, Bastian, Maarten, Anna, Rik, Benoit, Ruben, Dorett, Nikolas, Nong, Niels, Bart, Niru, Maria, Peter, Edoardo, Jeroen, Gerben, Martijn, Wen, Emmy, Catalina, Linde, Erika, Nhung, Melanie** and many more that have temporarily slipped my mind, thanks for everything.

I'm also very grateful to the Biometris staff for their warm welcome and their help and support in teaching. I haven't met a single unfriendly person here, which

explains the relaxed atmosphere in the group.

To my office mates at de Dreijen, Helix, and Radix: **Rob, Tjerko, Nikolas, Anna, Rik, Erika, Bob, Patrick, Jip, Bader,** and **Wenhao**: I feel fortunate I could always talk to you about work, but especially also about various non-work related topics (I remember subjects varied quite wildly). Thanks for entertaining my random thoughts and putting up with me eating at all times. **Anna**, you were certainly my office mate for the longest stretch, and it was always great to share my love for cycling, music, and especially cat pics with you.

Rob, my career in research started with you when you supervised my MSc thesis. Later, your help during my PhD was invaluable and without it I would not be at this point today. We bonded over our (excellent) music tastes, and I'm happy we got to share this love for music at the festivals and concerts we visited at various places in the world. I know I can count on you to always have my back. Thank you for being a great teacher, and more importantly a great friend.

Dr. **Becca**, I guess we are both doctors now! I couldn't have imagined a better person to work on my last and probably most challenging PhD project with. It has taught me so much, and thankfully you were always willing to suffer through it with me and answer my endless stream of questions about CRISPR, weird cells, microscope stuff, and science in general. Your positive energy will be missed on the day of my defence, and though I'm sad you can't be here to celebrate I know you and Jochem are having a blast back in New Zealand.

Aan al mijn homies van de Molenstraat 14, de locatie van de mooiste verzameling rotzooi in heel Wageningen, jullie hebben vele jaren mijn thuisfront en mijn vangnet gevormd. **Alexandra, Allard, Anne, Carsten, David, Emiel, Heleen, Jesse, Jorrit, Mark, Mylo**, adoptie-huisgenoot **Robin, Wouter**, en natuurlijk **Poekie** en **Stufi**. Bedankt voor de therapeutische banksessies, etentjes in het park, BBQs met pyromane trekjes, het altijd gezellige kippenhok (incl. kippen op stok), het leed, het vermaak, het leedvermaak, de noodgedwongen boost voor het immuunsysteem, de 5 mei chaos, spelletjesavonden met en zonder drakentorens, spontane doordeweekse feestjes, huisexcursies naar verre oorden, nieuwe culinaire ervaringen, en de eindeloze stroom aan kleurrijke mensen die jullie altijd meebrachten. Oost west, asbest.

In het bijzonder wil ik jullie nog noemen, **Heleen, Mylo**, en **Anne**: dankzij jullie weet ik wat het is om je onvoorwaardelijk gesteund te voelen. En ook hoe het is om te lachen tot de tranen over je wangen lopen. Ook al lachen jullie niet altijd even hard met me mee, een matig gevoel voor humor is jullie snel vergeven als je verder zo fantastisch bent. Ik bewonder hoe jullie in het leven staan.

Annemiek: ook jij was korte tijd mijn 'huisgenoot'. Naast dat je een enorme inspiratie bent op de fiets en qua doorzettingsvermogen, heb je ook nog eens een heel groot hart. Dankjewel dat je je huis voor me openstelde, het heeft ontzettend geholpen in de laatste hectische maanden van mijn PhD.

Floor, Jolijn, Kirsten, Lindsay, Lizzy en **Tessa**: ik zie ons nog staan tussen de kluisjes, tijdens de ontelbare pauzes in de eeuwigheid die middelbare school heet.

Samen deze vormende periode doorkomen schept een bepaalde band. Ondanks onze vele studies, verhuizingen en banen hebben we altijd contact gehouden en ik ben heel erg blij dat we nog steeds regelmatig de tijd maken om elkaar te zien. Stuk voor stuk zijn jullie mooie, lieve, slimme vrouwen met veel verschillende passies en talenten. Ik ben trots op de mooie dingen die jullie bereikt hebben sinds de Spieringshoek-dagen.

