Stochatic interacting systems in biophysics : immunology and development

Jonathan Desponds

THÈSE DE DOCTORAT
de l'Université de recherche
Paris Sciences Lettres –
PSL Research University

préparée à
l'École Normale Supérieure

Systèmes stochastiques en interaction en biophysique : immunologie et développement

par Jonathan Desponds

École doctorale n°564
Spécialité : Physique
Soutenue le 22.09.2016

Composition du Jury :

M. Hervé Isambert
Institut Curie
Rapporteur

M^me^ Aleksandra Walczak
École Normale Supérieure
Directrice de thèse

M. Thierry Mora
École Normale Supérieure
Directeur de thèse

M^me^ Nathalie Dostatni
Institut Curie
Membre du Jury

M. Vincent Hakim
École Normale Supérieure
Membre du Jury

M. Martin Weigt
UPMC
Membre du Jury

ENS ÉCOLE NORMALE SUPÉRIEURE
PSL RESEARCH UNIVERSITY PARIS

Résumé

Nous présentons deux problèmes de biologie faisant appel à un traitement de données et des modèles issus de la physique statistique : la dynamique des populations en immunologie et la régulation génétique dans le développement embryonnaire. En immunologie, nous étudions le problème de la sélection somatique dans le système immunitaire adaptatif: la sélection cellulaire et la compétition qui s'y opèrent, constituant un système quasi Darwinien au sein de l'organisme. Dans un premier temps, nous considérons différentes hypothèses sur la dynamique selective : signaux déclenchant la division ou la mort cellulaire par liaison antigénique ou par cytokines, paramètres dynamiques de division, mort et fluctuations environnementales. Nous explorons leur influence sur la taille des clones dont la distribution à queue lourde a été observée à travers les espèces et les types de cellules. Deux familles de modèles émergent : un premier dans lequel le bruit est cohérent à l'échelle du clone et un second dans lequel le bruit varie de cellule à cellule. Nous montrons dans quelle mesure la distribution de taille de clones permet de déterminer le meilleur modèle et relions la forme de la distribution ainsi que l'exposant apparent de la loi de puissance aux paramètres biologiques. Dans un second temps, nous explorons les caractéristiques du réseau complexe et aléatoire formé par les clones et les antigènes : dimension, adjacence, dynamique. Nous nous intéressons à l'effet de la sélection dans le temps et à la vitesse d'évolution des clones.

La deuxième partie de cette thèse est consacrée au développement embryonnaire. Dans l'embryon, il est essentiel pour le noyau de déterminer sa position avec une grande précision pour orienter la différentiation et construire un organisme structuré viable. Cette information positionnelle est acquise, transmise et conservée par la diffusion de protéines et l'activation de circuits génétiques. Plus précisément, la formation de l'axe antéro-postérieur chez la Drosophile est déterminée entre autres par l'activation du gène *hunchback* par la protéine Bicoid. Nous analysons des données issues d'expériences d'imagerie fluorescente dynamique dans les premiers cycles cellulaires de l'embryon. Nous contruisons un modèle spécifique permettant d'analyser la fonction d'autocorrélation des traces temporelles de fluorescence qui prend en compte toutes les difficultés biologiques et expérimentales (bruit, calibration traces courtes, structure du gène artificiel) pour extraire les paramètre dynamiques d'activation de *hunchback*. Nous examinons différentes dynamiques potentielles (poisonnienne, markovienne ou non markovienne) et leur implication pour l'information dont la cellule dispose sur sa position ainsi que la précision de la lecture du gradient de Bicoid.

Mots clés : Immunologie, Développement, Dynamique des populations, Circuits génétiques

Abstract

This work presents two problems of biology requiring data analysis and models from statistical mechanics: population dynamics in immunology and gene regulation in embryo development. In immunology I study the problem of somatic evolution in the adaptive immune system: selection of and competition among cells that form a close-to-Darwinian system within one individual. First, I consider different potential hypotheses for selective dynamics: division and death signals through antigen binding or cytokines, dynamical parameters for division, death and fluctuations of the environment. I explore their impact on clone sizes. Experimentally, these clone sizes show heavy tail distributions for different species and different pools of cells. Two families of models emerge: models where noise is consistent at the level of the clone and models where it varies from cell to cell. I show how clone size distributions help discriminate between these models and relate the shape of the distribution and the exponent of the power law to biological parameters. Second, I explore the specifics of the complex stochastic network of clones and antigens: its dimensionality, connectivity and dynamics. I study the effect of selection at different time scales and the speed of evolution of the clones.

The second part of this dissertation concerns embryo development. In the fly embryo, it is crucial that nuclei can evaluate their position within the organism accurately to determine cell fate and build a healthy organism. This positional information is obtained, transferred, and maintained through diffusion of proteins and activation of genetic networks. More specifically, the patterning of the antero-posterior axis in drosophila requires the *hunchback* gene, activated by the Bicoid protein. I analyze data from fluorescent live imaging in the early cell cycles of the embryo. I build a tailor-made model to analyze autocorrelation functions of fluorescence time traces overcoming all biological and experimental challenges (noise, calibration, short traces, transgene construct) to extract the parameters of *hunchback* activation. I examine several potential types of dynamics for gene switiching (Poisson, Markovian or non-Markovian) and predict their impact on positional information and the accuracy of bicoid gradient readout.

Keywords : Immunology, Development, Population dynamics, Gene regulatory network

Remerciements

Je tiens en premier lieu à remercier mes directeurs de thèse Aleksandra Walczak et Thierry Mora sans qui la biophysique serait sans doute restée pour moi une science obscure, apanage de physiciens aventureux aux prises avec le vivant. C'est donc avec eux que j'ai découvert ce domaine à travers le cours de Master et la thèse qui a suivi. Je tiens à les remercier pour les innombrables discussions que nous avons eues (qu'il s'agisse de science ou de la qualité des différentes options gastronomiques de la rue Mouffetard), pour leur patience tandis qu'ils devaient m'enseigner la biologie, la physique statistique et les statistiques tout court, les méthodes numériques et surtout comment opérer ce miracle qui permet de mettre en oeuvre sur un système réel les belles théories que j'avais accumulées pendant ma scolarité. Je tiens particulièrement à les remercier pour leur présence, leur disponibilité à toute heure, leur efficacité et encore une fois leur infinie patience à travers les différentes périodes plus ou moins difficiles de cette thèse. Je tiens encore à leur exprimer ma gratitude pour m'avoir aidé à intégrer la communauté scientifique en biophysique à travers discussions et conférences et m'avoir donné les moyens de faire valoir mon travail (présentation, article) jusqu'à leur soutien pour l'obtention de mon post-doctorat.

J'ajoute bien sur que ca a été pour moi un plaisir de collaborer scientifiquement avec eux, que j'ai beaucoup appris de leur rigueur, de leur culture et de leur créativité.

Enfin, merci d'avoir fait vivre ce groupe avec votre énergie et votre enthousiasme : Aleksandra par son ironie et ses talents de conteuse d'histoires et Thierry par son humour et son esprit sportif en toutes circonstances (des pistes de ski des Houches aux sorties surf de Santa Barbara).

Je tiens à remercier mes deux rapporteurs Hervé Isambert et Ilya Nemenman pour leur lecture attentive du texte de thèse au beau milieu de l'été et la pertinence de leurs commentaires. Ils m'ont permis d'améliorer la qualité de mon texte de thèse et d'envisager de nouvelles directions de recherche. Je voudrais remercier Ilya dont nous n'avons pu organiser la venue pour ses nombreux commentaires et son sens du style. Je souhaite remercier les autres membres du jury Nathalie Dostatni, Vincent Hakim et Martin Weigt qui se sont aussi penchés sur le manuscrit en cette période de rentrée et acceptent de braver la menace des coupures d'électricité pour venir m'écouter à l'ENS.

Je voudrais dire un grand merci aux collaborateurs du projet qui a occupé la seconde partie de ma thèse sur la morphogénèse. A Nathalie pour avoir porté si énergiquement la partie expérimentale de ce projet et nous avoir éclairé de son expertise biologique, à Huy (sans qui nos résultats seraient restés bien théoriques) pour sa science de l'analyse de données et sa rigueur, à Teresa (qui nous a quittés trop tôt pour Jussieu) pour avoir été la première à mettre en place cette analyse, à Tanguy sans qui rien n'aurait été possible pour avoir fait les expériences et avoir su calmer les controverses, à Mathieu pour ses avis toujours éclairés et ses idées brillantes,

à Michaël pour ses connaissances computationnelles, à Philippine pour sa (trop) courte mais enthousiaste présence dans notre groupe et à tous les collaborateurs. Merci au group Axomorph puis Reflex pour de longs vendredis matins de fructueuse discussion !

Merci aux autres membres du LPT et du LPS pour les discussions que nous avons parfois pu avoir au fil des années, au labo, au hasard des bureaux ou en conférence. Je remercie en particulier Jesper Jacobsen pour avoir été le premier à m'introduire il y a fort longtemps à l'élégante équivalence entre différents problèmes géométriques en physique statistique et à Bernard Derrida pour avoir encadré mon stage de Master et m'avoir fait découvrir la dynamique des populations. Je tiens à remercier le laboratoire et en particulier Viviane Sebille et Sandrine Patacchini pour leur aide inestimable.

J'aimerais remercier les autres membres du groupe Mora-Walczak. Dans l'ordre : Yuval pour son humeur égale, son hédonisme subtil et son amour des choses simples qui décomplexe, Andreas pour son aide en informatique, son art dramatique et les périples américains (du Golden Gate à Flagstaff), Rhys pour les discussions sur l'information, son amour du Rubik's Cube et son humour, Quentin, le médecin du groupe pour son flegme légendaire, ses ressources musicales et sa créativité scientifique, Paulina, toujours joyeuse, pour m'avoir fait re-découvrir le basket et les mystères de la Pologne, Huy pour nous avoir fait rêver de contrées aussi variées que la Finlande et le Vietnam, Max pour son art de la danse et son introduction au Nord-Américain vernaculaire et Christophe que nous aurions souhaité voir plus si un couloir ne nous avait séparés.

Merci aux autres membres du labo, doctorants et post-doctorants avec qui j'ai pu partager des bureaux ou des cafés. Merci dans un ordre chaotique à Martin, Antoine, Alice, Alaa, Harold, l'autre Antoine, Mathieu, Sophie, Manon et tant d'autres que j'oublie sans doute.

Merci à tous les chercheurs rencontrés en conférence avec lesquels j'ai eu tant plaisir à discuter et qui sont trop nombreux pour être tous nommés (Michael Lässig, Rob de Boer, Ilya Nemenman, Anton Zilman, Vincent Hakim, Curt Callan, Alan Perelson, Massimo Vergassola, Shenshen Wang, Matteo Marsili, Gasper Tkacik, Michael Desai, Hernan Garcia, Benny Chain pour en citer quelques uns). Ainsi qu'à tous les doctorants et post-doctorants de la communauté.

Merci à mes professeurs de Saint-Mandé, Vincennes et du lycée Henri IV et en particulier à Serge Francinou qui m'a enseigné la rigueur mathématique.

Je tiens à remercier mes amis les plus proches. Les rares et chers amis du lycée que j'ai gardé à travers les années. Mes amis d'Henri IV avec lesquels j'ai découvert les sciences et qui m'ont permis de traverser ces deux difficiles années: mon premier groupe de colle, Baptiste et Clément mais aussi Thomas, Alexandre, Aurélien, Pierre, Yo, Astrid, Julien, Bruno, Joon, Mathilde et tant d'autres. Merci au groupe du B1 de l'ENS: Pierre, Hélène, Ségolène, Robert, Marine, Xavier, Nofer, Thibaut, Elsa, Ruben, Clothilde, Brice, Mathieu et leurs assimilés pour m'avoir tant appris sur les sciences, la littérature, la philosophie. Merci à Anne-Laure et Brice pour avoir été si présents pendant ces quatre ans. Merci enfin à Sarah pour m'avoir été un soutien inestimable dans les moments difficiles.

Pour terminer, je tiens à remercier ma soeur, ma mère, mon père et ma grand-mère pour tout ce qu'ils m'ont apporté et dont la liste serait bien plus volumineuse que la thèse qui suit. Cette thèse est un peu de vous même si vous répétez sans cesse que vous ne pouvez la comprendre ...

Contents

Chapter 1

Introduction

If the field of biophysics is hard to define, it is even harder to define theoretical biophysics. The use of theory of biology is not new (no one would argue that the theory of evolution, for instance, is not a cornerstone of biology), but a quantitative theory of biology is a recent idea.

Over the last decade experiments in biology and medical sciences that include quantitative measurements with the physics-level precision have exploded. Large amounts of data are available in many subfields of biology, and the need for tools and models to analyze them and make sense of them is great. Large scale genetics experiments can create ever increasing amount of data using deep sequencing, while single cell experiments give us access to the finer structure of living organism. Physics, with its long tradition of modeling complex systems, has the tools required to efficiently build simple descriptions of biological systems.

Most tools available in physics to model complex systems were developed in the field of statistical mechanics. The emergence of collective behaviour from a large number of individual subsystems and their stochastic description are common in many biological systems. This collective behaviour in the stochastic description of the group arises from the (almost) deterministic laws of physics and chemistry at the molecular level because many units or many interactions are involved. A lot of biological problems involve large numbers of cells or proteins, or long time scales and the most efficient way to describe them is the framework of statistical mechanics (and in some cases thermodynamics).

Historically, the fields of biophysics and stochastic processes are closely related. Robert Brown, the first person to observe and describe Brownian [1] motion was a botanist. The explanation of the irregular movement of particles he observed came a century later, from physicists. As pointed out in [2] a key point of Albert Einstein's [3] (and Smoluchowski's [4]) explanation of the phenomenon is that "the motion of these molecules is so complicated that its effects can only be explained probabilistically". While it is obvious to physicists nowadays that this is one of the foundations of descriptions of complex systems using statistical mechanics, it is also an early beginning of a long tradition of stochastic models in biophysics. The use of statistical mechanics is particularly strong in the two subfields of this thesis: population dynamics and gene regulatory networks. In population dynamics, while some deterministic models can capture the essence of the dynamics (such as the famous Lotka-Volterra equations [5, 6]) most models follow Wright, Fisher or Kimura [7, 8, 9] in population genetics and include the effect of stochastic fluctuations [10]. Similarly, in gene regulatory networks, most chemical events happening in cells are known to be stochastic (e. g., binding and unbinding [11], conformation

changes)[12]. Including the effect of noise is the only way to go beyond the law of mass action and understand the details of gene regulation [13]. The stochasticity of gene activation is paramount for studies of the accuracy of gene regulatory systems [11],.

Different subfields of biophysics developed as interaction and collaboration between physicists, medical doctors and biologists increased. One of the most impactful fields of theoretical biophysics is theoretical immunology. The field takes it start from the publication of pioneering work on shape space of immune receptors by Perelson and Oster in 1979 ([14]), which gave a first answer to the question: what fraction of the pathogenic environment can an immune cell react to? This work was based on the assumption that complex interactions of immune cells and antigens can be described by effective low dimensional variables that can be used in theoretical models without being explicitly described and specified. Such effective descriptions follow the tradition of models in statistical mechanics. The early success of theoretical immunology in improving the understanding of HIV and fighting it (see [15] for a thorough review of the topic) has made it a major field of biophysics. Development of sequencing and deep sequencing have provided large amounts of genomic data and stirred the explosion of new bioinformatics tools, and in particular their application to the theoretical immunology research. The work done during my PhD, while using such experimental genomic data, runs more along the lines of effective analytical models than heavy bioinformatics programming.

The vertebrate adaptive immune system - comprised of B-cells and T-cells - is the second line of defense of our organisms against pathogens such as viruses and bacteria. Pathogens are identified by the adaptive immune system through the recognition of the molecules they produce called antigens. Adaptive immune cells express antigen-specific receptors on their membrane that can recognize pathogens and trigger immune responses [16]. To face the wide variety of threats in the environment and the fast evolution of these pathogens, and to fulfill the need for a fast response to invasions, the adaptive immune system relies on a large repertoire of receptors with well-honed dynamics. Receptors are very specific (the binding affinities of receptors to antigens are high only for a very small fraction of couples) and cross-reactive (each receptor can bind to several antigens and each antigen to several receptors). The adaptive immune system maintains these specific features by using a mechanism for selection of cells able to fight off invasions and controlling the size of clones (groups of immune cells that share the same receptor). The distribution of clone sizes is a signature of selective dynamics in immune systems. Recent experimental techniques (single molecule barcoding) give us access to large and reliable data including clone size distributions. They show that, even though receptors are very different from cell type to cell type and from species to species, the distribution of clone sizes is systematically heavy tailed and resembles a power law. The main ingredients of adaptive immune systems have been determined experimentally (e. g., pathogen recognition triggering division, exchange of growth factor, hypermutations for B-cells) but the specifics of the dynamics remain quite inaccessible as available data is always a modulated result of the dynamical processes (division, death, differentiation, mutation, selection). In this thesis, I use clone size distributions as a probe into the selective dynamics of immune systems. I rely on stochastic models to bridge the gap between experimental data and theoretical understanding. In my models, each cell in a given environment can be attributed a fitness based on its ability to bind to pathogens or growth factors. Cells with high fitness are more likely to divide and less likely to die: cell lineages constitute a small Darwinian system within the organism. This Darwinian evolution is kept out of equilibrium by constant introduction of new clones and by

fluctuations of pathogenic environments. Previous analyses of Darwinian selection dynamics do not apply to the immune system because most new lineages stem from external production in the bone marrow or the thymus and not from branching processes and the genetic drift (with the notable exception of hypermutations).

In Chapter 4 (which follows [17]), I show that fitness fluctuations acting at the level of a clone and not at the level of a cell are necessary to reproduce the long-tailed distributions observed in data. More precisely, I show that models of adaptive immune systems fall into two classes:

- In clone-level noise models, the main signal to initiate cellular division is the recognition of pathogens by the receptor. Fluctuations of the environment are perceived consistently by cells across each clone, leading to large expansions of specific clones and the absence of a population scale in the system. Both numerical and analytical analyses show that such models produce power laws in clone size distributions. I express the power law exponent in terms of the biological parameters and discuss what can be learned from experimentally observed power law exponents.

- In cell-level noise models, the main signal to initiate cellular division or death is the exchange of cytokines (non specific proteins that have been shown to influence growth and division of lymphocytes). Cytokines do not bind to immune receptors and fitness can vary within a clone. In this model, fitness depends on a complex cell-state that does not require to be explicitly defined for the analysis of the model. The relevance of clone structures for the dynamics is no longer related to receptor-based antigen recognition but relies on its identification with cell lineages providing correlation in fitness between clonal cells that decays with time, division and population size. I show that these models do not produce power laws but could be mistaken for one when sequencing is not deep enough.

This work shows that physicists' tradition of classifying models can prove very powerful when dealing with large numbers of unknown biological parameters.

Another important question in theoretical immunology is understanding how the adaptive immune system matures and ages. The constant short-term selection of adaptive immune cells in their fluctuating environment also has long-term consequences for aging of the immune system. I analyze these long-term variations of immune repertoires by modeling the turnover of the fittest clones in receptor-space niches. This analysis requires a very fine description of receptor-antigen interactions. The analysis of Chapter 4 assumes steady-state at the level of the organism and relies on the assumption that the complex network of immune interactions can be described by sets of independent equations in a sort of mean-field approach. In Chapter 5, I explore the fine structure of the bipartite random graphs that represent B-cell and T-cell pools interacting with antigens. I consider the effect of selection on clone distributions and discuss the prediction of fitness drift models on aging of immune repertoires. I find that fitter and fitter clones are selected over time and that the variation of the fitness distribution of clones depends on the specifics of receptor space. To relate the analysis to the results of Chapter 4 and to experimental data I investigate the effect on clone size distributions of different features of the network of antigens and clones such as the dimension of receptor space, the adjacency of the antigen-receptor interactions and the niche structure of antigenic resources. I find that in certain parameter regimes, the description of Chapter 4 can break due to intraclonal competition.

The second part of my PhD was dedicated to a collaboration with experimentalists at the Curie institute on development of fly embryos. This follows an established line of research started

by Berg and Purcell. In [11], Berg and Purcell gave their celebrated formula on limits of precision of sensing molecular signals that spread through diffusion processes. It shows how simple arguments based on scaling laws can be used to determine natural limits on performance of biological systems. We can now revisit these questions thanks to great experimental developments that allow us to study gene expression in living systems [18, 19].

During development, cells need to determine their fate to contribute to building a functioning organism. Cell fate is defined by activation or repression of several genes resulting in the production of different proteins [20] as represented by the famous Waddington landscapes [21]. The decision to activate or repress these gene regulatory networks is based on a reading of the cell's position in the embryo through protein gradients. Specifically, the formation of the antero-posterior axis in the *Drosophila melanogaster* embryo happens in the early cell cycles by *hunchback* genes in the nuclei reading off the concentration of the Bicoid maternal gradient [22], which decays exponentially along the antero-posterior axis of the embryo. During these early cell-cycles, the embryo is made of one large cytoplasm and many nuclei. Recent progress in imaging techniques has made it possible to record gene expression in live organisms with exceptional temporal resolution [18, 19]. The goal of my work was to analyze this new live fluorescent imaging data to infer the structure and dynamics of *hunchback* gene expression.

Fluorescence is accumulated at the locus as the gene is read. Our model includes a description of the accumulation process and several competing hypotheses on the dynamic of gene switching (Poisson, Markov, high dimensional Markov, and non Markov). It predicts the behaviour of the autocorrelation function of the fluorescent signal overcoming all experimental and biological challenges: short time traces, experimental noise, and fluorescent background. We infer the parameters of gene activation at different positions along the antero-posterior axis and compare the validity of the competing assumptions on gene activation. We find that *hunchback* gene dynamics are bursty with several events of activation and deactivation within one cell cycle. We show that the precision of the readout of Bicoid gradient expected from the inferred dynamical parameters is consistent with experimental results but far below the observed accuracy of antero-posterior boundary patterning. These results imply the potential recovery of missing information further downstream in the regulatory pathway. This work is submitted and available on the arXiv [23].

The rest of this Dissertation is structured as follows.

In Chapter 2, I introduce the tools from statistical mechanics and the numerical simulation methods used in the rest of the Dissertation.

In Chapter 3, I give a brief introduction to immunology and discuss some important theoretical immunology models from the last two decades.

In Chapter 4, I derive the equation of clone sized distributions in fluctuating environments and show that only clone level noise can explain the scale free experimental data. This Chapter is a direct copy of the work published in [17].

In Chapter 5, I explore selection over long time scales in the immune system and derive clone size distributions in explicit competition models of the naive immune repertoire.

In Chapter 6 I give the necessary notions on development to understand the related parts of this dissertation, present experimental methods on gene expression imaging, and discuss limits to sensing in diffusion limited processes.

In Chapter 7, I build a model of autocorrelation functions of gene expression time traces to extract the parameters of gene activation from fluorescent live imaging. I compare the estimates

for accuracy of protein gradient readouts with experimental data and the predictions from the different models of stochastic gene switching. This chapter is a direct copy of the work submitted for publication and available on the arXiv and bioRxiv [23].

Chapter 8 contains the conclusion, discussion and the outlook of the future research.

Chapter 2

Stochastic processes and simulation methods

In this Chapter I present analytical and numerical tools required to tackle the challenges of the biological models used in the Chapters that follow.

2.1 Stochastic processes in statistical mechanics

In this section I present the tools from statistical mechanics used in the following parts of this work. See [24] for a more detailed presentation.

2.1.1 Markov processes

We define a stochastic process as a random path that is a function of time $X(t)$. The variable X itself can be multidimensional or even infinite dimensional. Note that the time t can either be continuous or discrete.

When X can take discrete values, the probability distribution of the process X at time t is written as $P(x,t)$. When the process can take continuous values the probability for X to be in the window dx is $p(x,t)dx$.

The simplest stochastic process is a process independent in time where the joint probability distribution factorizes as:

$$p(x_1,t_1;x_2,t_2...x_n,t_n) = \prod_{i=1}^{n} p(x_i,t_i). \tag{2.1}$$

A generalization of independent processes, known as Markov process, is not uncorrelated to its past, but a process where all the memory is captured in the variable itself at all times, so that the dependence of the conditional probability is reduced to the last point of the past

$$p(x,t|x_1,t_1;x_2,t_2...x_n,t_n) = p(x,t|x_n,t_n), \tag{2.2}$$

where $t_1 < t_2 < ...t_n < t$.

Markov processes are so useful because in a Markov process any joint probability can be written as a product of two-point conditional probabilities and one initial condition:

$$p(x_1,t_1;x_2,t_2...x_n,t_n) = p(x_n,t_n|x_{n-1},t_{n-1})p(x_{n-1},t_{n-1}|x_{n-2},t_{n-2})...p(x_2,t_2|x_1,t_1)p(x_1,t_1), \tag{2.3}$$

where $t_1 < t_2 ... < t_n$. So a continuous Markov process is entirely determined by an initial (or final) fixed time probability distribution and its two-point conditional probability (or transition probabilities).

Markov processes obey the probability conservation equation known as the Chapman-Kolomgorov equation (or Smoluchovski equation):

$$p(x_1, t_1|x_3, t_3) = \int dx_2 p(x_1, t_1|x_2, t_2)p(x_2, t_2|x_3, t_3), \tag{2.4}$$

where $t_1 > t_2 > t_3$.

In a time homogeneous Markov process the laws of the dynamics do not depend on time and the transition probability only depends on the time difference:

$$p(x_2, t_2|x_1, t_1) = f(x_1, x_2, t_2 - t_1), \tag{2.5}$$

where again $t_2 > t_1$.

Markov processes are ubiquitous in statistical mechanics and biophysics as most seemingly non-Markovian systems can be expanded to form Markov processes in higher dimensions [2].

2.1.2 Jump processes and the Master equation

In this subsection, I describe jump processes for continuous time. Jump processes in discrete time are much simpler to define and Master equations can be derived from the continuous time case. It so happens that a lot of processes in statistical mechanics and in biology are not continuous, or that the most efficient way to describe them is by discontinuous processes. Most specifically discontinuous processes fall into the class of jump processes.

A jump process is a random piecewise constant function that "jumps" from state to state. In a discrete state space, the process is defined by the states $\{\sigma\}_{\sigma \in S}$ and the transition rates or jump rates

$$W_{\sigma'\sigma} = \lim_{\Delta t \to 0} \left[\frac{1}{\Delta t} P(\sigma', t + \Delta t; \sigma t) \right], \tag{2.6}$$

which are assumed to be finite and defined for $\sigma \neq \sigma'$. It follows from Eq. 2.6 that

$$\partial_t P(\sigma, t) = \sum_{\sigma' \neq \sigma} \left[W_{\sigma\sigma'} P(\sigma', t) - W_{\sigma'\sigma} P(\sigma, t) \right]. \tag{2.7}$$

This equation is called the Master equation and represents a very detailed description of a stochastic process. In the most general case, the jump rates can be functions of time, although the analyses presented in this work are always set in a time homogeneous framework.

In a continuous state space $s \in S$ we define the jump rates from s to s' $W(s'|s)$ as density functions:

$$W(s'|s)\Delta s' = \lim_{\Delta t \to 0} \left[\frac{1}{\Delta t} p(s', t + \Delta t; st) \Delta s' \right] \tag{2.8}$$

and the Master equation is

$$\partial_t p(s, t) = \int ds' \left[W(s|s')p(s', t) - W(s'|s)p(s, t) \right]. \tag{2.9}$$

Note that the time spent in a state before jumping is exponentially distributed with the parameter $\sum_{\sigma'} W_{\sigma'\sigma}$ in the discrete case and $\int ds' W(s'|s)$ in the continuous case. It is also worth noting that all jump processes are Markov processes.

2.1.3 Transition matrices

In this paragraph we restrict our analysis to discrete state Markov systems known as Markov chains. The definitions and results can be extended to continuous states by replacing transition matrices with kernels.

Let X_n be a Markov chain with discrete time steps $n \geq 1$. We can encode the jump rates in a matrix $\mathbf{T}$ called a transition matrix:

$$T_{\sigma'\sigma} = W_{\sigma'\sigma} \tag{2.10}$$

if $\sigma \neq \sigma'$ and

$$T_{\sigma\sigma} = 1 - \sum_{\sigma'} W_{\sigma'\sigma}. \tag{2.11}$$

In its most general form the matrix $\mathbf{T}$ is a function of time. $T_{\sigma'\sigma}$ represents the probability to be in state σ' at time $n+1$ given that the system was in state σ at time n. Let $P(n)$ be the vector of probabilities with coordinates $P_\sigma(n)$. Then the Master equation can be rewritten in a simpler form as

$$P(n+1) = \mathbf{T}P(n). \tag{2.12}$$

If the system is time homogeneous, then $\mathbf{T}$ is independent of time, and we can write that

$$P(n+m) = \mathbf{T}^m P(n). \tag{2.13}$$

$\mathbf{T}$ is a left stochastic matrix (each column sums to 1). It means that 0 is an eigenvalue. The corresponding eigenvectors (once normalized) are called the stationary distributions of the system. The stationary distribution is unique under certain conditions of irreducibility of the Markov chain that are not discussed in this Dissertation.

For a continuous time system we have the equivalent of Eq. 2.12:

$$\partial_t P(t) = [\mathbf{T} - \mathbb{1}]\, P(t). \tag{2.14}$$

Defining $\mathbf{U} = \mathbf{T} - \mathbb{1}$ we get the equivalent of Eq. 2.13 for a time homogeneous system

$$\partial_t P(t) = e^{(\mathbf{T}-\mathbb{1})(t-s)} P(s), \tag{2.15}$$

where $s < t$.

2.1.4 Langevin equations

In this section, we present the Langevin formalism in the one-dimensional case for simplicity. See [24] for complete proofs and the multidimensional case.

A Langevin equation is a random process defined by a dynamical equation for the trajectories of the form:

$$\partial_t x(t) = a(x) + b(x)\xi(t), \tag{2.16}$$

where $\xi(t)$ is a Gaussian white noise with correlation function

$$\langle \xi(t)\xi(s) \rangle = \delta(t-s). \tag{2.17}$$

There are many definitions of the Gaussian white noise as a limit of time-correlated noise. See [24] for a more complete discussion of these definitions and their implications on the definition of stochastic integrals.

The formalism of Langevin equations is convenient because it is close to the type of equations that physicists are used to writing for deterministic dynamics with the addition of the random Langevin force. It is often easier to derive Langevin equation from the microscopic description of trajectories in large systems and then extract information about the whole population by moving to the Fokker-Planck framework or manipulating directly the Langevin equation. This is in particular the approach followed in Chapter 4 and Appendix A.

Eq. 2.16 is ambiguous. When building Riemann integrals the choice of discretization of space for summing (of which the integral is the continuous limit) does not matter as the difference between conventions vanishes for small tilings. In stochastic integrals, however, Gaussian white noise is δ correlated, and so the point where the functions are evaluated when building the integral matters. The convention depends on a parameter traditionally named α to define the integral solution of Eq. 2.16

$$x(t+\Delta t) - x(t) = a(x_\alpha)\Delta t + b(x_\alpha)\int_t^{t+\Delta t} \xi(s)ds, \tag{2.18}$$

where

$$x_\alpha(t) = (1-\alpha)x(t) + \alpha x(t+\Delta t). \tag{2.19}$$

Choices of α are related to the fine details of the model and will be discussed more thoroughly in section 2.1.8. It is enough to say here that changing α is equivalent to systematically adding an extra term in Eq. 2.16.

Langevin equations are equivalent to the mathematical formalism of stochastic differential equations, where Gaussian white noise is replaced with the Wiener process. The choice of systematically using the Itô formalism ($\alpha = 0$) in mathematics is inconvenient to physicist making Langevin equations a more natural framework.

2.1.5 The Fokker-Planck equation: expanding the Master equation

The Fokker-Planck equation is an approximate description of a given Markov processes. It can be seen as a limit of Master equations in jump processes where jumps are so numerous and small that the system becomes continuous.

Let X be a jump process on a continuous state space described by Eq. 2.9. We rewrite the jump rate $W(s'|s)$ as a function of the the initial point and the jump length $W(s, s'-s)$. We assume that jumps are very small (so $W(s,r)$ is very peaked around 0 in r). We then Taylor expand W in 2.9 to get the Kramers-Moyal expansion:

$$\partial_t p(x,t) = \sum_{n=1}^{+\infty} \frac{(-1)^n}{n!}\left(\frac{\partial}{\partial_x}\right)^n (a_n(x)p), \tag{2.20}$$

where the moments of the jump rates $a_n(x)$ are

$$a_n(x) = \int dr W(x,r)r^n = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int dy (y-x)^n p(y, t+\Delta t|x,t). \tag{2.21}$$

The second right hand side in Eq. 2.21 shows how the Fokker-Planck equation emerges as a limit of jump processes. In a lot of processes used in statistical mechanics and biophysics (and, in particular, processes deriving from the Gaussian white noise or the Wiener processes), the moments vanish for $n \geq 3$ (even when they do not, the Fokker-Planck equation is often an

efficient simplifying assumption that still describes the dynamics very well). This is equivalent to saying that the variation of the process within a time step Δt are bounded by a term of order Δt^2. Under this assumption Eq. 2.20 reduces to the Fokker-Planck equation

$$\partial_t p(x,t) = -\partial_x \left[a_1(x)p(x,t)\right] + \frac{1}{2}\partial_x^2 \left[a_2(x)p(x,t)\right], \tag{2.22}$$

also known as the Kolmogorov forward equation. Eq. 2.22 is particularly convenient as it has reduced a very high dimensional problem (the Master equation) into a low dimensional continuous equation with a very specific form. At equilibrium, the solution to the Fokker-Planck equation is simply given by

$$p_{eq}(x) = \frac{K}{a_2(x)} e^{2\int_{x_0}^{x} dy \frac{a_1(y)}{a_2(y)}}, \tag{2.23}$$

where K is determined by normalizing the distribution.

However, a lot of physical systems and most biological systems are not at equilibrium, and the Fokker-Planck formalism is very convenient in these situations. It is very easy to add a source term to Eq. 2.22 when the system is kept out of equilibrium by the introduction of new particles.

Indeed, let us assume, for instance, that $p(x,t)$ represents the probability for a population to have x individuals at time t in an environment with insufficient resources (we ignore competition). The drift term $a_1(x)$ represents the decay of the population in these harsh conditions and can be written as $-\nu x$ (where ν is a death rate) and the diffusion term $a_2(x)$ is the result of birth death fluctuations in the population and so is proportional to $\sqrt{x}$. We can see that the equilibrium solution of this system is a delta function in 0 because the population is bound to go extinct at long times (which is consistent with Eq. 2.23 as $\sqrt{x}$ is not integrable around 0). If we now keep the system out of its equilibrium by constantly adding new species at random or deterministic times with rate s with a distribution of introduction sizes $\theta(x)$, the system can still be described by a Fokker-Planck equation. Then Eq. 2.22 is modified by adding a source term

$$\partial_t \rho(x,t) = \partial_x \left[a_1(x)p(x,t)\right] + \frac{1}{2}\partial_x^2 \left[a_2(x)p(x,t)\right] + s\theta(x), \tag{2.24}$$

and θ can even depend on time. Some details of the source of new species are not included in the equation, such as the arrival time distribution, which could be rigorously accounted for in the Master equation. At the level of a large population such details do not necessarily matter. If θ is independent of time, there exists a steady state solution to Eq. 2.24. Its properties will be discussed in 2.1.9.

There exist different variations of the Fokker-Planck equation (backward and forward) that can tackle a very wide range of problems (e. g., moments, escapes, first passage times). I will not discuss these here.

2.1.6 From Langevin to Fokker-Planck

The Langevin and the Fokker-Planck formalisms are equivalent, and it is very useful to have a set of rules to go from one description to the other as most problems are easier to solve (or to define) in one of the two descriptions.

From Eq. 2.16 we can extract the moments of the Kramers-Moyal expansion

$$a_n(x) = \lim_{\Delta t \to 0} \frac{\Delta x^n}{\Delta t} \tag{2.25}$$

by Taylor expanding Eq. 2.16 in x and t and computing the average of the powers of ξ. We find that

$$\begin{aligned} a_1(x) &= a(x) + \alpha b(x)b'(x), \\ a_2(x) &= b^2(x), \end{aligned} \tag{2.26}$$

where α defines the type of stochastic integral used (Itô and Stratonovitch for instance). We immediately get the Fokker-Planck equation associated with Eq. 2.16:

$$\partial_t p(x,t) = \partial_x \left[(a(x) + \alpha b(x)b'(x))p(x,t) \right] + \frac{1}{2}\partial_x^2 \left[b^2(x)p(x,t) \right]. \tag{2.27}$$

The role of α in this equation will be discussed in 2.1.8 but we can already see that changing α will simply add an extra drift term: $\alpha \neq 0$ means that the noise can influence itself and create systematic drift.

2.1.7 Boundary conditions

In many problems of biophysics the range that the stochastic process can reach is not infinite, but bounded. Species populations or numbers of proteins cannot be negative (as particle physicists haven't yet produced anti-proteins), and many populations are bounded by a carrying capacity defined by available space or homeostatic constraints. In all these cases, it is crucial to determine what happens to the process when it reaches the boundary of the available domain. These "details" can make a lot of seemingly simple problems intractable. For continuous systems, these conditions are easier to implement in the Fokker-Planck formalism. The most common types of boundary conditions include:

- Absorbing boundary conditions: when reaching the wall, the process is stopped and stays at the wall. This is very common in population dynamics as a species reaching a population size of 0 usually cannot expand back to positive values without exterior help (since Pasteur and the end of spontaneous generation). Formally it is equivalent to imposing that $p(x,t) = 0$ at the boundary. The use of absorbing boundary conditions can also be used to compute the statistics of species extinctions. The mass loss in the probability distribution represents extinction rates as species hit the wall. These problems can be tackled very elegantly with the formalism of path integrals.

- Reflecting boundary conditions: when reaching the wall, the process is sent back to the bulk, usually in the opposite direction with the same speed. This type of boundary conditions arise in mechanics when elastic collisions happen at the walls. In population dynamics carrying capacities can be represented by reflecting absorbing conditions and the population going through the carrying capacity wall should be assumed to be 0. Technically this means that the flux of probability at the wall is set to 0.

- Periodic boundary conditions: when hitting one wall, the process reenters the domain through another wall. There is little biological motivation for using periodic boundary conditions as biological systems rarely behave that way (or any real system in general). This type of conditions has its use though in representing very large systems in numerical simulations and equation solving. This is because periodic boundary conditions usually create less problems than having the distribution vanish at the edge of a tiled space. Periodic boundary conditions also preserve some translational symmetries.

2.1.8 The old problem of Itô and Stratonovitch

The introduction and the choice of the discretization parameter α in section 2.1.4 can seem a bit mysterious. We have seen that the choice of α influences the form and the solutions of the Fokker-Planck equation corresponding to the Langevin dynamics and so it is clear that it has an impact on the trajectories and cannot be ignored. The two most common choices for α are Itô ($\alpha = 0$) and Stratonovitch ($\alpha = 1/2$) [25]. The chain rule only applies to Langevin equations in the Stratonovitch formalism, while most mathematical work is done with the Itô one. Some people advocate the use of $\alpha = 1$ as a fully anticipating noise.

The choice of α is part of the model and should be based on a careful analysis of the physical or biological phenomena at hand. Attempts to determine the "right" α date back to 1974 [26] and after more than fourty years there is still no absolute answer although the consensus is that in most physical situations the Stratonovitch rule applies. What emerges is that intrinsic and extrinsic noise as defined below should not be treated the same way. I give below a summary of the main points of [25] where a detailed analysis of intrinsic and extrinsic noise is given.

α represents how much the noise anticipates the future (an important point for financial markets in particular): a non-zero value of α means that, at the microscopic level, the noise can act on itself and be self-correlated even at very small time scales. In that sense, the discretization of an extrinsic noise (such as the random external force driving a mechanical system or the introduction of new pathogens in the immune system) should lead to $\alpha = 1/2$ as the noise, being produced outside of the system, does not have to "wait" for the system to evolve. There is no reason for the effect of the noise on itself not to be centered. The continuous time equation does not need to include the limit of a discrete delay of the effect of the noise on itself.

In intrinsic noise (such as birth death noise in population dynamics), the effect of the noise on the dynamics should be delayed as it is produced by the noisy system itself. In this case, the noise should not be anticipating the future and the Itô convention should work.

Thirty years later, after a lot of discussion, a pretty similar picture still holds [27], with the important addition that noise in continuous limits of discrete discontinuous processes should also obey the Itô rule (as the discontinuity and discreteness mean that events are scarce and cannot affect the noise itself without delay).

The effect of these choices will be illustrated and discussed at length in the context of population dynamics in immunology in Chapter 4 and Appendix A.

2.1.9 In and out of equilibrium

Thermodynamics was initially formulated at equilibrium, where all variables are well defined. Since the 1970's, methods to deal with the difficulty of out-of-equilibrium systems have flourished, and we are now armed with a variety of tools to tackle these challenges. Most biological systems are set out of equilibrium including both the immune system in Chapter 4 and Appendix A and the gene regulatory network of Chapter 7.

A state of equilibrium is reached when stochastic processes are at the detailed balance. In the language of 2.1.2, at any steady state we have

$$\partial_t P(\sigma, t) = 0 \Rightarrow P(\sigma, t) \sum_{\sigma'} W_{\sigma'\sigma} = \sum_{\sigma'} P(\sigma', t) W_{\sigma\sigma'}. \tag{2.28}$$

Beyond this, what detailed balance ensures is that "each elementary process is equilibrated by

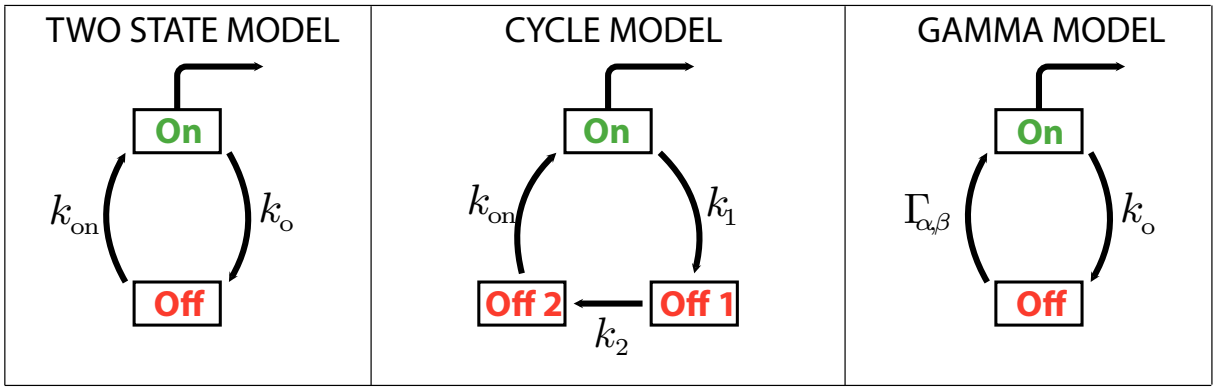

Figure 2.1: Different jump process models of promoter binding and gene activation. In these models, the gene can be in different states. In some states, labeled ON in the figure, the gene can be transcribed. Other states are labeled OFF, and in these states the gene cannot be transcribed. The arrows from state to state correspond to jump probabilities. In the left panel, the process is Markovian and jump times are exponentially distributed. The second model (central panel) is also Markovian, but irreversible, as the process cycles through the states. The third model (right panel) is non-Markovian as the jump times from OFF state to ON state are Γ distributed (with parameters α and β) and not exponentially distributed. The Γ function for jump times is peaked and has intrinsic memory. These models are studied in Chapter 7.

its reverse process":

$$P(\sigma, t)W_{\sigma'\sigma} = P(\sigma', t)W_{\sigma\sigma'}. \tag{2.29}$$

When at equilibrium, in particular, the transition matrix can be symmetrized and its eigenvalues become real. It is then much easier to directly access the equilibrium distribution. At equilibrium the system is reversible, no entropy is produced, and no energy is dissipated. In a system out of equilibrium, the forward path and backward path do not have the same probability to happen (see [28] for details), and this irreversibility is related to entropy production and energy dissipation [29].