Arlette, jij bent een van mijn oudste vriendinnen en the coolest girl I know. Ik leerde je kennen als een felle debater die weet wat ze wil, en nog altijd ga je altijd recht op je doelen af. Jou wil je aan je kant hebben, al is het maar om deel te kunnen nemen aan je creatieve en spontane plannen, en gelukkig voor mij zitten we in hetzelfde team. Ik heb super veel mooie herinneringen aan alle plekken waar we elkaar hebben opgezocht door de jaren heen, en kijk enorm uit naar onze aankomende trip (die we nu toch echt moeten gaan plannen).

Bregje, mijn roei-, blessure-, PhD- en fietsmaatje. Ik benijd je om hoe je situaties (en jezelf) altijd met door humor gemaskeerde wijsheid benadert, een gouden combinatie die enorm relativerend werkt. Ik heb veel gehad aan je steun, adviezen, en nuchtere kijk op dingen de afgelopen jaren. Dankjewel dat je de innerlijke schoonheid kunt zien zelfs (juist) in onze lelijkste foto's.

Laurens, jij hebt het unieke talent om elke plek aan te laten voelen als een bruine kroeg, waar er altijd een luisterend oor is en een schijnbaar eindeloze voorraad bier. We kunnen altijd makkelijk switchen tussen diepe discussies en onzin, een evenwicht dat naarmate de avond vordert altijd wat verder naar dat laatste wordt verschoven. Hoewel ik vaak goede verhalen hoor over je kookkunsten, krijg je het toch altijd voor elkaar dat ik uiteindelijk pannenkoeken voor je sta te bakken: ook dat is een talent.

Melina, dankjewel voor de gezelligheid, voor de leuke logeerpartijtjes, knuffels met Ollie, en je steun in moeilijke tijden. Je bent creatief, slim, en gedreven, en ik weet zeker dat je wel je weg gaat vinden in én na je PhD.

Mijn adoptie-nichtjes **Fenny, Noortje, Sianne, Eline** en **Franka**, jullie zijn me zo lief. Als ik eraan denk hoe lang ik jullie al ken (mijn hele leven) voel ik me bijna oud, maar alle herinneringen aan onze vaak idiote (maar nog vaker succesvolle) plannen, kampeervakanties, knutselsessies en sinterklaasweekendjes doen me dat gelukkig per direct weer vergeten.

Aan de mannen en vrouwen van **Toerclub Wageningen**: het was heerlijk om bij tijden de benen eens goed vol te laten lopen (met melkzuur welteverstaan) en tegelijkertijd het hoofd leeg te maken. Bedankt voor het vele kopwerk, de af en toe zeer welkome duwtjes, en de altijd gezellige borrel na afloop.

Sjoerd, dankjewel voor je liefde en je geduld. De vele avonturen samen met jou hebben me voor zover mogelijk gegrond gehouden tijdens de soms pittige PhD-jaren.

Lieve zus, lieve **Sianne**, je bent mijn meest favoriete persoon op deze hele wereld. Ik heb het grootste geluk dat ik zij aan zij met jou heb mogen opgroeien,

en dat we nog steeds alle mooie en ook moeilijke momenten kunnen delen. Jij en **Robert** staan altijd voor iedereen klaar en daar bewonder ik jullie om. Het is zo mooi om te zien met hoeveel liefde en plezier jullie **Flor**, mijn meest favoriete mini-mensje, grootbrengen; het is geen wonder dat hij zoveel lacht met jullie als ouders.

Lieve papa en mama, mijn rijke opvoeding is het grootste goed wat jullie me hebben gegeven, en een van de meest kostbare dingen die ik bezit. Jullie hebben me altijd gestimuleerd en gefaciliteerd om het beste uit mezelf te halen, en doen dat nog steeds door middel van de steun die ik dagelijks voel. Bedankt dat ik bij jullie altijd kan thuiskomen.

Lastly, to those who I have not mentioned by name, if you are reading this you have likely been in one way or another a part of this journey. Whether it was at the beginning, at the end, for only for a short time, or throughout, I feel most grateful for the company and the friendships I have gained along the way.

About the author

Emma Mathilde Keizer was born on the 23rd of August 1991 in Vlaardingen, the Netherlands. She attended SG Spieringshoek in Schiedam, where she obtained her gymnasium degree *cum laude* in 2009. During her childhood, Emma enjoyed playing the harp and climbing in trees. At secondary school, she participated in the Model European Parliament, a simulation of the working of the European Parliament for students, first as a delegate and in later years as chair. Plagued by a wide range of interests, she opted to study at University College Roosevelt in Middelburg. Here on the canals of Walcheren, she developed a love of rowing. She specialised in Mathematics and Physics and obtained her BSc degree *cum laude* in 2012. She went on to study a master in Applied Mathematics at the Katholieke Universiteit Leuven in Belgium, and after one year she decided to continue her studies at Wageningen University where she obtained her MSc degree in Bioinformatics, specialising in Systems Biology. After an internship with the Stochastic modelling group of Ramon Grima at the University of Edinburgh, she started her PhD research at Wageningen university with the Systems & Synthetic Biology group. After three years, she joined the Mathematical and Statistical methods group, where her PhD research continued with a collaborative project between TU Delft, AMOLF and Wageningen University. After her PhD, Emma will join the Physical and Organic Chemistry group of prof. dr. Wilhelm Huck as a postdoctoral researcher.