Systems in biology are out-of-equilibrium essentially for three potential reasons: they are transient and have not yet reached a steady state or equilibrium, there is a source disturbing the system, or the system is at steady state, but going through cycles that are not reversible.

In Chapter 4, the immune system is kept out of equilibrium by the constant introduction of new lymphocytes and new antigens, as well as by the decay of antigenic concentration in the body. The bath (here the outside world and the bone marrow or thymus) creates a flux of cells and clones through the system. One of the recent ideas in the field of immunology is to use this population or fitness flux to quantify how far out of equilibrium the system is [30].

In Chapter 7, the activation of the gene state by the Bicoid protein is represented by a Markov chain with different assumptions (or a Gamma model). When the chain has only two states with exponential jumps (Fig 2.1 A), the system is reversible as it can jump from any state to any other state directly, and the steady state solution is at equilibrium. In the Markov models with a higher number of states (Fig 2.1 B), the chain is irreversible (in the specific models of Chapter 7). The cycle structure creates a flux and dissipates energy. The theoretical and biological meaning of fluxes in gene regulatory networks is studied at length in [31].

2.1.10 About non-Markovian processes

Almost all the processes presented in this work are Markov processes. The tools that have been developed for Markov analysis are so powerful, that in most situations it is worth going through the trouble of finding a Markov description for the system of interest. In some cases where the non-Markovian nature of the process is deep (in a sense that we will try to define later), it can prove impossible to map the problem onto a Markov case.

Let us consider a simple example that illustrates mappings between Markovian and non-Markovian processes. In Chapter 4, we study the integrated Ornstein-Uhlenbeck process that can be defined as

$$\frac{dx}{dt} = f_0 + y(t), \tag{2.30}$$

$$\frac{dy}{dt} = -\lambda y(t) + \sqrt{2}\gamma\xi(t), \tag{2.31}$$

where ξ is a Gaussian white noise. This process is clearly Markov as it is defined by a traditional Langevin equation. Another way of writing the same process is

$$\frac{dx}{dt} = f_0 + \eta(t), \tag{2.32}$$

where η is not a Gaussian white noise, but a correlated noise with mean 0 and the same autocorrelation decay as the Ornstein Uhlenbeck process. The process described in Eq. 2.32 is clearly not Markovian as its memory is hidden in its derivative (the noise η). Simply adding one extra dimension to the process is enough to make it Markovian. Such processes are not deeply non-Markovian and can usually be rewritten as Markov processes in higher dimensions. Methods have been developed to solve first passage time problems directly for such processes using the Chapman Kolomogorov equation.

Some processes can also be deeply non-Markovian as their memory is inifinite dimensional, and there is no way to reduce them to a Markov case by adding a finite number of extra dimensions. This is the case of the Gamma model in Fig 2.1 C. The Gamma distribution can be seen as the convolution of several Markov steps with the same exponential jump parameter k in the limit of small k and so is the limit of higher dimensional Markov models but cannot be written exactly as a finite-dimensional Markovian process. In the general case it is impossible to reduce it to a Markov chain in higher dimension. In Chapter 7, I give a method to compute the autocorrelation function of the Γ model fluorescent traces in Fourier space exactly.

2.1.11 Correlation, covariance, and autocorrelation functions

Chapter 7 is mostly concerned with the autocorrelation of stochastic processes. A few theoretical definitions can help before trying to understand the empirical problem.

The covariance of two random variables X and Y is defined as

$$\mathrm{cov(X,Y)} = \langle(\mathrm{X} - \langle\mathrm{X}\rangle)(\mathrm{Y} - \langle\mathrm{Y}\rangle)\rangle = \langle\mathrm{XY}\rangle - \langle\mathrm{X}\rangle\langle\mathrm{Y}\rangle \tag{2.33}$$

and measures how much one process determines the other.

The average $\langle . \rangle$ here is taken over realizations of the process. In ergodic systems this is equivalent to taking the time average of the process. The stochastic processes in the following chapters are all ergodic. The equivalence empirically breaks when the time available for averaging

the process is too short. In this case, the average over the finite time of the process available $\langle . \rangle_t$ is very different from the average over realizations $\langle . \rangle_\omega$ (also defined as the "true" or theoretical average $\bar{X}$). These notions will be expanded in Chapter 7.

The (Pearson's) correlation between two random variables is very similar and defined as

$$\rho_{X,Y} = \frac{\langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle}{\sigma_X \sigma_Y}, \tag{2.34}$$

where σ_Z is the standard deviation of the probability distribution of the variable Z.

For steady state stochastic processes, we can define the autocorrelation function of the process $X(t)$ as a function of the time distance $\tau = t - s$ between the two time points involved

$$\rho_X(\tau) = \frac{\langle (X_t - \langle X \rangle)(X_{t+\tau} - \langle X \rangle) \rangle}{\sigma_X^2} = \frac{\langle X_t X_{t+\tau} \rangle - \langle X \rangle^2}{\sigma_X^2}, \tag{2.35}$$

where averages and variances are independent of time because the system is in the steady state. Chapter 7 deals with autocorrelation (and cross-correlation) when the average $\langle X \rangle$ cannot be estimated properly. Then the second equality in Eq. 2.35 no longer holds, and new definitions are required.

2.2 Computational methods

This section describes the numerical tools used for this work. All simulations and computational methods have been implemented in Matlab. The methods are standard and have been implemented independently of specialized numerical libraries with the exception of some optimization routines.

2.2.1 Simulating random population models: Gillespie method and fixed time step methods

The models of Chapter 4 and Chapter 5 both contain equations for the dynamics of a population of cells that divide and die at rates that depend on the other actors of the system, sometimes the other cells and sometimes the antigens (that also evolve randomly). Concurrently, new clones and new antigens enter the system at random times.

When the system has to be described cell by cell by modeling explicitly cell division and clone size as an integer (and not in a continuous approximation of the populations of the clones), the model is formally equivalent to a reaction process. In a reaction process, different components react with each other at rates that can vary with time and be functions of the concentrations of components, just like in the population dynamics problem.

The most efficient way of simulating this type of problem is usually the family of Gillespie and Gillespie-like algorithms. The Gillespie algorithm ([32, 33]) popularized by Gillespie in 1976 and invented by Doob in 1945, simulates the trajectories of a reaction process. The necessary parameters are the initial quantities of the different reagents, the set of possible reactions, and formulas for computing the reaction rates between the reagents. The waiting time before each reaction happens is assumed to be exponentially distributed. The algorithm repeats the following steps:

- The algorithm computes the rate k_i of each reaction from the concentrations of reagents (or the absolute number of reagents if the system is described this way). Note that here the rates are defined as already containing the reagent concentration factor, which is usually not the convention in chemistry or in most descriptions of this algorithm.

- The algorithm computes the waiting time before the next reaction. The minimum of n exponentially distributed random variable with rates $k_1, k_2 ... k_n$ has the same distribution as an exponentially distributed random variable with a rate $\sum_j k_j$. Thus the algorithm draws an exponential random variable with a rate equal to the sum of the reaction rates. Time is updated.

- The algorithm decides which reaction happens. Once the time is picked, the algorithm randomly picks the reaction that happens. Each reaction happens with a probability $k_i / \sum_j k_j$. The reaction happens and concentrations or absolute numbers are updated accordingly.

This method is very accurate because it exactly simulates the Master equation defined above. It is computationally very expensive. Faster but approximate versions of the algorithm exist. One of the most famous ones is the τ leaping method [34] adapted to systems where reactions rates do not vary much when only one reaction happens but require many reactions to change [35]. This algorithm computes a time that statistically corresponds to the number of events required to macroscopically change the reaction rates. It then computes the number of times each reaction has happened during the interval using Poisson distributed random variables.

I will not describe here all the variants of the Gillespie algorithm. The original version was implemented for simulating clones of adaptive immune systems in Appendix A.8. Each clone is a reagent with a population and the reactions are divisions and deaths of cells. Division and death are assumed to be memoryless and can be represented by exponential processes. The Gillespie algorithm was also implemented to simulate Markov chains in Chapter 7 as it simulates exactly the Master equation of Markov processes with discrete states.

All other simulations in population dynamics in this work are based on Langevin equations and continuous descriptions of population sizes (and their fitnessess when they are defined). In these cases the Gillespie algorithm does not apply as the system is not described in terms of discrete reaction events. Simulating a stochastic differential equation or a Langevin equation requires two parts: one of them is drawing the random variables and the other one is having a method for integrating the differential equation.

All algorithms need to use a time step that is small enough to ensure that the deterministic increment is small and that the random increment has very high probability to be small. Other than that, the methods are very similar to Euler integration in ordinary differential equations. There exist different methods for Itô and Stratonovitch formalisms. All methods are simpler for Itô because the functions are evaluated at the previous point exactly. In all the simulations in this work, any equation with the Stratonovitch convention was analytically transformed into its Itô equivalent and then simulated using Itô based methods. For this reason, I only present methods for the Itô formalism here.

When simulating the trajectories $x(t)$ of a Langevin equation (with the Itô convention),

$$\partial_t x(t) = a(x) + b(x)\xi, \tag{2.36}$$

the simplest algorithm called the Euler-Maruyama method uses a given time step Δt to integrate the process

$$x(t+\Delta t) = x(t) + a(x(t))\Delta t + b(x(t))\Delta\xi, \tag{2.37}$$

where $\Delta\xi$ is a Gaussian random variable with mean 0 and variance Δt. This algorithm requires the use of a short time step because the order of convergence is 0.5, which is a very slow convergence (i. e., the expectation value of the difference between the process and its approximation by the Euler-Maruyama method is proportional to Δt). The Euler-Maruyama method is called "the order 0.5 strong Taylor scheme".

The Euler-Maruyama method can be improved by including higher order derivatives in the integration scheme. Including the first order derivative is called the Milstein method, and for the Itô convention, the integration step is

$$x(t+\Delta t) = x(t) + a(x(t))\Delta t + b(x(t))\Delta\xi + \frac{1}{2}b(x)b'(x)(\Delta\xi^2 - \Delta t). \tag{2.38}$$

Intuitively, the Gaussian white noise ξ has variations of size $\sqrt{\Delta t}$ so the expansion in Eq. 2.37 is of order $\sqrt{\Delta t}$. In Eq. 2.38 the expansion is of order Δt so it includes square terms from ξ and $b(x)$, but only linear terms from $a(x)$. This method has faster convergence for small time steps (of order 1) than the Euler-Maruyama method. It is possible to avoid explicitly computing the derivative of b in Eq. 2.38 if it is not analytically available by replacing it with a finite difference approximation without reducing the performance of the algorithm.

Higher order methods can be implemented with Runge-Kutta schemes that take into account the Δt^2 terms in the Taylor expansion. The formulas are heavier and heavier as the order increases, but the time steps can be longer, which is usually advantageous when simulating large populations. The simulations of populations of clones of the immune system in this work were done with the Euler-Maruyama method with small time steps as they were never computationally very heavy.

2.2.2 Numerical solutions to partial differential equations (PDEs)

Numerically solving a stochastic differential equation can be done in two ways. The first one, described in the previous section, randomly generates enough trajectories to estimates the probability distribution. The second type of methods solves a partial differential equation to obtain directly the probability distribution of the process, usually solving the corresponding Fokker-Planck equation.

This section explains how to find approximate numerical solutions to a Fokker-Planck equation

$$\partial_t p(x,t) = -\partial_x \left[a_1(x)p(x,t)\right] + \frac{1}{2}\partial_x^2 \left[a_2(x)p(x,t)\right], \tag{2.39}$$

on a domain Ω with specified boundary conditions on $\partial\Omega$ using finite difference methods. Finite difference methods build a discretized version of Eq. 2.39 to solve it on a grid of Ω. Finite difference method is applied to two-dimensional equations in Chapter 4, where one dimension is clonal population and the other one is fitness. I describe the method for a one-dimensional problem in this section for simplicity. In all the equations of Chapter 4, there is a source term and the goal is to find a steady-state solution, so the left-hand side is set to zero. In this context, Eq. 2.39 becomes

$$0 = -\partial_x \left[a_1(x)p(x)\right] + \frac{1}{2}\partial_x^2 \left[a_2(x)p(x)\right] + s(x), \tag{2.40}$$

where $s(x)$ is the source term. The equations have an absorbing boundary at 0 for x, but the probability distribution is defined for populations going to $+\infty$. Of course it is impossible to define a grid for an infinite space numerically. Thus I use the fact that the probability distribution vanishes for very large values of x to restrict Ω to a bounded domain $0 \leq x \leq x_{\max}$ (the latter is picked to be wide enough for the probability distribution to be very small at the boundaries representing infinity). The boundary conditions at the boundaries representing infinities are set to be reflecting to avoid losing probability mass artificially.

To find steady-state solutions of Eq. 2.40, we reintroduce time, compute a discretized version of the differential operator acting on $p(x,t)$ on the right-hand side of Eq. 2.40, and propagate it until the distribution reaches a fixed point of the operator. This fixed point is the steady-state distribution.

For each (discrete) time point n, there is a vector of probability u^n for the J discrete x values $x_j = jh$, where $h = x_{\max}/J$ (i. e., $u_j^n = P(x(n) = x_j)$). The first step is to build the discrete version of $\partial_x [a_1(x,t)p(x,t)]$:

$$\partial_x [a_1(x,t)p(x,t)] \Rightarrow \frac{a(u_{j+1}^n)u_{j+1}^n - a(u_j^n)u_j^n}{h}. \tag{2.41}$$

It is sometimes safer to write the discretization for derivatives of only one function by expanding $\partial_x [a_1(x,t)p(x,t)] = p(x,t)\partial_x a_1(x,t) + a_1(x,t)\partial_x p(x,t)$. The same is done for time

$$\partial_t p(x,t) \Rightarrow \frac{u_j^{n+1} - u_j^n}{\Delta t}, \tag{2.42}$$

where Δt is the time discretization constant. This choice of discretization is called the forward discretization.

The second step requires discretizing a second order derivative. There are several ways to do so, the simplest one being the explicit central difference:

$$\partial_x^2 [a_2(x,t)p(x,t)] \Rightarrow \frac{a(u_{j+1}^n)u_{j+1}^n - 2a(u_j^n)u_j^n + a(u_{j-1}^n)u_{j-1}^n}{h^2}. \tag{2.43}$$

The explicit central difference is stable (i. e., it does not amplify small errors in the probability vector), but is not necessarily the most efficient way of discretizing the equation. Other choices of discretization include in particular the implicit scheme, where the spatial derivative is written for the next time step. It is usually more accurate, but requires more calculations. The equations of Chapter 4 are numerically solved using the simple explicit method.

The operator is built by adding the contributions of Eq. 2.41, 2.42 and 2.43 into one matrix that acts on the probability vector. The solution is obtained by having this matrix act on the probability vector until a fixed point is reached.

2.2.3 Optimization and Maximum Likelihood estimation

In Chapter 7 the goal is to infer the parameters of a model by comparing the prediction of the model with experimental or simulated data. The optimal parameters are found by maximizing a function called Likelihood which is proportional to the probability of having a given set of parameter given the data.

More precisely the data is a set of observations $x_1, x_2, ..., x_n$. The model depends on a set of parameters θ that are the variables that need to be determined (or optimized). For each set of

parameters θ, the model provides us with the probability to see an observation $p(x_i|\theta)$. If the different observations are independent, then the probability to see a given series of observations is

$$\mathcal{L}(\theta, x_1, x_2...x_n) = p(x_1, x_2...x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta). \tag{2.44}$$

This function is called the likelihood of the parameters θ for the observations $x_1, x_2, ..., x_n$. Note that it reverses the roles of parameters and observations (justified by the Bayes formula [36]). Maximum Likelihood estimation (MLE) developed by R. A. Fisher almost a century ago consists of looking for the value of θ that maximizes $\mathcal{L}$. This value is called the MLE. Because $\mathcal{L}$ is a probability of a large number of events, it varies over a wide range, and it is easier numerically to look for its maximum in log space (and the problems are equivalent because the logarithm is a strictly increasing function of its argument). If the MLE θ_{MLE} exists, it is a local maximum of the log-likelihood function and must satisfy

$$\frac{\partial}{\partial_\theta} \log \mathcal{L}(\theta, x_1, x_2...x_n) = 0 \tag{2.45}$$

and

$$\frac{\partial^2}{\partial_\theta^2} \log \mathcal{L}(\theta, x_1, x_2...x_n) < 0. \tag{2.46}$$

Numerical methods look for maxima of the log-likelihood using many different techniques. The simplest method called gradient ascent follows the most positive direction of the gradient at each point to find the maximum (in cases where it applies, conjugate gradient methods are much faster). The three major computational difficulties that can arise in trying to find this maximum are high dimensionality of parameter space, log-likelihood functions that are computationally expensive to compute, and the existence of multiple local minima. All optimizations in Chapter 7 were done for a small number of parameters (maximum of 2) and with an analytical log-likelihood function that is fast to compute. The only potential problem is multiple local maxima and it is solved by starting several optimization routines from different points of parameter space to find the global minimum. The number of optimization that needs to be performed to find the global minimum can be determined by reducing the unit of the grid until no new minima are found.

2.2.4 The example of least squares

The optimization problems of Chapter 7 are solved in terms of the method of least squares to fit the data. Each observation is a vector $\mathbf{x_i}$ with components x_i^j, $1 \leq j \leq m$. Each coordinate of the observations is assumed to be normally distributed with a variance σ_j^2 that is known and a mean θ_j that is the parameter to be fitted. The variance σ_j^2 is given by the model (or, in the case of Chapter 7, by the data since it is very hard to compute the expected variance from the model). Note that in the method of least squares, the variance can also be a parameter to be optimized.

If the observations are normally distributed, the probability to see $\mathbf{x_i}$ is

$$p(\mathbf{x_i}|\theta) = \frac{1}{\sigma\sqrt{2\pi}} \prod_{j=1}^{m} e^{-(x_j-\theta_j)^2/2\sigma_j^2}. \tag{2.47}$$

From this we obtain the log-likelihood:

$$\log \mathcal{L}(\theta, \mathbf{x_1}) = -\sum_{j=1}^{m} \frac{(x_j - \theta_j)^2}{\sigma_j^2}. \tag{2.48}$$

Defining the least square error as the sum of squared residuals $x_j - \theta_j$

$$S(\mathbf{x_i}) = \sum_{j=1}^{m} \frac{(x_j - \theta_j)^2}{\sigma_j^2}, \tag{2.49}$$

the problem of Maximum Likelihood is equivalent to finding the minimum of S. Equivalent methods to finding the maximum of $\mathcal{L}$ can be applied to finding the minimum of S.

Chapter 3

Introduction to immunology

3.1 The immune system

This section provides the reader with some introductory biological knowledge about the adaptive immune system that is specifically aimed at understanding the use of models in immunology. For a complete introduction to immunology for physicists see [37] and [16] for a thorough guide to the field.

3.1.1 The role of the immune system

The immune system is as essential to the definition of the self as our skin: it is the one of the few systems that can distinguish what is us from what is not. It has the enormous task of clearing the body of foreign invasions, but also of fighting cancer cells [16].

A pathogen left unchallenged in an organism can very quickly become a threat and so immune systems must be very efficient and fast when dealing with invasions. At the same time, the variety of pathogens existing at one time point is extremely wide and the immune system must be prepared to face them all. Moreover, bacteria and viruses evolve quickly and keep eluding existing immune defenses. For that reason, the immune system needs diversity and plasticity.

The efficiency and diversity of the immune system is the result of two selection processes: one at the evolutionary time scale with selection of individuals through generations and one at the cellular time scale with selection of the most efficient cells within one organism. In this work, I focus on the second one.

3.1.2 Actors of the immune system

All blood cells - including red and white blood cells - derive from the same type of precursor: the pluripotent hematopoietic stem cell located in the bone marrow. This precursor cell can differentiate into two types of progenitors: the common lymphoid progenitor later produces B-cells and T-cells, while the myeloid progenitor later turns into red blood cells, monocytes, dendritic cells and granulocytes.

The immune system is divided into two parts: the innate immune system is a non specific first line of defense against pathogenic invasions. The cells of the innate immune system can recognize targets as not being part of the self, but have no affinity to a specific subset of them.

On the other hand, adaptive immune cells (T-cells and B-cells) are highly specific fighters that recognize a small subset of targets that can bind to receptors located on their membrane called T-cell receptors (TCR) and B-cell receptors (BCR). Cells that share the same TCR or BCR form a clone. These targets of immune reactions are called antigen. The definition of antigen is very loose, and pretty much any kind of protein can qualify as an antigen. The adaptive immune system triggers powerful responses to infections by greatly increasing the number of cells specific to the recognized pathogen. The information about the invasion transits the innate immune system by what is called antigen presenting cells (APC), mostly dendritic cells and macrophages that gather epitopes (the part of the antigen recognized by the immune cell) of encountered pathogens everywhere in the organism and transfer them to the lymph nodes and germinal centers. Many types of cell can work as an antigen presenting cell through their Major Histocompatibility Complex (class I MHC), but only so called "professional" antigen presenting cells (macrophages, dendritic cells and B-cells) include the class II MHC and the co-stimulatory signals that increase T-cell recognition.

Cells communicate with each other through the use of cytokines, small proteins that can act as growth inducer or repressor on immune cells. Binding of cytokines does not depend on the specific receptor shared by the cells of the clone (the TCR or the BCR), but on receptors that are specific for each cytokine and the same on all cells (although their number can vary from cell to cell). The implication of this statement on the role cytokines can have on clonal dynamics is part of Chapter 4. These cytokines can be produced by immune cells (among others). There is a wide variety of cytokines, and our understanding of their role changes and improves very fast. One important cytokine that is mentioned at length in Chapter 4 is IL-7. IL-7 is a cytokine that influences production of new lymphocytes by the bone marrow and is an essential growth factor for T-cells. It regulates T-cell homeostasis and has been used in treating HIV.