Emma enjoys working within multi-disciplinary, international teams combining experimental research with mathematical modelling. Her projects have focused on the design of stochastic models of biological systems and the analysis of single-

cell data. Besides scientific research, she developed experience in supervising (student) projects, teaching BSc and MSc courses, and mentoring student teams as a supervisor of the Wageningen UR iGEM (international Genetically Engineered Machine) team, the World's largest international synthetic biology competition for student teams in Boston (USA). Outside of work, Emma enjoys spending her time on the bike, cycling on the road, in the forest, or during bikepacking holidays abroad.

List of publications

This thesis

I.S. Araújo, J.M. Pietsch, **E.M. Keizer**, B. Greese, R. Balkunde, C. Fleck, and M. Hülskamp. Stochastic gene expression in *Arabidopsis thaliana*. *Nature communications*, 8(1), 1-9, 2017.

E.M. Keizer†, B. Bastian†, R.W. Smith, R. Grima, and C. Fleck. Extending the linear-noise approximation to biochemical systems influenced by intrinsic noise and slow lognormally distributed extrinsic noise. *Physical Review E*, 99(5), 052417, 2019.

R.E. McKenzie†, **E.M. Keizer**†, J.N.A. Vink, J. van Lopik, F. Büke, V. Kalkman, C. Fleck, S.J. Tans and S.J.J. Brouns. Single Cell Variability of CRISPR-Cas Interference and Adaptation. *Molecular Systems Biology*, 2022 *(accepted)*.

Other publications

M.A. Prusicki, **E.M. Keizer**, R.P. van Rosmalen, S. Komaki, F. Seifert, K. Müller, E. Wijnker, C. Fleck, and A. Schnittger. Live cell imaging of meiosis in *Arabidopsis thaliana*. *Elife*, 8, e42834, 2019.

M.A. Prusicki, **E.M. Keizer**, R.P. van Rosmalen, C. Fleck, and A. Schnittger. Live Cell Imaging of Male Meiosis in *Arabidopsis* by a Landmark-based System. *Bio-protocol*, 10(9), 2020.

D. Lähnemann, J. Köster, E. Szczurek, D.J. McCarthy, S.C. Hicks, M.D. Robinson, ... and A. Schönhuth. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1), 1-35, 2020.

†These authors contributed equally

Overview of completed training activities

Discipline specific activities	**Organised by**	**Year**
iGEM 2017 world jamboree (supervisor)	BioBricks foundation	2017
BioSB 2016	BioSB	2016
Uncertainty Propagation in Spatial Environmental Modeling	PE&RC	2016
International Workshop on Control Engineering and Synthetic Biology	SynBioControl2017 – SISOS	2017
SEB Florence	The Society for Experimental Biology	2018
Single Cell Data Science: Making Sense of Data from Billions of Single Cells	Lorentz Center	2018
CompSysBio 2019	Université de Lyon	2019
SWI 2019	SWI	2019
CRISPR 2021	Institut Pasteur	2021
SMB 2021	Society of Mathematical Biology	2021

General courses	**Organised by**	**Year**
VLAG PhD week	VLAG	2016
Competence Assessment	WGS	2016
Soft skills training with role play	Aude	2016
PhD Workshop Carousel	WGS	2017
Reviewing a scientific paper	WGS	2018
Career perspectives	WGS	2018
Presenting with impact	WGS	2018
Machine Learning for Research	eScience Center	2020

Optionals	**Organised by**	**Year**
Preparation of research proposal	SSB	2015
Weekly group meetings	SSB	2015–2019
Seminar series	SSB	2015–2019
PhD trip 2017	SSB / MIB	2017
SB&E colloquium series	Biometris	2019–2021
Modelling and Simulation discussion group	Biometris	2019

Colophon

Cover design by Emma Keizer

Lay-out by Emma Keizer and ProefschriftMaken

Printed by ProefschriftMaken