3.1.3 B cells and T cells

The adaptive immune system is made of B-cells that mature in the bone marrow and T-cells that mature in the thymus. Throughout life, the organism keeps producing new B-cells and T-cells although production rates fall drastically with age. The thymus in particular shrinks in adult humans and bone marrow activity reduces to let the task of de novo production rest on flat bones. A very complete analysis of T-cell production at the different stages of life is given in [38].

B-cells and T-cells have different ways of contributing to immunity. When activated, B-cells produce antibodies, Y shaped proteins that recognize antigens and both tag target microbes or infected cells for attack by other parts of the immune system or directly kill them by blocking essential pathways for survival and division. T-cells are divided into two subtypes: cytotoxic T-cells or killer T-cells (known as $CD8^+$ T-cells) and Helper T-cells (also known as $CD4^+$ T-cells). Killer T-cells bind to infected cells or tumor cells by recognition of antigens present on the membrane of the target and then destroy cells by release of cytotoxins or triggering apoptosis. Helper T-cells play a role in immune response by enhancing proliferation of B-cells and their differentiation into a antibody secreting state. Helper T-cells are the main targets of HIV. The lethality of the disease demonstrates how important they are to the immune system.

Without going too far into details what one can get from this picture is that B- and T-cells have different action modes and rely in an asymmetric way on each other for activation

and immune response. Although their interaction is asymmetric they are both essential to the immune response.

3.1.4 The adaptive immune system and the immune response

B-cell and T-cell specificity is based on the receptor they carry on their membrane. The group of cells sharing the same receptor is called a clone. The central, variable part of the receptor (or immunoglobulin) is produced once for the clone through a complex process of random gene recombination implying the choice of gene templates for three genes (V, D and J) and a high number of insertions and deletions on a small DNA region. This random process has been studied very carefully (see for instance [39, 40]) and the probability to produce twice the same receptor is sufficiently low for cells of the same clone to be considered to belong to the same lineage. After this recombination the receptor is maintained throughout cell divisions (with the exception of hypermutations). To make sure that they do not target the self, B-cells and T-cells undergo a round of negative selection where cells that respond too strongly to self antigen stimuli are discarded. At the same time cells that do not reply at all are also discarded as not efficient enough, in a process called positive selection. After these two rounds of selection, cells are released into the periphery (blood and lymphoid organs) where they can divide or die depending on the type of signal they receive.

When a pathogen enters the organism the first defenses it will usually encounter belong to the innate immune system as macrophages for instance patrol the tissues. Bits of the invading pathogens are carried by antigen presenting cells and will eventually reach T-cells (if the infection is too severe to be fought off by the innate system alone). The message is amplified and lymphocytes specific to the invader proliferate in lymph nodes forming germinal centers. Proliferation during a full fledge immune response can increase enormously and clone sizes be multiplied several powers of ten fold.

Once activated, B-cells trigger a mechanism called somatic hypermutations where an enzyme called AID changes specific nucleotides and error-prone polymerases are recruited to repair the modifications. These mutations are targeted specifically at the variable part of the BCR and quickly produce different cell lineages with slightly different receptors. Each receptor has a different binding affinity to the antigens of the invasive pathogen and the fittest clones will be selected through a process known as affinity maturation to form effector and memory cells.

Effector B-cells and T-cells actively fight off pathogenic invasions while memory cells are stored ensuring that the organism can react very fast to another invasion by a pathogen similar to one that has already been seen. This acquired immunity has varying time scales that are not very well understood. In that sense the organism can react both immediately and on the long term to the fluctuations of its environment by adapting its immune repertoire (explaining the name adaptive immune system).

The first person to use acquired immunity was Edward Jenner in 1796. He noticed that people that had been infected with cowpox were immune to smallpox, a disease that caused several hundreads of thousand of deaths each year in Europe. The similarity between the two strains gave human immune systems an edge when fighting infections. He advocated the inoculation of vaccinia (the other name of cowpox) and named the process after it. Two centuries later smallpox had officially disappeared. The use of vaccines has spread and improved so that they can be designed for various diseases. The discovery of Edward Jenner marks the beginning of

the field of immunology.

3.1.5 The naive and memory pool

Adaptive immune cells undergo heavy phenotypic changes when activated (i.e when encountering an antigen with high enough binding affinity to stay bound for a time exceeding a certain threshold). Cells that have been activated are called memory and effector cells. Effector cells have fast dynamics with a lot of division and death. The dynamics of memory cells is hard to access, some of them are believed to be dividing while others seem to be dormant.

The naive pool contains cells that have not yet encountered an antigen. It does not include an overwhelming fraction of immune cells but it does contain the main reservoir of diversity of the immune system. Naive cells dynamics are slower than that of activated cells. Division is very rare and cells are long lived. Different estimates have been given for the different pools (in particular with labeling experiments [41, 42]) and in all of them naive cells are more dormant than other repertoires with cell life span estimates going up to 2 years in mice and a turnover about 30 times slower than for effector cells. Still, the stability of the naive pool size even after thymectomy (or using lymphopenic mice for B cells) shows that there is a control of the population of the naive repertoire (called homeostatic control).

Using mice living in sterile environment and irradiated mice where clonal diversity has been reduced Freitas et al ([43]) showed that there is competition between naive B-cells and the resources required for this competition include antigens. It is still believed that in normal conditions (in healthy individuals) the main limiting resource for naive cells are cytokines.

3.1.6 Immune systems across species

Immune systems vary across species. Bacteria are protected using a system called CRISPR that inserts bits of phage genome (spacers) in their DNA to interfere with phage invasion of the cell. Non-jawed vertebrates rely on their own specific type of adaptive immune system but jawed vertebrate all rely on adaptive immune systems that are quite similar in structure. This attribution of immune systems can be seen as the optimal solution to adapting to fluctuating environments [44]. However the specifics of immune system differ across species even among mammals.

First of all the size of the immune system scales with the size (i.e the number of cells scales with the mass) of the organism while the diversity of repertoires seems to scale as the logarithm of the mass [45]. It results that clone sizes and dynamics vary from species to species. The mouse is the most common system studied (apart from humans) in immunology and their immune system seems to be less diverse than the human immune system. Whether this is a simple consequence of scaling laws or has deeper biological meaning has not been elucidated.

3.2 Experimental clone size distributions

High-throughput data of immune repertoires has been available since 2009 ([46]) with the sequencing of antibodies in zebrafish. High-throughput methods have multiplied the number of receptors that can be analyzed in an experiment making it possible to gather large statistics on the distribution of clones (among other things). Since then the method has been expanded to

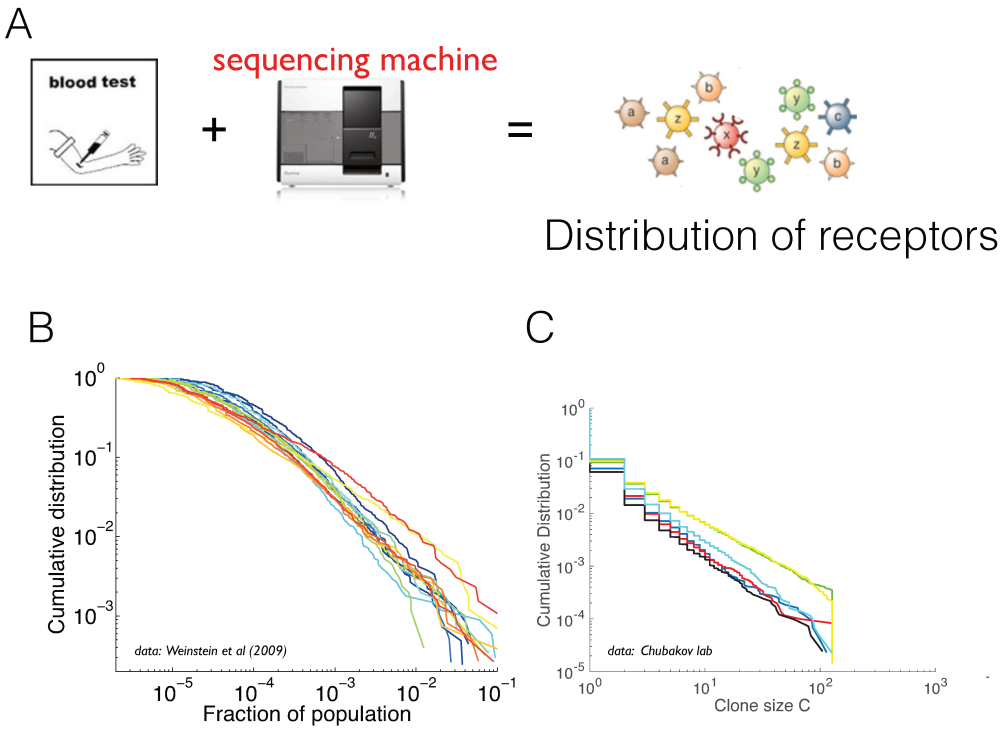

Figure 3.1: A. Extracting receptor distribution from repertoires. B. Clone size distribution in antibody repertoires in zebrafish from [46] Each colour is a different fish. C. Clone size distribution from human α chain in TCR. Experimental data from Chudakov lab, Moscow (from [47])

other cell types and other species ([48, 49, 50, 51]). The observed clone size distributions are heavy tailed and resemble power laws (see Fig. 3.1 B and C for some examples).

In humans it is now possible to obtain clone size statistics with tens of thousands of reads from a single blood sample (Fig. 3.1 A). Deep sequencing experiments also have their downfalls as the method requires multiple replications of the genetic material (DNA or RNA) trough a process called polymerase chain reaction (PCR). Replication errors in PCR can be accounted for just like in any other sequencing method but the main problem of PCR is that it may replicate different pieces of DNA or RNA at different rates making the counts of reads of each receptor much less reliable and challenging the validity of the obtained repertoire distribution. Much of the observed variability in immune repertoires could then actually derive from PCR bias in replication rates.

These doubts have been lifted in the last three years by the development of experimental methods using unique molecular identifiers or barcodes. With these methods it both possible to read the sequences of a large number of cells accurately and to count the number of times each sequence is seen. These methods have been applied successfully to immune repertoires ([46, 52]).

It is now possible to state with much confidence that the effector and memory cells in mice and human immune repertoires show distributions that are very close to power laws. Experiments on naive cells are unfortunately not that numerous and the results are still controversial although heavy tailed distributions have been reported ([50]).

Power laws in physics are well known to be associated with critical phenomena and scale-free problems. The work presented in Chapter 4 explains how scale free laws can emerge in immune dynamics but does not answer the question of criticality of immune systems. The idea that biological systems are the result of an optimization through evolution and as such tend to be poised at criticality has surfaced in the last decade [53]. It does not seem that these power laws are produced by finely tuned parameters (at least not for any exponent of the power law). The emergence of power law distributed data is not necessarily enough to justify criticality ([54]) and can simply be due to random parameters in the dynamics.

There is of course much more information in the result of a deep sequencing experiment than a simple clone size distribution. The exact sequence of all the cells can also be analyzed. Such analysis requires the use of much more complex bioinformatics tools and has lead to a quantification of selection as a function of position on the variable part of the receptor ([39, 55]) leading to a better understanding of selection before and after the release of lymphocytes in the periphery. Linking the sequence structure to function remains an open problem although progress has recently been made in this direction [56]. The choice to stay clear of explicitly modeling receptor sequence in this work is motivated by the analytical difficulty of dealing with high dimensional variables such as sequences. In the models that follow the sequence is hidden in an abstract shape space that is more closely related to function and easier to understand. Fortunately results can be derived within this convenient framework independently of the hidden genetic variables.

Recent estimates of the size of the immune repertoire in humans have shown that it is extremely large. The number of T-cells in a healthy adult is estimated to be around 4.10^{11}. Such values are sufficiently large to motivate the use of statistical mechanics tools in modeling the immune system.

3.3 Models of the immune system

3.3.1 Why model the immune system?

What can we hope to achieve by modeling immune systems and immune repertoires? Beyond the sheer interest for knowledge and understanding models of the immune systems are central for many contemporary medical questions. I have mentioned above how crucial models of the immune system have been in fighting HIV. There are many unknowns in the dynamics of human immune systems (and of mice, of fish and other animals). One of the most exciting prospects of understanding the dynamics of the immune system is a better understanding of auto-immune diseases. Auto-immune disorders are rising in western countries and constitute one of the main threats to public health as pollutions of all kind increase. It has been suggested that insufficient stimulation of the immune system leads to auto-immune disorders but no clear quantitative understanding of this phenomenon is available yet. Getting a precise idea of the dynamics of adaptive immune cells would help greatly move in that direction. All over the world vaccination is still a necessary tool to prevent epidemics. As viruses evolve and elude vaccines protecting

populations from pandemics becomes harder and harder. To improve vaccines one needs to manipulate the antigens to produce weakened strains that still trigger acquired immunity in hosts. This can only be achieved with a clear understanding of immune reactions and memory. Last but not least the success of immunotherapy has put focus on immunology in the fight against cancer.

From the point of view of evolutionary dynamics models of immune repertoires are very exciting as they constitute a variant of classic Darwinian systems: in immune systems new species are produced de novo in the bone marrow or in the thymus and do not evolve from existing ones (with the exception of hypermutations). Most of the other features of the system are similar to Darwinian dynamics with the existence of niches with specific antigenic resources.

3.3.2 A model of antigenic stimuli

In this section I present a very important model of competition between adaptive immune cells for antigenic resources. It was introduced and expanded in a series of papers by De Boer, Perelson and Freitas between 1994 and 2001 ([57, 58]). The model is theoretically and experimentally well established and constitutes a safe basis for further investigations (although not without detractors). For consistency I present the model with the notations of Chapter 4.

Let us consider a peripheral repertoire of B-cells or T-cells clones. We have a number M of clones with index $i \in [1, M]$. Each clone contains a population of cells $C_i(t)$ that is a function of time. At the same time in the organism a certain number N of antigens are present with absolute count $a_j(t)$ with $j \in [1, N]$.

The interaction between antigens and lymphocytes happens through binding. The binding affinity of antigen j with the clonotype i is a number $K_{i,j}$ and the collection of these numbers is encoded in a matrix called interaction matrix that corresponds to a weighted adjacency matrix of the bipartite graph of clonotypes and antigens. De Boer and Perelson show that the availability F_j of antigen j is inversely proportional to the sum of binding probabilities of all the cells in the system:

$$F_j(t) = \frac{1}{1 + \sum_{i=1}^{M} K_{i,j} C_i(t)}. \tag{3.1}$$

Here it is important to note that all cells of a clone have the same properties because they have the same receptor. The antigenic stimulus received by a cell of clone i is

$$S_i(t) = \sum_{j=1}^{N} a_j(t) F_j(t) K_{i,j}. \tag{3.2}$$

Note that including thymic or bone marrow output of new clones and renewal of antigens in the body would also make N, M and K functions of time. The choice of the entries of the matrix K is complex. A very reliable assumption to make is that K is sparse because adaptive immune cells are very specific. Beyond sparsity the distribution of the K matrix entries is biologically unclear. The structure of K and of the underlying shape space are discussed in the following chapters.

Here we must leave the precise framework of [57] to write a more general equation on clonal dynamics. The idea is to write an equation for the dynamics of the clonal population based on the assumption that antigenic stimulus enhances division or prevents death. Following [57] I

include the effect of antigen as a death rate decrease:

$$\partial_t C_i(t) = \mu C_i - \nu \frac{C_i}{S_i}. \tag{3.3}$$

The average stimulus can be seen as a local in time fitness of the clone. These equations can be considered in a fixed antigenic environment or coupled to antigen fluctuations. In all cases they include competition for antigens and resource niches with partial competitive exclusion. We discuss the notion of competitive exclusion in Chapter 5.

3.3.3 Current models of the immune system

In the last years there has been several attempts at modeling immune repertoires but none of them mentioned the relationship to clone size distributions. A lot of focus in particular has been put on proving that neutral dynamics (i.e equal fitness of clones) can explain the observed data [59] and trying to learn parameters from steady state of neutral equations. In the neutral model the variability of clone sizes relies entirely on birth death noise. In particular, no quantitative comparison of the predictions of neutral models with observed clone size distributions had been made. We show in Appendix A.1 that the assumptions of neutral models are not compatible with observed distributions in effector and memory cells.

This means that the variability of clone sizes cannot be explained by simple fluctuations of birth and death within equally fit populations. This conclusion leads to examining the other possible mechanisms for sources of fluctuations. Another main difference between the models developed in the following Chapters and pre-existing models is that they do not only try to extract some parameters from clone size distributions, they also help distinguish between different mechanisms for the dynamics of selection in the adaptive immune system.

In many aspects, our analysis widens the range of possible models to include more realistic dynamics. We include the neutral model in a much larger class of dynamical models based on cytokine exchange with cell to cell variability. When talking about power laws one must be careful as many people define power laws as power laws with exponential cutoff. This is not what we mean here. Of course all biological systems have cutoffs as they cannot be infinite but when a power law behaviour is observed over several decades one must check that a model of it that predicts a cutoff to the power law is indeed far enough to be consistent with the data.

Recent discussions and presentations at the 2016 Santa Barbara program on Quantitative Immunology (https://www.kitp.ucsb.edu/activities/immuno16) showed that neutral dynamics could be successfully applied to the naive repertoire and its aging. We discuss this in more detail in Chapter 5.

Chapter 4

Fitness shapes clone size distributions of immune repertoires

This chapter was published in PNAS (2016) vol 113, no. 2, 274-279 [17]

Jonathan Desponds[1], Thierry Mora[2] and Aleksandra M. Walczak[1]

[1] Laboratoire de physique théorique, CNRS, UPMC and École normale supérieure, 24, rue Lhomond, 75005 Paris, France
[2] Laboratoire de physique statistique, CNRS, UPMC and École normale supérieure, 24, rue Lhomond, 75005 Paris, France

4.1 Significance

Receptors on the surface of lymphocytes specifically recognize foreign pathogens. The diversity of these receptors sets the range of infections that can be detected and fought off. Recent experiments show that, despite the many differences between these receptors in different cell types and species, their distribution of diversity is a strikingly reproducible power law. By introducing effective models of repertoire dynamics that include environmental and antigenic fluctuations affecting lymphocyte growth or "fitness," we show that a temporally fluctuating fitness is responsible for the observed heavy tail distribution. These models are general and describe the dynamics of various cell types in different species. They allow for the classification of the functionally relevant repertoire dynamics from the features of the experimental distributions.

4.2 Abstract

The adaptive immune system relies on the diversity of receptors expressed on the surface of B and T-cells to protect the organism from a vast amount of pathogenic threats. The proliferation and degradation dynamics of different cell types (B cells, T cells, naive, memory) is governed by a variety of antigenic and environmental signals, yet the observed clone sizes follow a universal power law distribution. Guided by this reproducibility we propose effective models of somatic

evolution where cell fate depends on an effective fitness. This fitness is determined by growth factors acting either on clones of cells with the same receptor responding to specific antigens, or directly on single cells with no regard for clones. We identify fluctuations in the fitness acting specifically on clones as the essential ingredient leading to the observed distributions. Combining our models with experiments we characterize the scale of fluctuations in antigenic environments and we provide tools to identify the relevant growth signals in different tissues and organisms. Our results generalize to any evolving population in a fluctuating environment.

4.3 Introduction

Antigen-specific receptors expressed on the membrane of B and T cells (BCRs and TCRs) recognize pathogens and initiate an adaptive immune response [16]. An efficient response relies on the large diversity of receptors that is maintained from a source of newly generated cells, each expressing a unique receptor. These progenitor cells later divide or die, and their offspring make up clones of cells that share a common receptor. The sizes of clones vary, as they depend on the particular history of cell divisions and deaths in the clone. The clone size distribution thus bears signatures of the challenges faced by the adaptive system. Understanding the form of the clone size distribution in healthy individuals is an important step in characterizing the antigenic recognition process and the functioning of the adaptive immune system. It also presents an important starting point for describing statistical deviations seen in individuals with compromised immune responses.

High throughput sequencing experiments in different cell types and species [46, 48, 49, 60, 61, 62, 63, 51] have allowed for the quantification of clone sizes and their distributions [46, 64, 50, 51]. Previous population dynamics approaches to repertoire evolution have taken great care in precisely modeling these processes for each compartment of the population, through the various mechanisms by which cells grow, die, communicate, and change phenotype [65, 66, 67, 68, 69, 70]. However, one of the most striking properties of repertoire statistics revealed by high-throughput sequencing is the observation of power laws in clone size distributions (see Fig. 4.1A-B), which holds true for various species (human, mice, zebrafish), cell type (B and T cells) and subsets (naive and memory, CD4 and CD8), and seems to be insensitive to these context-dependent details. It remains unclear, however, what universal features of these dynamics lead to the observed power-law distributions. Here we identify the key biological parameters of the repertoire dynamics that govern its behavior.

The wide range and types of interactions that influence a B or T cell fate happen in a complex, dynamical environment with inhomogeneous spatial distributions. They are difficult to measure *in vivo*, making their quantitative characterization elusive. Motivated by the universality of the observed clone size distribution, we describe the effective interaction between the immune cells and their environment as a stochastic process governed by only a few relevant parameters. All cells proliferate and die depending on the strength of antigenic and cytokine signals they receive from the environment, which together determine their net growth rate (Fig. 4.1 C). This effective fitness that fluctuates in time is central to our description. We find that its general properties determine the form of the clone size distribution. We distinguish two broad classes of models, according to whether these fitness fluctuations are clone specific (mediated by their specific BCR or TCR) or cell specific (mediated by phenotypic fluctuations such as the

number of cytokine receptors). We identify the models that are compatible with the experimentally observed distributions of clone sizes. These distributions do not depend on the detailed mechanisms of cell signaling and growth, but rather emerge as a result of self-organisation, with no need for fine-tuned interactions. Performing a series of validated approximations we find a simple algebraic relationship constraining the different timescales of the problem by the experimentally observed exponent of the clone-size distribution. This result allows for testable predictions and estimates of the rates that govern the diversity of a clonal distribution.

4.4 Results

4.4.1 Clone dynamics in a fluctuating antigenic landscape

The fate of the cells of the adaptive immune system depends on a variety of clone-specific stimulations. The recognition of pathogens triggers large events of fast clone proliferation followed by a relative decay, with some cells being stored as memory cells to fend off future infections. Naive cells, which have not yet recognized an antigen, do not usually undergo such extreme events of proliferation and death, but their survival relies on short binding events (called "tickling") to antigens that are natural to the organism (self-proteins) [71, 72]. Because receptors are conserved throughout the whole clone (with the exception of B cell hypermutations), clones that are better at recognizing self-antigens and pathogens will on average grow to larger populations than bad binders. By analogy to Darwinian evolution, they are "fitter" in their local, time-varying environment.

We first present a general model for clonal dynamics that accounts for the characteristics common to all cell types, following previous work by de Boer, Perelson and collaborators [73, 67, 74]. We later explore the effect of specific features such as hypermutations, memory/naive compartmentalization and thymic output decay on the clone size distribution.

We denote by $a_j(t)$ the overall concentration of an antigen j as a function of time. We assume that after its introduction at a random time t_j, this concentration decays exponentially with a characteristic lifetime of antigens λ^{-1}, $a_j(t) = a_{j,0}e^{-\lambda(t-t_j)}$ as pathogens are cleared out of the organism, either passively or through the action of the immune response. Lymphocyte receptors are specific to certain antigens, but this specificity is degenerate, a phenomenon refered to as cross-reactivity or poly-specificity. The extend to which a lymphocyte expressing receptor i interacts with antigen j (foreign or self) is encoded in the cross-reactivity function K_{ij}, which is zero if i and j do not interact, or a positive number drawn from a distribution to be specified, if they do. In general, interactions between lymphocytes and antigens effectively promote growth and suppress cell death, but for simplicity we can assume that the effect is restricted to the division rate. In a linear approximation, this influence is proportional to $\sum_j K_{ij}a_j(t)$, *i.e.* the combined effect of all antigens j for which clone i is specific. This leads to the following dynamics for the evolution of the size C_i of clone i (Fig. 4.1 C):

$$\frac{dC_i}{dt} = \left(\nu + \sum_j K_{ij}a_j(t) - \mu\right) C_i + B\xi_i(t), \tag{4.1}$$

where ν and μ are the basal division and death rates, and where $B\xi_i(t)$ is a birth-death noise of intensity $B^2 = (\nu + \sum_j K_{ij}a_j(t) + \mu)C_i$, with $\xi_i(t)$ a unit Gaussian white noise (see Appendix A.1 for details about birth death noise).

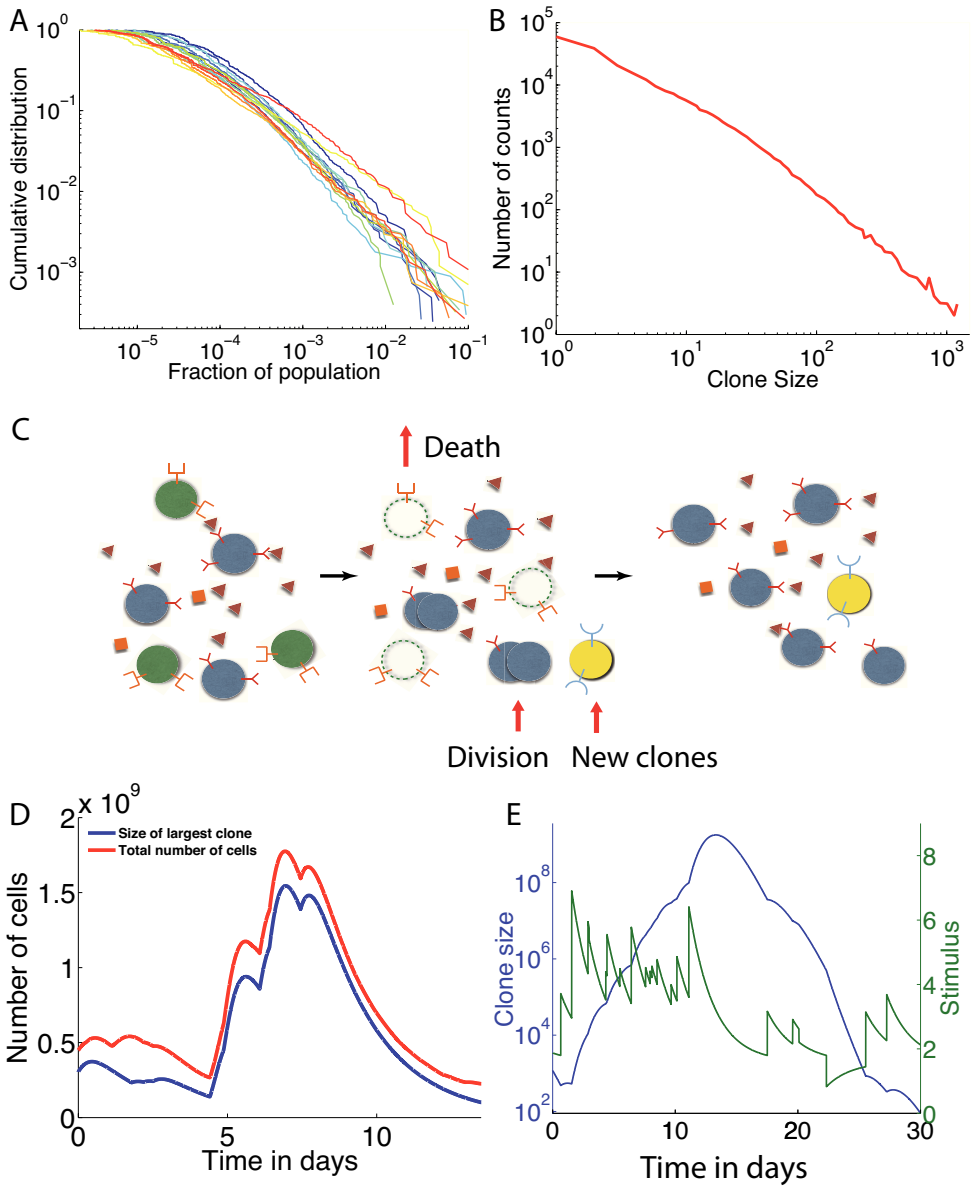

Figure 4.1: Experimental clone size distributions have heavy tails. **A.** B cell zebrafish experimental cumulative clone size distribution for fourteen fish as a function of the fraction of the population occupied by that clone from data in Weinstein et al. [46]. **B.** Clone size distribution for murine T-cells from Zarnitsyna et al. [50] (data plotted as presented in original paper). **C.** The dynamics of adaptive immune cells include specific interactions with antigens that promote division and prevent cell death. New cells are introduced from the thymus or bone marrow with novel, unique receptors. Division, death and thymic or bone marrow output on average balance each other to create a steady state population. **D-E.** Example trajectories from simulations of the immune cell population dynamics in Eq. 4.1. The total number of cells (D) shows large variations after an exceptional event of a large pathogenic invasion. One or a few cells that react to that specific antigen grow up to a macroscopic portion of the total population, and then decrease back to normal sizes after the invasion. A typical clone size trajectory along with its pathogenic stimulation $\sum_j K_{ij} a_j(t)$ shows the coupling between clone growth and variations of the antigenic environment (E). Parameters used: $s_C = 2000$ day^{-1} , $C_0 = 2$, $s_A = 1.96 \cdot 10^7$ day^{-1}, $a_{j,0} = a_0 = 1$, $\lambda = 2$ day^{-1}, $p = 10^{-7}$, $\nu = 0.98$ day^{-1}, $\mu = 1.18$ day^{-1}.

New clones, with a small typical initial size C_0, are constantly produced and released into the periphery with rate s_C (Fig. 4.1 C). For example, a number of the order of $s_C = 10^8$ new T-cells are output by the thymus daily in humans [38]. Since the total number of T cells is of the order of 10^{11}, this means that the net effect of cell death and proliferation results in a negative average growth rate of 10^{-3} days^{-1} in homeostatic conditions [38]. Because the probability of rearranging the exact same receptor independently is very low ($< 10^{-10}$) [40], we assume that each new clone is unique and comes with its own set of cross-reactivity coefficients K_{ij}. Assuming a rate s_A of new antigens, the average net growth rate in Eq. 4.1 is $f_0 = \nu + \langle a_{j,0}\rangle\langle K\rangle s_\mathrm{A}\lambda^{-1} - \mu < 0$, and the stationary number of clones should fluctuate around $N_C \approx s_C|f_0|^{-1}$ clones. This is just an average, and treating each clone independently may lead to large variations in the total number of cells (*i.e.* the sum of sizes of all clones). To maintain a constant population size, clones compete with each other for specific resources (pathogens or self-antigens) and homeostatic control can be maintained by a global resource such as Interleukin 7 or Interleukin 2. Here we do not model this homeostatic control explicitly, but instead assume that the division and death rates ν, μ are tuned to achieve a given repertoire size. We verified that adding an explicit homeostatic control did not affect our results (see Fig. S2 and Appendix A.2).

We simulated the dynamics of a population of clones interacting with a large population of antigens. Each antigen interacts with each present clone with probability $p = 10^{-7}$, and with strength K_{ij} drawn from a Gaussian distribution of mean 1 and variance 1 (truncated to positive values). Although it has been argued that the breadth of cross-reactivity and affinity to self-antigens are correlated [75, 76], here for simplicity we draw them independently, as we do not expect this correlation to qualitatively affect the results. A typical trajectory of the antigenic stimulation undergone by a given clone, $\sum_j K_{ij}a_j$, is shown in Fig. 4.1E (green curve), and shows how clone growth tracks the variations of the antigenic environment. When the stimulation is particularly strong, the model recapitulates the typical behaviour experimentally observed at the population level following a pathogenic invasion [77, 78], as illustrated in Fig. 4.1D: the population of a clone explodes (red curve), driving the growth of the total population (blue curve), while taking over a large fraction of the carrying capacity of the system, and then decays back as the infection is cleared.

On average, the effects of division and death almost balance each other, with a slight bias towards death because of the turnover imposed by thymic or bone marrow output. However, at a given time, a clone that has high affinity for several present antigens will undergo a transient but rapid growth, while most other clones will decay slowly towards exctinction. In other words, locally in time, the antigenic environment creates a unique "fitness" for each clone. Since growth is exponential in time, these differential fitnesses can lead to very large differences in clone sizes, even if variability in antigen concentrations or affinities are nominally small. We thus expect to observe large tails in the distrubution of clone size. Fig. 4.2A shows the cumulative probability distribution function (CDF) of clone sizes obtained at steady state (blue curve) showing a clear power-law behaviour for large clones, spanning several decades.

The exponent of the power-law is independent of the introduction size of clones (see inset of Fig. 4.2A), and the specifics of the randomness in the environment (exponential decay, random number of partners, random interaction strength) as long as its first and second moment are kept fixed (See Fig. S3 and Appendix A.3).

4.4.2 Simplified models and the origin of the power law

To understand the power-law behavior observed in the simulations, and its robustness to various parameters and sources of stochasticity, we decompose the overall fitness of a clone at a given time (its instantaneous growth rate) into a constant, clone-independent part equal to its average $f_0 < 0$, and a clone-specific fluctuating part of zero mean, denoted by $f_i(t)$. This leads to rewriting Eq. 4.1 as:

$$\frac{dC_i}{dt} = [f_0 + f_i(t)]C_i(t) + B\xi_i(t), \tag{4.2}$$

with $B^2 \approx (|f_0| + 2\mu)C_i$.

The function $f_i(t)$ encodes the fluctuations of the environment as experienced by clone i. Because antigens can be recognized by several receptors, these fluctuations may be correlated between clones. Assuming that these correlations are weak, $\langle f_i(t)f_j(t')\rangle \approx 0$, amounts to treating each clone independently of each other, and thus to reducing the problem to the single clone level. The stochastic process giving rise to $f_i(t)$ is a sum of Poisson-distributed exponentially decaying spikes. This process is not easily amenable to analytical treatment, but we can replace it with a simpler stochastic process with the same temporal autocorrelation function. This autocorrelation is given by $\langle f_i(t)f_i(t')\rangle = A^2 e^{-\lambda|t-t'|}$, with the antigenic noise strength $A^2 = s_A p a_0^2 \langle K^2\rangle \lambda^{-1}$, and where we recall that λ^{-1} is the characteristic lifetime of antigens. The simplest process with the same autocorrelation function is given by an overdamped spring in a thermal bath, or Ornstein-Uhlenbeck process,

$$\frac{df_i}{dt} = -\lambda f_i + \sqrt{2}\gamma\eta_i(t), \tag{4.3}$$

with $\eta_i(t)$ a Gaussian white noise of intensity 1 and $\gamma = A\sqrt{\lambda}$ quantifies the strength of variability of the antigenic environment (see Appendix A.4). This is also the process of maximum entropy or caliber [79] with that autocorrelation function (see Appendix A.5 and [80]).

The effect of the birth death noise $B\xi_i(t)$ is negligible when compared to the fitness variations for large clones and it has no effect on the tail (see Fig. S5 and Appendix A.6). It can thus be ignored when looking at the tail of the distribution and its power law exponent, but it will play an important role for defining the range over which the power law is satisfied.

The population dynamics described by Eqs. 4.2 and 4.3 can be reformulated in terms of a Fokker-Planck equation for the joint abundance ρ of clones of a given log-size $x = \log C$ and a given fitness f:

$$\frac{\partial\rho(x,f,t)}{\partial t} = -(f_0 + f)\frac{\partial\rho}{\partial x} + \lambda\frac{\partial(f\rho)}{\partial f} + \gamma^2\frac{\partial^2\rho}{\partial f^2} + s(x,f), \tag{4.4}$$

where the source term $s(x,f)$ describes new clones arriving at rate s_C with size C_0 and normally distributed fitnesses of variance $\langle f^2\rangle = \gamma^2/\lambda$. This Fokker-Planck equation can be solved numerically with finite element methods with an absorbing boundary condition at $x = 0$ to account for clone extinction. The solution, represented by the black curve in Fig. 4.2A, matches closely that of the full simulated population dynamics (in blue). The power-law behaviour is apparent above a transition point that depends on the distribution of introduction sizes of new clones and the parameters of the model (see below). Intuitively, the microscopic details of the noise are not expected to matter when considering long time scales, as a consequence of the central limit theorem. However, the long tails of the distribution of clone sizes involve rare events and belong to the regime of large deviations, for which these microscopic details may be important. Therefore, the agreement between the process described by the overdamped spring

and the exponentially decaying, Poisson distributed antigens is not guaranteed, and in fact does not hold in all parameter regimes (see Fig. S8).

We can further simplify the properties of the noise by assuming that its autocorrelation time is small compared to other timescales. This leads to taking the limit $\gamma, \lambda \to \infty$ while keeping their ratio constant $\sigma = \gamma/\lambda$ constant, so that $f_i(t)$ is just a Gaussian white noise with $\langle f_i(t) f_i(t') \rangle = 2\sigma^2 \delta(t - t')$ (see Appendix A.6 and Fig. S4). The corresponding Fokker-Planck equation now reads

$$\partial_t \rho(x,t) = -f_0 \partial_x \rho(x,t) + \sigma^2 \partial_x^2 \rho(x,t) + s(x), \tag{4.5}$$

with $s(x) = s_C \delta(x - \log(C_0))$. This equation can be solved analytically at steady state, and the resulting clone size distribution is, for $C > C_0$:

$$\rho(C) = \frac{s_C}{\alpha \sigma^2} \frac{1}{C^{\alpha+1}}, \tag{4.6}$$

with $\alpha = |f_0|/\sigma^2 = \lambda |f_0|/A^2$ (details in Appendix A.6). The full solution, represented in Fig. 4.2A in red, captures well the long-tail behaviour of the clone size distribution despite ignoring the temporal correlations of the noise, and approaches the solution of the colored-noise model (Eq. 4.3) as $\lambda, \gamma \to \infty$, as expected (see Fig. 4.2A).

The power law behaviour and its exponent depend on the noise intensity, but are otherwise insensitive to the precise details of the microscopic noise, including its temporal properties. Fat tails (small α) are expected when the average cell lifetime is long (small $|f_0|$) and when the antigenic noise is high (large σ or A). The explicit expression for the exponent of the power law $1 + \alpha$ as a function of the biological parameters can be used to infer the antigenic noise strength A^2 directly from data. The typical net clone decay rate $|f_0| \approx 10^{-3}$ can be estimated from thymic output and repertoire size, as discussed earlier. The characteristic lifetime of antigens λ^{-1} is harder to estimate, as it corresponds to the turnover time of the antigens that the body is exposed to, but is probably of the order of days or a few weeks, $\lambda \approx 0.1$ day^{-1}. We estimated $\alpha = 1 \pm 0.2$ from the zebrafish data of Fig. 4.1A [46, 64] using canonical methods of power-law exponent extraction [81] (see Appendix A.7 for details), and also found a similar value in human T cells [82]. The resulting estimate, $A = 10^{-2}$ day^{-1}, is rather striking, as it implies that fluctuations in the net clone growth rate, A, are much larger than its average f_0.

While the distribution always exhibits a power law for large clones, this behavior does not extend to clones of arbitrarily small sizes, where the details of the noise and how new clones are introduced matter. We define a power-law cut-off C^* as the smallest clone size for which the cumulative distribution function (CDF) differs from its best power-law fit by less than 10%. Using numerical solutions to the Fokker-Planck equation associated to the colored-noise model, we can draw a map of C^* as a function of the parameters of the system. In Fig. 4.2B-C we show how C^* varies as a function of the introduction size for different values of the dimensionless parameter related to the effective strength of antigen fluctuations relative to their characteristic lifetime at fixed power law exponents. In principle one can use this dependency to infer effective parameters from data. In practice, when dealing with data it is more convenient to consider the value of the cumulative distribution at C^*, rather than C^* itself. For example, fixing $C_0 = 4$ and fitting the curve of Fig. 4.1A with our simplified model using λ as an adjustable parameter, we obtain $\lambda \approx 0.14$ day^{-1} (see Appendix A.7), which corresponds to a characteristic lifetime of antigens of around a week. Although this estimate must be taken with care, because of possible PCR amplification biases plaguing the small clone size end of the distribution, the procedure

described here can be applied generally to any future repertoire sequencing dataset for which reliable sequence counts are available.

4.4.3 A model of fluctuating phenotypic fitness

So far, we have assumed that fitness fluctuations are identical for all members of a same clone. However, the division and death of lymphocytes do not only depend on signaling through their TCR or BCR. For example, cytokines are also growth inducers and homeostatic agents [83, 84], and the ability to bind to cytokines depends on single-cell properties such as the number of cytokine receptors on the membrane of a given cell, independent of their BCR or TCR receptor. Other stochastic single-cell factors may affect cell division and death. These signals and factors are *cell* specific, as opposed to the *clone* specific properties related to BCR or TCR binding. Together, they define a global phenotypic state of the cell that determines its time-varying "fitness," independent of the clone and its T-cell or B-cell receptor. This does not mean that these phenotypic fitness fluctuations are independent across the cells belonging to the same clone. Cells within a clone share a common ancestry, and may have inherited some phenotypic properties of their common ancestors, making their fitnesses effectively correlated with each other. However, this phenotypic memory gets lost over time, unlike fitness effects mediated by antigen-specific receptors.

We account for these phenotypic fitness fluctuations by a function $f_c(t)$ quantifying how much the fitness of an individual cell c differs from the average fitness f_0. This fitness difference is assumed to be partially heritable, which we model by:

$$\frac{df_c}{dt} = -\lambda_c f_c(t) + \sqrt{2}\gamma_c \eta_c(t), \tag{4.7}$$

where λ_c^{-1} is the heritability, or the typical time over which the fitness-determining trait is inherited, γ_c quantifies the variability of the fitness trait, and $\eta_c(t)$ is a cell-specific Gaussian white noise of power 1. Despite its formal equivalence with Eq. 4.3, it is important to note that here the fitness dynamics occurs at the level of the single cell (and its offspring) instead of the entire clone. The dynamics of the fitness $f_i(t)$ of a given clone i can be approximated from Eq. 4.7 by averaging the fitnesses $f_c(t)$ of cells in that clone, yielding:

$$\frac{dC_i}{dt} = [f_0 + f_i(t)]C_i(t) + \sqrt{(\nu+\mu)C_i(t)}\xi_i(t), \tag{4.8}$$

$$\frac{df_i}{dt} = -\lambda_c f_i(t) + \frac{1}{\sqrt{C_i(t)}}\sqrt{2}\gamma_c \eta_i(t), \tag{4.9}$$

where $\eta_i(t)$ and $\xi_i(t)$ are clone-specific white noise of intensity 1, and ν and μ are the average birth and death rates, respectively, so that $f_0 = \nu - \mu$ (details in Appendix A.9). The difference with Eq. 4.3 is the $1/\sqrt{C_i(t)}$ prefactor in the fitness noise $\eta_i(t)$, which stems from the averaging of that noise over all cells in the clone, by virtue of the law of large numbers. Because of this prefactor, the fitness noise is now of the same order of magnitude as the birth-death noise, which must now be fully taken into account. Taking Eq. 4.8 and Eq. 4.9 at the population level gives a Fokker-Planck equation with a source term accounting for the import of new clones. We verify the numerical steady state Fokker–Planck solution against Gillespie simulations (Fig S6, see Appendix A.8 for details).

Fig. 4.3A-B show the distribution of clone sizes for different values of the phenotypic relaxation rate λ_c and environment amplitude γ_c. These distributions vary from a sharp exponential

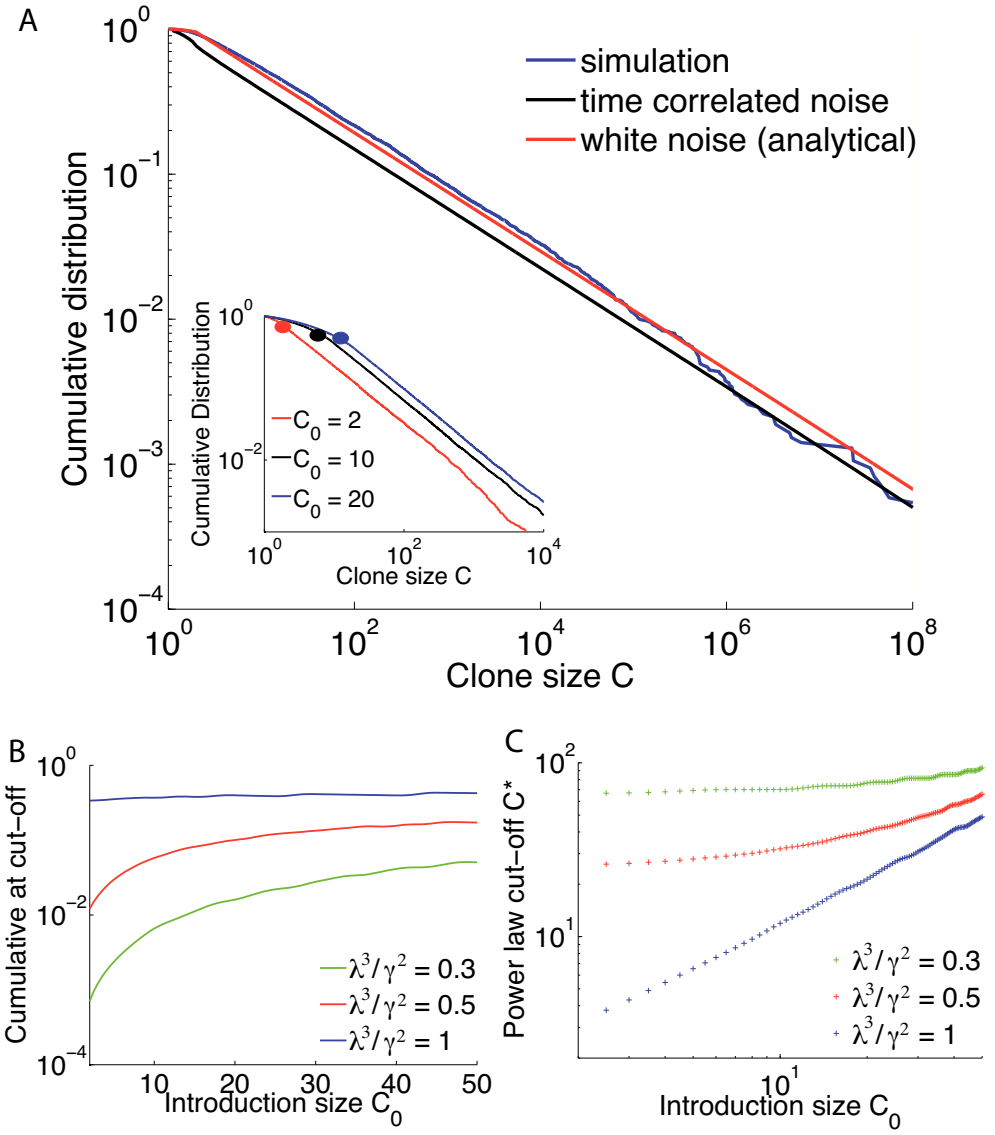

Figure 4.2: Clone size distributions for populations with fluctuating antigenic, clone-specific fitness. **A.** Comparison of simulations and simplified models of clone dynamics. Blue curve: cumulative distribution of clone sizes obtained from the simulation of Eq. 4.1. Black curve: a simplified, numerically solvable model of random clone-specific growth, also predicts a power-law behaviour. Red curve: analytical solution fo the Gaussian white noise model, Eq. 4.4. Parameters used: $\nu = 0.98$ day^{-1}, $\mu = 1.18$ day^{-1}, $\lambda = 2$ day^{-1}, $s_C = 2000$ day^{-1} , $C_0 = 2$, $s_A = 1.96 \cdot 10^7$ day^{-1}. Inset: the exponent is independent of the initial clone size. Results from simulation with different values of the introduction clone size. The cut-off value of the power law behaviour, represented here as a dot, is strongly dependent on the value of C_0. Parameters are $\nu = 0.2$ day^{-1}, $\mu = 0.4$ day^{-1}, $\lambda = 2$ day^{-1}, $\gamma = 1$ day$^{-3/2}$ and $s_C = 5000$. **B.** Value of the cumulative distribution function at the point of the power law cut-off as a function of the introduction clone size C_0 for different values of a dimensionless parameter related to the effective strength of antigen fluctuations relative to their characteristic lifetime λ^3/γ^2 for a fixed power law exponent α. We use the cumulative distribution function because it is robust, invariant under multiplicative rescaling of the clone sizes. This way we do not need to correct directly for PCR multiplication or sampling. Parameters are for B and C $\nu = 4.491$ days^{-1}, $\mu = 5.489$ days^{-1} and $\alpha = -0.998$. **C.** Power-law cut-off as a function of the introduction clone size.

drop in the case of low heritability (large λ_c) to heavier tails in the case of long conserved cell states (small λ_c). To quantify the extend to which these distributions can be described as heavy-tailed, we fit them to a power law with exponential cut-off, $\rho(C) \propto C^{-1-\alpha} e^{-C/C_m}$, where C_m is the value below which the distribution could be interpreted as an (imperfect) power law. Fig. 4.3C shows a strong dependency of this cut-off with the phenotypic memory λ_c^{-1}. The longer the phenotypic memory λ_c^{-1}, the more clone-specific the fitness looks like, and the more the distribution can be mistaken for a power law in a finite-size experimental distribution. Larger birth-death noise also extends the range of validity of the power-law. As a result, and despite the absence of a true power-law behaviour, these models of fluctuating phenotypic fitnesses cannot be discarded based on current experimental data.

The model can be solved exactly at the two extremes of the heritability parameter λ_c. In the limit of infinite heritability ($\lambda_c \to 0$) the system is governed by selective sweeps. The clone with the largest fitness completely dominates the population, until it is replaced by a better one, giving rise to a trivial clone-size distribution. In the opposite limit, when heritability goes to 0 ($\lambda_c \to +\infty$), the Fokker-Planck equation can be solved analytically (see Appendices A.9 and A.10), yielding an exact power-law with exponential cutoff, $\rho(C) \propto C^{-1-\alpha} e^{-C/C_m}$, with $\alpha = -[1 + (\mu+\nu)\lambda_c^2/2\gamma_c^2]^{-1}$ and $C_m = (\mu-\nu)^{-1}[(\mu + d\nu)/2 + \gamma_c^2/\lambda_c^2]$. The numerical solution of Fig. 4.3B is close to this limit. Note that even with a negligible exponential cutoff, the predicted $\alpha < 0$ contradicts experimental observations.

4.5 Discussion

The model introduced in this paper describes the stochastic nature of the immune dynamics with a minimal number of parameters, helping interpret the different regimes. These parameters are effective in the sense that they integrate different levels of signaling, pathways, and mechanisms, focusing on the long timescales of clone dynamics. We assumed that they are general enough that different cell types (B and T cells) or subsets (naive or memory) can be described by the same dynamical equations despite their differences. How do refined models including these differences affect our results?

Naive and memory cells differ in their turn-over rate, *i.e.* their death rate, memory cells being renewed at a pace 10 times faster than naive ones [42]. In our model, this difference is reflected in a higher birth-death noise for memory cells. We have shown that this noise had no effect on the tail of the clone-size distribution for clone-specific fitness (SI Fig. S5), while it was important for the case of a cell-specific fitness, where birth-death noise contributed to the distribution to the same extend as fitness fluctuations. However, some repertoire datasets mix both naive and memory sets, and one could wonder whether our results hold for such mixtures. To examine this question, we simulated a simple two-compartment model where naive cells get irreversibly converted into memory cells when their stimulation is above a certain threshold (see SI, Appendix K for details). We found that, when fitness was clone specific, the clone-size distribution of the mixture and that of memory cells alone still follows a power law, while that of naive cells only does so when conversion to memory upon stimulation is partial (SI Fig. S12). Repeating the same analysis for the cell-specific fitness model, we found that clone-size distributions for each phenotype differed according to their respective birth-death noises, with a longer tail for memory cells as expected from their higher turn-over rate.

The main difference between B and T cells ignored by our model is that B cell receptors

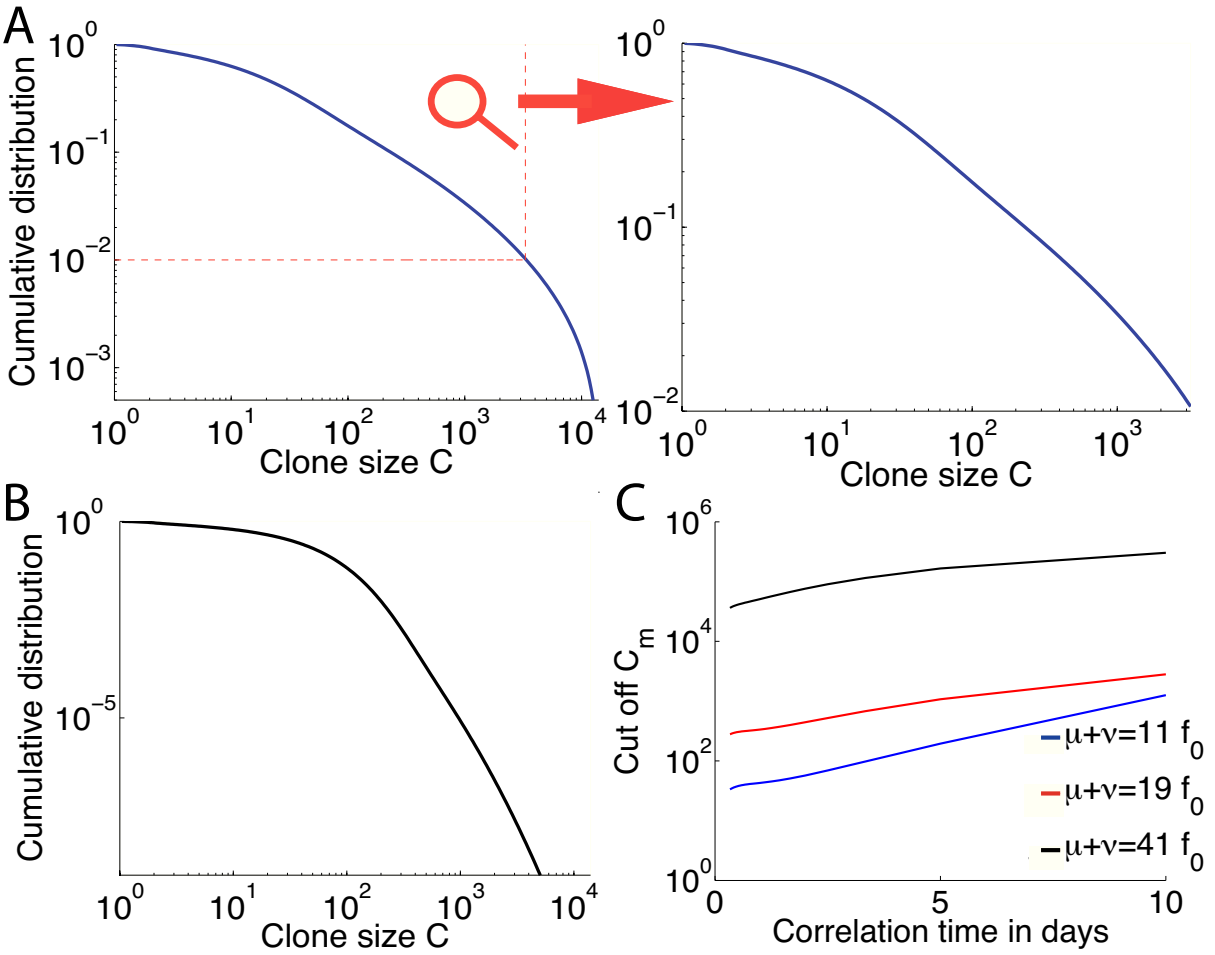

Figure 4.3: Clone size distributions for populations with a cell specific fluctuating phenotypic fitness. **A.** Cumulative distribution of clone sizes for moderate phenotypic heritability (λ_c^{-1}). The distribution is power-law like for small clone values and drops above a cut-off around 0.01 of clone size probability. An experiment that does not sequence the repertoire deeply enough could report a power law behavior (see zoom). Parameters are $\nu = 0.17$ days^{-1}, $\mu = 0.3$ day^{-1}, $\lambda_c = 0.4$ days^{-1} and $\gamma_c = 0.5$ days$^{-3/2}$. $C_0 = 2$ for all three graphs. **B.** An example of a distribution of clone sizes from a cell-specific model with very low environmental noise, close to the pure birth-death limit. The distribution is flat ($\alpha = 0$) and then drops exponentially. It does not resemble experimental data. Parameters are $\nu = 0.1$ days^{-1}, $\mu = 0.3$ days^{-1}, $\lambda_c = 2$ days^{-1} and $\gamma_c = 5$ days$^{-3/2}$. **C.** Value of the cumulative distribution at the exponential cut-off as a function of the speed of environment variations λ_c, for different birth-death noise levels. Parameters are $f_0 = -0.998$ days^{-1} and $f_0\lambda_c^2/\gamma_c^2 = 0.998$.

accumulate hypermutations upon proliferation. We studied this effect by allowing proliferating clones to spawn new clones with slightly modified affinities to antigens (see SI, Appendix L). The resulting clone-size distribution still follows a power-law (Fig. S13), although with a slightly smaller exponent due to increased stochasticity.

Another simplifying assumption of our model is that the dynamics reaches a steady state. This may be challenged by the decay of the thymic output s_C with age. To estimate the importance of this effect, we simulated the model of a clone-specific fitness with an exponentially decaying source term, combined with a decreasing $|f_0|$ chosen to keep the population constant on average (see SI Appendix M). The clone-size distributions at different points in time, shown in Fig. S14, still follow a power law. Interestingly, the exponent α is predicted to decrease with age, consistent with $\alpha \propto |f_0|$.

We showed that the relevant sources of stochasticity for the shape of the clone–size distributions fall into two main categories, depending on how cell fate is affected by the environment. Either the stochastic elements of clone growth act in a clone-specific way, through their receptor (BCR or TCR), leading to power-law distributions with exponent ≥ 1, or in a cell-specific way, *e.g.* through their variable level of sensitivity to cytokines (and more generally through any phenotypic trait affecting cell fitness), leading to exponentially decaying distributions with a power-law prefactor. These two types of signals (clone specific and cell specific) are important for the somatic evolution of the immune system [83, 84, 85, 86, 87, 74] and our analysis shows that the shape of the clone size distribution is informative of their relative importance to the repertoire dynamics. It provides a first theoretical setting and an initial systematic classification for modeling immune repertoire dynamics. Our method applied to high-throughput sequencing data can be used to quantify how much each type of signal contributes to the overall dynamics, and what is the driving force for the different cell subsets. For example, although it is reasonable to speculate that clone-specific signals should dominate for memory cells (through antigen recognition), and cell-specific selection for naive cells (through cytokine-mediated homeostatic division), the relative importance of these signals for both cell types is yet to be precisely quantified, and may vary across species. A clear power law over several decades would strongly hint at dynamics dominated by interactions with antigens, while a faster decaying distribution would favor a scenario where individual cell fitness fluctuations dominate. Applying these methods to data from memory cells can give orders of magnitude for the division and half-life of memory lymphocytes, as well as the typical number of cells C_0 from a clone that are stored as memory following an infection.

The application of our method to data from the first immune repertoire survey (B cell receptors in zebrafish [46]) suggests that clone-specific noise dominates in that case, allowing us to infer a relation between the dynamical parameters of the model from the observed power-law exponent ≈ 2. However, there are a few issues with applying our method directly to data in the current state of the experiments. First, the counts (*i.e.* how many cells have the same receptor sequence and belong to the same clone) from many high-throughput repertoire sequencing experiments are imperfect because of PCR bias and sampling problems. New methods using single-molecule barcoding have been developed for RNA sequencing [88, 89, 63], but they do not solve the problem entirely, as the number of expressed mRNA molecules may not faithfully represent the cell numbers because of possible expression bias. In addition, most studies (with the exception of [90]) have been sequencing only one of the two chains of lymphocyte receptors, which is insufficient to determine clone identity unambiguously. As methods improve, however,

our model can be applied to future data to distinguish different sources of fitness stochasticitiy and to put reliable constraints on biological parameters. Studying clone size distributions in healthy individuals allows us to characterize signatures of normally functioning immune systems. By comparing them to the same properties in individuals suffering from immune diseases or cancer, our approach could be used to identify sources of anomalies.

Thanks to its generality, our model is also relevant beyond its immunological context, and follows previous attempts to explain power laws in other fields [91, 92, 93]. The dynamics described here corresponds to a generalization of the neutral model of population genetics [9] where thymic or bone marrow outputs are now reintepreted as new mutations or speciations, and where we have added a genotypic or phenotypic fitness noise (receptor or cell-specific noise, respectively). It was recently shown that such genotypic fitness noise strongly affects the fixation probability and time in a population of two alleles [94, 95]. Note that, since new thymic or bone marrow clones are unrelated to existing clones, there are no lineage histories, in contrast to previous theoretical work on evolving populations in fluctuating fitness landscapes [96, 97, 98]. Our main result (Eq. 4.6) shows how fitness noise can cause the clone-size distribution (called frequency spectrum in the context of population genetics) to follow a power law with an *arbitrary* exponent > 1 in a population of fixed size, while the classical neutral model gives a power law of exponent 1 with an exponential cut off (as shown in our exact solution with $\gamma_c = 0$). Our results can be used to explain complex allele frequency spectra using fluctuating fitness landscapes.

Chapter 5

Random networks of immune systems: structure and selection

This work is destined to publication once more material has been produced, some of the questions are not answered yet.

This section contains unpublished work that goes beyond the steady state mean-field approach of the previous chapter. Indeed, it has been proven that non-obvious effects can arise in population dynamics due to discretization and spatial dimensions [99] for instance. I first describe how the niche structure of the immune system is related to an increase in fitness of clones with time. In the second part, I investigate the structure of the graph of interactions between cells of the immune system and the antigenic environment and study its effect on clone size distributions. Chapter 4 was mostly dedicated to the dynamics of effector cells and fluctuating pathogenic environment. This chapter focuses on a fixed environment and the dynamics and the competition for self-peptides (although some results can be used in both systems).

5.1 Selection and fitness change with de novo mutations

5.1.1 Introduction

In standard Darwinian competition dynamics, mutations occur from existing clones and are selected through competition. Sweeps and selection of fitter and fitter individuals have been studied in this case extensively (see for instance [100] for a recent example). In many systems, however, most new genotypes do not come from the existing clones but are produced by independent external sources. Such situations can arise in populations of animals or bacteria if members of new clones wander into the system at a rate much higher than the rate of mutations. The main motivation for this model is the immune system. The most obvious fitness factor for adaptive immune cells is their receptor (BCR or TCR). This receptor is inherited by daughter cells during clonal expansion (with the exception of hypermutations), so no new genetic material is produced by direct mutations from existing clones. Of course, errors in replication can occur but the contribution from such errors compared to introduction of new materials by the thymus or the bone marrow is very small with the exception of the enhanced level of mutations during B-cell clonal expansion with hypermutations. Indeed, the effect of point-mutations on receptors is very small compared to the high variability of VDJ recombinations [39]. The thymus and bone

marrow continuously provide the system with new genetic lineages. Antigen binding is clearly essential for the division and survival of effector cells and it also has been proven to contribute to naive repertoire homeostasis [43]. The fight for antigenic resources calls for a definition of clone fitness. In this section, I present a model of fitness change in systems with de novo mutations and discuss its application to the immune system.

The main difference between this model and the previous one is that it no longer assumes that the environment fluctuates. This model is mostly motivated by the naive immune system, where self-peptides can be assumed to be constant. It can also be used as a basis for a model of the pool of memory cells,where the pathogenic environment varies slowly.

5.1.2 From biology to model

This model assumes the existence of a shape space of receptors and antigens that is closely related to fitness. This shape space is assumed to have a given dimension d, and both clones and resources (antigens) are represented by a position in $\mathbb{R}^d$ in this shape space. Clones compete for resources but can only access them if they are close, where closeness is defined in terms of an interaction kernel. The interaction kernel is a function of the positions of the antigen and the receptor in shape space and gives a value that represents the strength of their interaction. It is also assumed that the clonal reservoir, from which newcomers are introduced, is large enough and varied enough to never send the same genetic material twice. So the equations that describe the evolution of existing clones have no external source terms because the only source of new cells in the clone is division. The model assumes that the resources in shape space are drawn from a given probability distribution. This recognition space is an effective projection of the high dimensional genotypic space onto the function of the lymphocytes that is the binding to antigens. In the simple case where the expected density of resources in shape space is low enough (and there is no region in shape space where the probability density accumulates), then the resources can be assumed to be far enough from each other to create independent niches. In this limit it is possible to derive exactly the distribution of fitness of the main clones occupying these niches. It can then be used as a first building block for more complex models, including niches that partially overlap (some clones can access different resources) in the direction of [65], but on a global population level.

One of the goals of this model is to discuss aging in the immune system and see if clone size distributions contain hints of the structure of shape space and of the resource distributions. The model presented here has a fixed antigenic space, extension to varying antigenic space are discussed at the end of the section.

5.1.3 Model of a niche

Let us consider a system made of N independent niches. Independent here means that each niche is approximatively undisturbed by whatever is happening in other niches. A system of independent niches is locally reduced to one niche. To develop ideas, let us consider a system of clones $i \in I$ (cells or individuals with the same receptor sequence) with population size C_i and fitnesses f_i. Let us also assume exponential growth as a function of fitness:

$$\partial_t C_i = f_i C_i + \sqrt{(\mu(f_i) + \nu(f_i))C_i}\xi, \tag{5.1}$$

where μ and ν are respectively the birth and death rate of the cells of the clone, and ξ is the standard Gaussian white noise. In general there can be different parameterizations of fitness, the simplest one is to define it as the exponential growth factor (as is done here). The fitness is directly related to the birth and death rates: $f_i = \mu(f_i) - \nu(f_i)$. f_i is a decreasing function of the number of competitors and their fitness. The second term accounts for the birth-death randomness in the population and can be ignored in the first approximation as being much smaller for large clones than the fitness variations.

In a fixed environment with no arrival of new clones, the system will equilibrate to a state where one clone C_{i^*} outcompetes all others in the limit of long times. In a system kept out of equilibrium by the arrival of new clones, at any given time, one clone will have the highest fitness. In a first approximation, we assume that the arrivals are separated by long enough time intervals. Then at any given time, there is one clone dominating the niche and all the other ones are outcompeted and decaying at rates depending on their relative fitness compared to the dominant clone. As time goes by and fitter and fitter individuals are selected (assuming the fitness distribution of new individuals is constant), the expectation time for arrival of a clone fitter than C_{i^*} goes to infinity and we are in a regime of the separation of the time scales.

Therefore, on long time scales, we can assume that in each niche one clone is selected as the dominant, while all the other ones are slowly decaying after entering the system. For each niche j, we set the winner clone index to be $i_j = i^*$ and so we get a bimodal distribution of clones. Some of them belong to a pool of winners

$$W = \{C_{i_j}\}_j, \tag{5.2}$$

and all the other clones belong to the (*a priori* much larger) pool of losers L.

5.1.4 The dynamics of the winners' pool

Let us consider a set of N niches (where $N \gg 1$) and the distribution $\rho(f,t)$ of fitness f in the W pool (and $\Gamma(f,t)$ the cumulative distribution of $\rho(f,t)$). We also define $\rho_0(f)$ as the distribution of fitness f in the newly introduced clones (assumed to be independent of time), $\Gamma_0(f)$ its cumulative distribution function, and θ the number of new clones introduced per time unit. Intuitively, $\Gamma(f,t)$ selects and keeps a limited number of the highest extreme values of $\Gamma_0(f)$ up to time t.

Having several independent niches is completely equivalent to replicating several times the niche process. The ensemble average - ρ, is the same as a time or realization average over one niche. The probability for one niche j to have its dominant clone replaced is equal to the probability of having a new clone entering the niche in this time unit (θ/N) times the probability for this clone to be better than the existing one $1 - \Gamma_0(f_{i_j})$.

Following the above argument we obtain the Master equation for the distribution of fitness of clones in the winners' pool:

$$\partial_t \rho(f,t) = \frac{\theta}{N}\left(\rho_0(f)\int_0^f df'\rho(f',t) - \rho(f,t)\int_f^{+\infty} df'\rho_0(f')\right), \tag{5.3}$$

assuming fitness can only be positive (other cases would only change the integration boundaries). In particular, if there is a cutoff in the fitness distribution, the integral can either be cut or

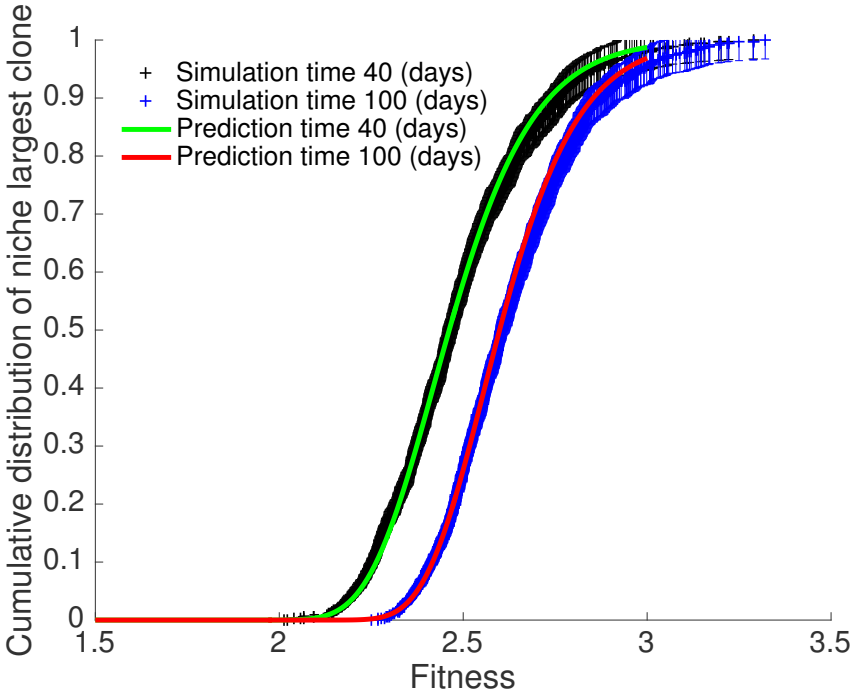

Figure 5.1: Simulation and prediction (Eq. 5.6) of the cumulative distribution of fitness of the largest clone in independent niches with a Gaussian distribution of fitnesses of new clones. In blue and black are the results of simulations with error bars.The dynamics are simple with exponential growth and linear fitness factor (Eq. 5.8). Parameters are: $N^* = 10^5$ individuals, introduction size 2 individuals, introduction rate is 10 new clones per day, new fitness is drawn from Gaussian with mean 1 and standard deviation 0.5. The number of niches is 1000. Note that increasing the number of niches will decrease the size of the error bars.

assumed to be 0 above a certain value. Using the definitions of the cumulative distribution functions and rewriting Eq. 5.3, we obtain

$$\partial_t \rho(f,t) = \frac{\theta}{N}\left(\rho_0(f)\Gamma(f,t) - \rho(f,t)(1-\Gamma_0(f))\right). \tag{5.4}$$

The right hand side is the derivative of a product. Integrating over f on both sides we get that

$$\partial_t \Gamma(f,t) = -\frac{\theta}{N}\Gamma(f,t)(1-\Gamma_0(f)). \tag{5.5}$$

Renormalising time to define $s = \theta t/N$ we find

$$\Gamma(f,s) = \Gamma(f,0)e^{-\Lambda_0(f)s}, \tag{5.6}$$

where $\Lambda_0 = 1 - \Gamma_0$. We obtain for the distribution of fitness of clones in the winner pool

$$\rho(f,s) = e^{-\Lambda_0(f)s}\left(\rho(f,0) - \Gamma_0(f,0)\rho_0(f)s\right). \tag{5.7}$$

5.1.5 The case of independent niches

We simulate the dynamics of independent niches. Each individual belongs to one niche and competes only with the other individuals of the niche. Each new clone is introduced with a

new random fitness drawn from a fixed distribution. In Fig. 5.1, the distribution of new fitness is Gaussian but the result holds for any distribution. A non-Gaussian fitness distribution will simply lead to a different shape. The dynamics are taken to be

$$\partial_t C_i = f_i C_i - \frac{N}{N^*} C_i, \tag{5.8}$$

where N^* is the carrying capacity of the system when the mean fitness is equal to 1 (and is proportional to the carrying capacity for other values of the mean fitness). Different coefficients or exponents for the homeostasis term (the second term) will yield the same results. We simulate this system and find perfect agreement with the theory for the distribution of fitness of the largest clone of each niche (see Fig. 5.1).

Let us now consider a slightly more complex case. We describe the space of genotypes as a low dimensional space. Resources are allocated in this genotype space and each new clone comes in with a random position in this space. The fitness is then a decreasing function of the distance to resources, and each clone competes with other clones only through fitness. So effectively clones compete only with clones that are close enough to the same resources. In this limit, clones that are not drawn close enough to any resource quickly die, and there are empty regions in shape space (see Fig 5.3 A). We consider the limit in which the fitness function decays faster than the typical distance between two resources (in the shape space), and statistically each resource creates its own independent niche.

For each niche, the relevant fitness parameter is simply the distance to the resource. We simulate this system for a two dimensional genotype space with a Gaussian interaction kernel (fitness decays as the exponential of the square of the distance to the resource). Clones locally compete for resources through the interaction matrix K. The interaction term between clone i and resource j is given by

$$K_{i,j} = e^{-d_{i,j}/l^2}, \tag{5.9}$$

where l is the typical distance, above which clone-resource interaction decays. It is smaller than the typical distance between two resources. $d_{i,j}$ is the Euclidian distance. The choice of the Euclidian distance is arbitrary because recognition space is an effective space and any other L^p norm could also be valid. In the simulation example of Fig. 5.2 I have chosen this to be the Euclidian distance. We assume periodic boundary conditions to avoid artificial boundary effects on the distribution of clones. Resources are randomly and uniformly distributed in space, and so are the positions of initial and new clones. We then describe a typical Lotka-Volterra dynamics with competition where each resource has availability

$$F_j = \frac{1}{1 + \sum_i K_{ij} C_i} \tag{5.10}$$

and each clone feels the stimulus

$$S_i = \sum_j K_{ij} F_j. \tag{5.11}$$

This model is inspired by the classic models of [57, 58] described in Chapter 3.3.2.

Instead of writing the equations in terms of fitness, we write them directly in term of the distance to the local resource (as fitness here is a decreasing function of distance). In order to keep the function increasing, we write the equations in terms of the inverse distance. The goal is now to determine what is the probability for a newcomer to enter a niche and beat all its

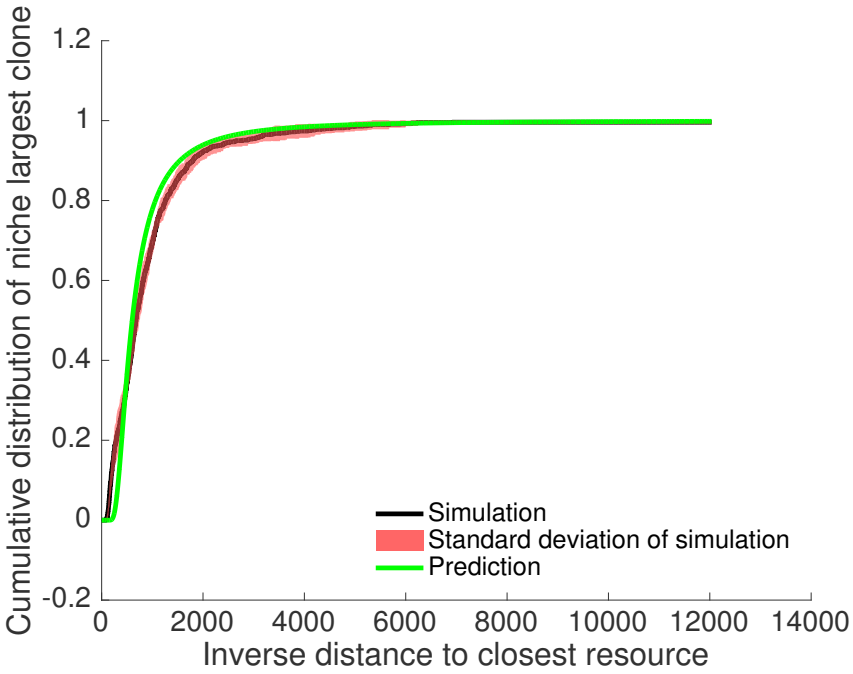

Figure 5.2: Simulation and prediction for Γ, the cumulative distribution of fitness of the largest clone in independent niches on a two dimensional shape space with Gaussian kernel. The prediction is given by Eq. 5.6 with Γ_0 defined as in Eq.5.12. Parameters are: $V = 1$, $l = 0.01$, $\theta = 8$ day^{-1}, $\alpha = 1$ day^{-1}, $\beta = 0.01$ day^{-1}.

competitors in this model. To beat all competitors, a newcomer must enter a disk of radius smaller than the best existing clone's distance to the resource.

We expect again competitive exclusion in each niche. The probability for a newcomer to be at a distance smaller than d from the resource is the ratio of the volume of a sphere of radius d to the total volume V of shape space (in the case of infinite shape space volume and infinite introduction rate V is a ratio of these constants). So the cumulative distribution function of the fitness of the newly introduced sequences is a function of the inverse distance

$$\Gamma_0(1/d) = \frac{\pi^{n/2} d^n}{\Gamma(\frac{n}{2}+1)V}, \tag{5.12}$$

where d is the distance to the resource, and n is the dimension of shape space. A discussion of the value of n in realistic biological terms is given in [14]. The formula is simply the ratio of the probability to be within the hypersphere of radius d to the total volume.

We check the validity of Eq. 5.12 with simulated data in Fig. 5.2. The cumulative distribution solution is not exact, but only true in the limit of low resource density, which explains the small discrepancy between the prediction and the simulation. Nevertheless we find good agreement with the theory.

5.1.6 Prospects and discussion

This simple model can be used as a building block of more complex models of niche interactions and selection. These possibilities have not been explored yet due to lack of time.

A first direction is to relax the assumption of fully independent niches and allow the density of resources to increase. The network becomes much more complex. A realistic goal is then to derive an approximate formula based on the connectivity of the network by accounting for the number of resources available to each clone and thus the number of competitors a newcomer would have to overcome. However, as soon as clones rely on several resources for survival, then non-trivial niche occupation replacement dynamics occur (see Fig. 5.3 A for an example). As local competitors have extra resources with availability depending on clones that do not compete for the local resource, the simple equations of the previous sections do not always apply.

This could constitute a good model of the naive repertoire, as it is reasonable to assume that the self-antigen naive cells use as resources are quite stable over a clone's lifetime. Such a model would give insight into the selection of fitter and fitter naive clones with aging, and it would predict an increase in clone size with age that has been observed.

However, aging also includes the reduction of thymic and bone marrow outputs that are alternative explanations to increase of clone size with age. Recent studies have proved that infection by Cytomegalovirus (CMV) is strongly correlated with immunosenescence, the decay of immune functions with aging (see [101] for a complete review on the topic). CMV triggers recurrent inflammation at old age. The exact mechanism that relates it to immunosenescence, and in particular, to the decrease of clonal diversity is still a mystery. The model presented above could be used to represent the memory pool with a fixed antigenic memory. In that case CMV could be an overabundant resource or a resource that has an especially large interaction radius. This model could be used to investigate the effect of this abundant resource on clone size distributions and shape space coverage by the immune system on long time scales and find out if constant inflammation is enough to trigger large-scale population effects in fitness.

A difficulty that arises is that, at long times, the waiting time for niche winner replacement goes to infinity. At the same time, the fitness difference between the newcomer and the existing large clone goes to 0, and so the time it would take a fitter clone to grow bigger than the existing one and outcompete it goes to infinity. The interaction of these time scales must be included with care in model building.

A second direction for model building is to introduce time-varying resources. A model with separation of time scales between competitive exclusion and resource variations could still be tractable. It would provide an interesting model of the complete immune system (including effectors and memory) and should be compared to the steady state of Chapter 4.

5.2 Fine structure of networks and clone size distributions

Some of the main actors of the immune system - B-cells and T-cells - are divided into three different pools: naive, effector and memory cells. The naive cells, being those that have not yet encountered a pathogen they bind to, rely on large diversity to ensure an immune response to any new pathogen invading the organism. The binding ability is encoded in a receptor located on the surface of the cell and adapted to a few pathogens. This receptor consists of an alpha and a beta chain almost uniquely built during lymphocyte production (in the thymus or in the bone marrow) and transmitted through divisions. The aim of this analysis is to explain and predict clone size distributions as a result of assumptions made about interclonal and intraclonal competition. I will show that those two types of competition have very different effects on the clones in the populations.

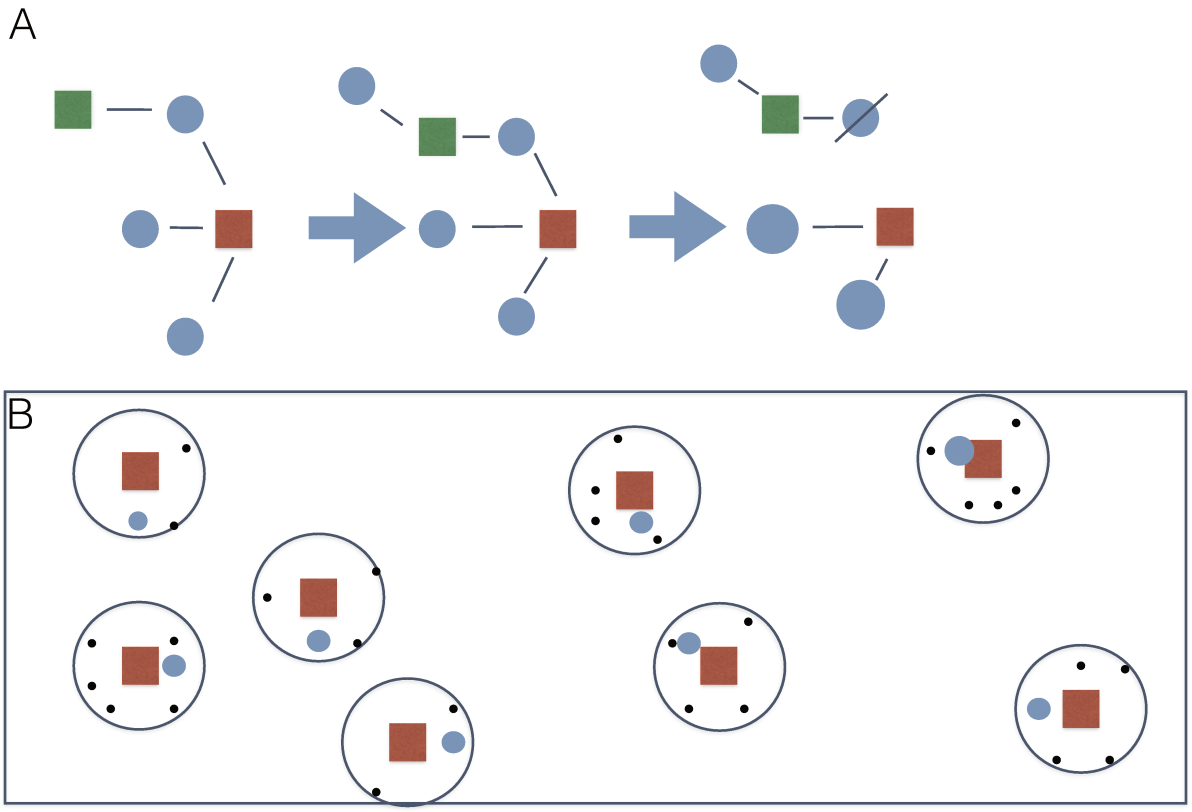

Figure 5.3: A. Evolution of a niche network: random addition of element to the graph destroys local niche equilibrium. Squares represent different antigens, circles represent clones and black lines represent high binding affinity between clones and antigens. On the left the dynamical graph is assumed to be locally stable. The introduction of a new clone binding to the green resource increases the pressure on the middle clone leading to its extinction. The two remaining clones that can bind to the red antigen grow. B: An example of loser and winner clones on a two dimensional shape space with a rotation invariant interaction kernel. Squares represent antigens, small dots are decaying clones of the losers pool and large blue circles are winners of the local niche. The black circles represent the typical decay distance of binding affinities.

How does this competition occur? In order to remain alive, naive cells require short binding events with peptides presented by the self so the organism can be rid of the least active clones. Here again, binding affinity depends on the receptor that usually responds to a few peptides. This necessary encounter with the self - called tickling - is a source of competition between clones and within clones for the limited resources that are these self-peptides. The intricate network of interacting cells and peptide reservoirs generates complex, non-linear dynamics for the clone populations while the constant introduction of new clones due to thymic or bone marrow output keeps the system out of equilibrium. A perturbation around mean-field solutions is explored here, trying to answer some of the following questions. What are the typical sizes and lifetimes of a clone? What is the essential competition mechanism? What is the typical distribution of clone sizes? In what range of parameters are these results valid and what should we expect for different values of these parameters?

The model is quite faithfully adapted from [67] though used in a different context, under different assumptions and to a broader applicability. This model differs from the ones in Chapter 4 because they do not directly average out the competition to form independent sets of equations. The competition is modeled explicitly and the mean-field independent equations are derived as a limit of the model. This means that the fluctuations in fitness can come from different sources. One of them is fluctuations of the resources as explored in Chapter 4. The possibility explored here is to assume the environment is fixed but the system is still out of equilibrium because new clones are introduced, disturbing the graph structure and creating fluctuations.

5.2.1 Modeling competition: antigens and lymphocytes

The naive lymphocyte population (typically $4 \cdot 10^{11}$ for T-cells) in adult humans is divided into p clones labeled with an index $i \in [1; p]$ with populations C_i. The antigens presented by the self are indexed by $j \in [1, q]$, their important variable is their availability to bind F_j which represents the fraction of antigens that are not currently bound to a lymphocyte. These antigens do not change over time since this model does not intend to represent invasions or a variable environment but the interaction with the self. The information on the network is the binding probability of each antigen with a lymphocyte. It is encoded in a $p \times q$ matrix K called the interaction matrix, where $K_{i,j}$ is the binding probability of the antigen j with lymphocyte i.

Writing down the partition function for binding equilibrium gives immediately the availability of antigens:

$$F_j = \frac{1}{1 + \sum_i K_{i,j} C_i}, \tag{5.13}$$

as described in 5.1.5. The lymphocytes are maintained alive by binding to self-proteins so we need to compute the probability of these encounters. Each lymphocyte is tickled by the encounters with the self proportionally to a tickling factor S_i that only depends on the clone it belongs to. The tickling factor S_i is given (again in a similar way as in 5.1.5) by

$$S_i = \sum_j K_{i,j} F_j. \tag{5.14}$$

5.2.2 Dynamics of the system

With these definitions at hand, it is possible do define the dynamics of the system. Let us consider a single clone i. The cells can divide with rate α (in divisions per day) and die with a

modified rate $\frac{\beta}{S_i}$ that is lower if the binding events are more numerous. The clone population dynamics are given by

$$\frac{dC_i}{dt} = \alpha C_i - \frac{\beta}{S_i} C_i. \tag{5.15}$$

S_i acts on the death rate because tickling is a survival signal rather than a proliferation signal. The form of Eq. 5.15 is a strong assumption. Different choices should be explored to prove that the results do not depend strongly on the specifics of Eq. 5.15.

Another important variable is N, the total size of the lymphocyte population. Here it is assumed that it is constant since its variations scale over years and are irrelevant in this short time-scale dynamics [38]. New clones are constantly produced in the bone marrow (B-cells) and the thymus (T-cells). We will assume that these new clones all present original receptors (the probability of seeing a new clone with an already existing receptor is extremely low) and that their population when entering the system is given by a known distribution $s(C)$ (the choices of distributions will be discussed later).

5.3 Clone size distributions in limits of the niche structure

Within an individual, clones compete and are selected, fitness being here the ability to bind to the self. In this small evolutionary system, a wide diversity is preserved and the fittest clone does not sweep over the whole population. The reason for this is essentially three-fold and quite similar to the reasons why sweeps do not happen in traditional clonal evolution (without the spatial aspect). First, due to birth and death stochasticity, the life span of a clone is too short, and the division rate too small (in the naive repertoire) to allow its population to grow up to a macroscopic fraction of the cell population. The second argument (which can also be seen as a time scale argument) is that the constant introduction of new clones brings too many competitors for one to completely sweep over the others. The last argument - the essential argument for the dynamics - is that since clones are specific to a few antigens, niches harbour different clones that are locally fit. It is from this point of view that our perturbative analysis will be carried out.

5.3.1 Degenerated cases: fully specific and nonspecific models

In this first section, we derive results for two extremely simplified models that, though defined differently turn out to be equivalent.

The first model is a mean-field approximation of the interactions and competition. Let us assume that all clones can bind with equal probability to any antigen. The lymphocytes are nonspecific and mathematically speaking, this assumption means that K is a full matrix with all entries equal to a unique value k.

Let n_A be the number of clones that we assume to be constant. Then the availability and tickling factor are uniform and respectively equal to

$$F_j = F = \frac{1}{1+kN} \quad ; \quad S_i = S = \frac{kn_A}{1+kN}. \tag{5.16}$$

The clones are all equality fit so, summing over all clones in Eq. 5.15, at steady state the average total cell population N^* is the solution of

$$0 = \partial_t N^* = \left[\alpha - \frac{\beta(1+kN^*)}{kn_A}\right] N^* + s, \tag{5.17}$$

where s is the average number of new cells entering the system per unit time. The solution of the equation is

$$N^* = \frac{\alpha n_A}{2\beta} - \frac{1}{2k} + \sqrt{\left(\frac{1}{2k} - \frac{\alpha n_A}{2\beta}\right)^2 + \frac{n_A s}{\beta}}. \tag{5.18}$$

The model is fully deterministic except for birth-death noise. If the birth-death noise is small because the dynamics are slow and all new clones enter at size C_0, then the system has a source and deterministic dynamics with an absorbing barrier in 0. We obtain the distribution of clone sizes from detailed balance:

$$\rho(C) = N\frac{1}{C\log C_0} \text{ for } C < C_0, \text{ and } 0 \text{ for } C > C_0. \tag{5.19}$$

Note that this distribution has a cutoff at C_0 and is not heavy tailed. Including the birth-death noise reproduces the Langevin equation of drift with birth-death of Appendix A.1 and its solution.

The real system, however, is not so a simple because specificity generates much more complex dynamics.

Let us now assume that clones are entirely specific and can only bind to one and only one antigen. Then each clone is equivalent to a smaller version of the preceding case and the formula given above for the total population holds for the population of each clone.

5.3.2 Perturbation, global effects and fitness change

In this section I derive the central equation of Chapter 4 from the competition equations (the equations involving S and F explicitly) using a mean-field approximation. We will see that this mean-field approximation does not include the specific effect of intraclonal competition and fails to predict correctly the clone size distribution.

Let us write the affinity as a perturbation of the mean-field case for all $(i,j) \in [1, n_A] \times [1, q]$

$$K_{i,j} = k + \delta k_{i,j}, \tag{5.20}$$

where $k_{i,j}$ is small in comparison to k. This assumption is not unreasonable: as the clone growth is exponential, even a small difference in fitness will create large variations at long times.

We must now assume that all clones are are a priori equally fit. The idea is that no receptor has any intrinsic advantage (the ones that are non functional have already been ruled ruled out in negative selection) but that fitness is due to the random number of competitors and the fluctuations of the graph topology. It means that $\langle \delta k_{i,j} \rangle = 0$ (where the average is taken over different realizations of the clone-antigen interaction random variable).

The stochasticity induced by new arrivals and birth-death processes is amplified by the non-linear dynamics of the system. I present below a direct calculation that shows that it predicts power-laws. However we will see that the approximation upon which the calculation relies breaks when intraclonal competition starts to matter.

Expanding F and S using Eq. 5.20, the tickling factor can be divided into two parts:

$$S = S^* + \delta S_i, \tag{5.21}$$

where

$$S^* = \frac{k n_A}{1+kN} + \frac{k}{(1+kN)^3}\sum_j (\sum_l \delta k_{l,j} C_l)^2 + \frac{1}{N(1+kN)^2}\sum_{j,l} \delta_{l,j}^2 C_l \tag{5.22}$$

and

$$\delta S_i = \sum_j \frac{\delta k_{i,j}}{1+kN} - \frac{1}{(1+kN)^2}\sum_{j,l}\delta k_{l,j}\delta k_{i,j}C_l - \frac{1}{N(1+kN)^2}\sum_{j,l}\delta^2_{l,j}C_l. \tag{5.23}$$

This equation is obtained by Taylor expanding the definition of S around its average S^*. S^* does not depend on i, it is a global effect and is treated as the systematic drift of the clone population. δS_i is due to specific interactions with a varying environment, its correlation time scale is of the order of the clones average lifespan. We model this interaction using a white noise proportional to a factor computed from the variance of S_i and the lifespan of the clones.

5.3.3 Fokker-Planck equation

The expansion of the tickling factor can lead to a dynamical equation that is formally equivalent to the simplified (delta-correlated fitness noise) version of clone-level noise dynamics in Chapter 4. The precision equation is Eq. A.24, given in Appendix A.6. This perturbative approximation does not give an accurate description of the distribution of clone sizes for large clones as we will see in 5.3.4. The clone size follows the dynamical stochastic equation

$$\partial_t C_i = \alpha C_i - \frac{\beta}{S^*}C_i + \frac{\beta C_i}{(S^*)^2}\sqrt{\Gamma}\xi, \tag{5.24}$$

where ξ is a Gaussian white noise and the convention for stochastic integral is Stratonovitch because the noise for each clone is extrinsic. Eq. 5.24 corresponds to Eq. A.24 with parameters $f_0 = \alpha - \beta/S^*$ and $\sigma = \sqrt{\Gamma}\beta/(\sqrt{2}(S^*)^2)$.

The derivation of the solution is given in Appendix A.6. Under the simple assumption that all new clones have the same population C_0 we get the power law:

$$\rho(C) = \frac{1}{a\sigma^2}\frac{\tau}{C^a}\left(1 - \frac{1}{C_0}\right) \qquad \text{for } C > C_0 \tag{5.25}$$

and

$$\rho(C) = \frac{1}{a\sigma^2}\frac{\tau}{C_0}\left(1 - \frac{1}{C^a}\right) \qquad \text{for } C < C_0, \tag{5.26}$$

where

$$a = \frac{f_0}{\sigma^2}. \tag{5.27}$$

This solution has the same form as the main result of Chapter 4. Its novelty is that it is formulated in terms of a precise description of competition. The parameters Γ and S^* do not appear in the description of Chapter 4.

Here τ is the number of new cells produced daily. In the mean-field picture, the small, exponentially decaying clones constitute most of the repertoire but large population explorations happen often enough to compensate the most frequent behaviour. We will see in the next section that this picture is not accurate: intraclonal competition was not accounted for in the calculation and it creates a spontaneous scale in the distribution, breaking the power-law.

5.3.4 Interclonal and intraclonal competition

The picture of Eq. 5.24 explains how scale free distributions emerge from fluctuations of the interaction graph of clones and antigen through interclonal competition but does not explicitly account for intraclonal competition.

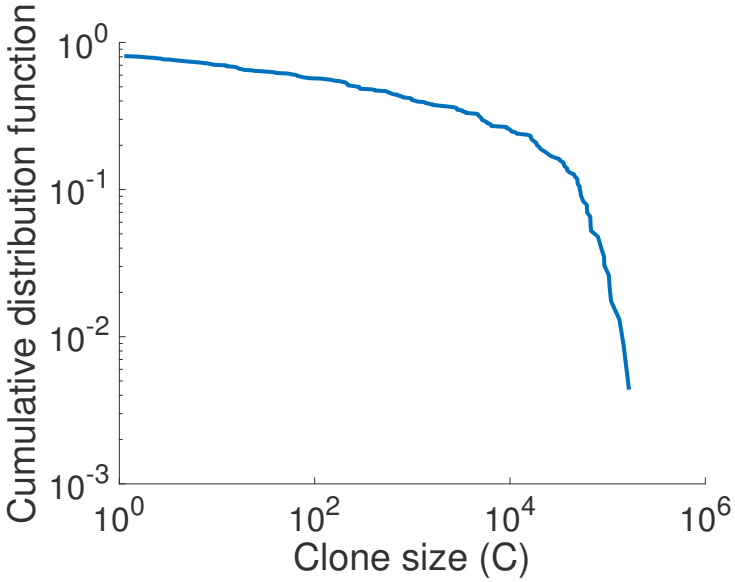

Figure 5.4: Distribution of clone sizes from the simulation of the competition model described in 5.2.1. The parameters of the simulations are $\alpha = 1$ day^{-1}, $\beta = 0.0001$ day^{-1}, $C_0 = 10$, $s = 20$ clones per day.

Even in the context of a scale free interclonal competition where clones can expand to very large populations thanks to exponential growth, intraclonal competition introduces a natural scale in the system. If each clone has a typical number C_c of competitors and a size C at which this competition makes its population stable, the model of the previous section assumes that all the competitors can randomly disappear, letting the clone grow indefinitely. Let us assume a situation where all competition for a resource is gone through random events in other niches. Only one clone is left alone in the niche, once this clone has grown large enough to reach a size of the same order as C_c, intraclonal competition is equivalent to the former interclonal competition and so the growth of the clone is stopped.

So the typical number of competitors creates a natural scale in the distribution and an exponential cutoff to the power law. This effect can be observed in simulations (see Fig. 5.4). The exponent of the power-law region of the distribution is very small since in explicit competition models, f_0 is always small because it is the result of the equilibration of birth and death through competition. In the distribution in Fig. 5.4 a power-law behaviour is observed for the first three decades of clone sizes with exponential cutoff due to the effect of intraclonal competition. Only a high number of competitors is consistent with large power-law distributed regions in experimental data.

In that context the picture of an infinite dimensional shape space for the matrix $K_{i,j}$ is not very reasonable and the matrix needs to have a specific topology: if two clones share one resource, they are more likely to share other resources than two random clones. For a random event to wipe out such a high number of competitors requires that their population evolutions are very correlated. This can only happen in a low dimensional shape space as the number of directions in the graph the competitors can use for extra resources is reduced.

The next step for this model is to quantify the cutoff of the power law from intraclonal competition and compare it to experimental distributions of naive repertoires. This has not

been done as not reliable data is yet available.

Naive repertoires have recently been sequenced and their distributions of clones sizes seem to have an exponential decay for large clone sizes (from Benny Chain, private communication). The models above could explain those distributions as competition in a fixed environment.

Chapter 6

Development in *Drosophila* embryos

Parts of this section directly use material from Ferraro et al in [102].

6.1 Patterning in early embryos

Adult organisms have a complex structure with several symmetries along different axes. Yet this structure is formed from an initial egg that one would naively imagine structureless and isotropic. Going from the egg to the adult requires several genetic and epigenetic mechanisms to break the symmetry of the egg and implement the blueprint of development. The structure is built by phenotypic variation of cells from one location to another in the embryo and so it is essential that cells estimate their position to influence their cell fates.

The same is true of *Drosophila melanogaster*. In the hours following egg laying cell divisions in *Drosophila* embryos are synchronous (they actually spread in a fast wave across the embryo) so that the early life of the embryo can be divided into numbered nuclear cycles corresponding to the time between mitoses. Divisions first happen in the center of the embryo. At nuclear cycle 6, nuclei start their migration from the center to the periphery of the embryo and spread in a single layer at the surface of the embryo to give rise after about one hour of development to the syncytial blastoderm (see Fig. 6.1A). During each nuclear cycle the number of nuclei is defined very precisely. Each nucleus needs to know where it is located in the embryo to determine its fate. Positional information is encoded and read in concentrations of different proteins called morphogens that activate or repress a set of initially activated genes called gap genes. Together, these concentration gradients form a map of the embryo that nuclei can read [103]. The mechanism that processes concentration of proteins to alter the fate of the nuclei is a series of gene regulatory networks. The accuracy and reproducibility of these processes is paramount to ensuring the build of a healthy organism as they bridge the gap between local processes such as migration or differentiation with global properties (e.g. shape, patterns).

Information theory is a natural framework to describe these phenomena. Positional information can be defined as the mutual information between the value of the cue the nucleus uses to determine its fate and the position in the embryo ([104, 105]) where mutual information I between two random variables X and Y is defined as

$$I(X,Y) = \int\int p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{6.1}$$

It quantifies the amount of information the morphogen gradient can provide about position.

In this framework it is possible to evaluate the performance of a readout of the morphogen gradient and ask what class of patterning is optimal from the information point of view ([106]). It is also possible to compute the mutual information between the input and the output of gene regulatory networks and evaluate their influence on the efficiency for gradient readout. Much work has been done in that direction in particular to define and find optimal networks with low complexity ([107, 108]).

The question of information is even more complex in the case of embryo development as each mitosis may erase most of the information stored during the nuclear cycle limiting the time available for the nucleus to accurately read its position.

In this work I do not focus on information theory as the goal is really to infer the dynamics of the gene networks from experimental data rather than study their efficiency from a theoretical point of view. I will to a certain extent discuss the implications of these results on positional information in the nucleus.

6.2 Development in fly embryos: Bicoid and *hunchback*

The *Drosophila* embryo has been used for decades as an excellent model to understand how cell identity is determined and maintained during development. The key regulatory networks in *Drosophila* include (among others):

1. two morphogenic transcription-factors: Bicoid and Dorsal and all the networks they activate downstream that are essential, respectively for Antero-Posterior (AP) [109] and Dorso-Ventral (DV) [110, 111] patterning,

2. the transcription cascade responsible for the formation of muscles (myogenic program)

3. the chain of reactions responsible for neural patterning.

Very early in embryonic life the mother deposits *bicoid* mRNA at one edge of the embryo. As the translated proteins diffuse through the embryo they create a gradient. Bicoid activates the transcription of the *hunchback* gene, a gap gene that codes for the development of the antero-posterior axis. The region where *hunchback* is strongly expressed will form the anterior part of the *Drosophila* body and the region where it is not expressed the posterior (see Fig 6.1 B). In the middle region of the antero-posterior axis only a fraction of nuclei express the *hunchback* gene. The slope of *hunchback* expression sets a border at the center (the exact definition of this border is ambiguous and will be discussed). It is the accuracy of the boundary location and its sharpness that define the efficiency of the readout.

The accuracy of the antero-posterior boundary formation process suffers from two limitations: one is the precision with which the regulatory network can read Bicoid concentration (it will be discussed in Chapter 7), the other are the fluctuations of the gradient of Bicoid concentration itself. The following section reviews results on the accuracy of diffusion limited processes.

6.3 The Berg and Purcell limit

In the setup of the *hunchback* readout of the Bicoid gradient, the *hunchback* gene sits in the nucleus surrounded by diffusing *bicoid* molecules at a concentration $\bar{c}$. Bicoid can bind to the

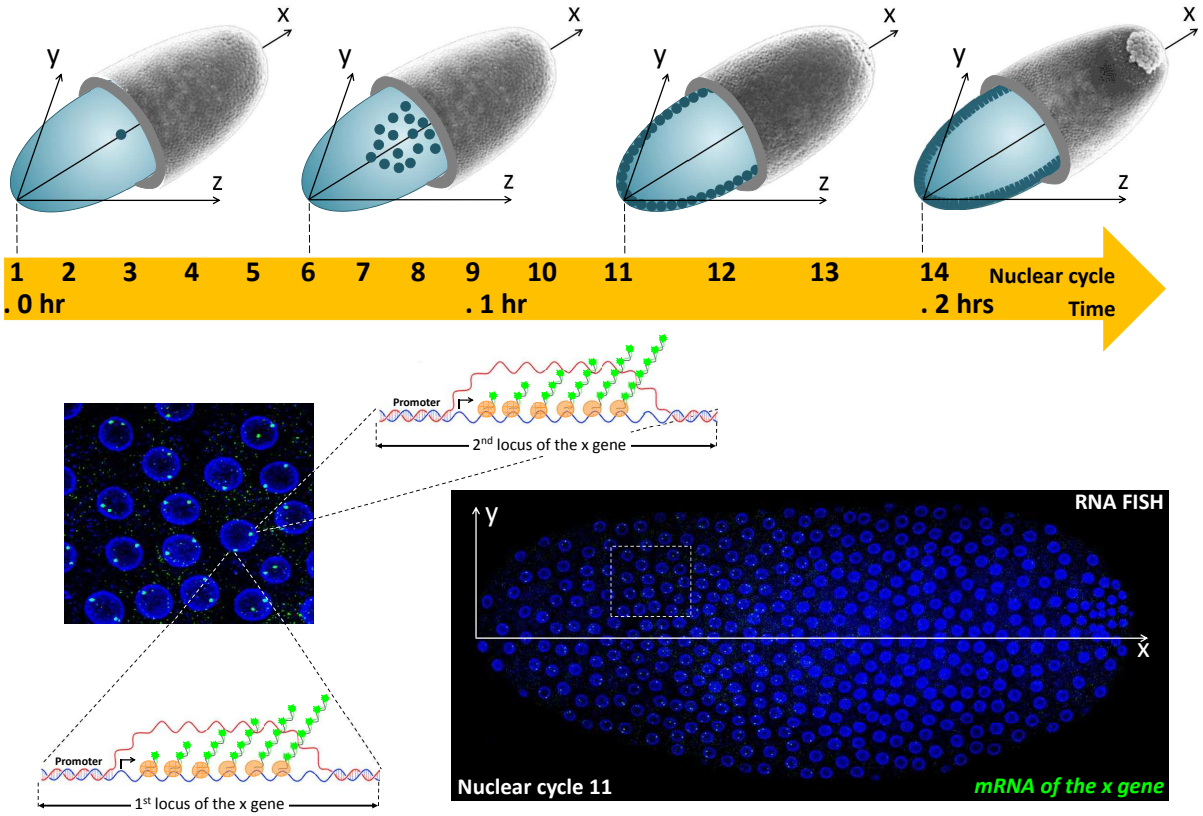

Figure 6.1: Top: Cell division and migration in the first two hours after egg laying. Bottom: On the right, nuclei (blue, nuclear envelope labeled with WGA-AlexaFluor-63315) are visualized at the surface of the whole embryo at nc11 and on the left, a close up of expressing nuclei (taken from the dashed square on the right). Expression of a given gene of interest (here hunchback) can be detected by RNA FISH with fluorescently labeled anti-sense RNA probes. Expression is revealed by two type of staining: speckle-like dots (arrow heads) corresponding to single mRNA and bright intense foci (arrow) corresponding to the accumulation of several nascent pre-mRNAs at their site of synthesis, as schematically diagrammed for the two hunchback loci.

hunchback site triggering the expression of the gene and the production of *hunchback* mRNA. In the limit of fast polymerase recruiting the total mRNA produced can be considered proportional to the time the promoter is occupied providing the nucleus with a way of integrating the binding signal it received. What is the variability of the estimate the nucleus can make of Bicoid concentration? More precisely can we compute $\delta c/c$ where c is the empirical concentration measured by the nucleus over a given time T?

The classic argument of Berg and Purcell in [11] gives an answer to that question in the case of a single binding site. Experiments and analysis have shown that between 5 and 7 Bicoid molecules can bind to the *hunchback* gene but the level of expression of the gene as a function of the number of bound Bicoid molecules is not known precisely [112, 113]. The Berg and Purcell limit provides us with an intuitive understanding of the different parameters on accuracy of concentration readouts as well as an upper bound on the sharpness of the boundary.

The Berg and Purcell limiting formula assumes that the time scales of binding and unbinding (the inverse of the jump rates of the Markov chain) are small compared the total time of integration T the nucleus has to read the concentration. It is equivalent to saying that a very high number of binding and unbinding events happen during the time T.

The movement of Bicoid in the nucleus is modeled by a simple diffusion equation

$$\partial_t c = D\nabla^2 c, \tag{6.2}$$

where c is the local concentration of Bicoid and D is the diffusion constant. Under those assumptions the flux of ligands arriving at a binding site is $4D\sigma\bar{c}$ where σ, the binding cross section of receptor and ligand represents the size of the target. Defining n as the empirical occupancy of the gene over the time T and $\bar{n}$ as the true average of this occupancy we see that the probability that the gene is free at any time point is given by $1-\bar{n}$. So the average number of binding events is

$$4D\sigma\bar{c}(1-\bar{n})T. \tag{6.3}$$

Due to averaging we expect the relative variability $\delta c/c$ to be proportional to the inverse of the square root of the number of events. The constants can be computed (see [11] for details) to get

$$\frac{\delta c}{c} = \sqrt{\frac{2}{4D\sigma\bar{c}(1-\bar{n})T}}, \tag{6.4}$$

which is the celebrated Berg and Purcell limit.

Eq. 6.4 is a fundamental tool in understanding the accuracy of diffusion limited processes but it also has a few problems. It assumes that any ligand finding the target will bind which is not true for most real biological systems. In [114] Bialek and Setayeshgar derive a correction to the Berg and Purcell formula by including two additional effects: the three dimensional geometry of the target and the possibility for the ligand to fail at binding the receptor even in contact. The new formula they derive is

$$\frac{\delta c}{c} = \sqrt{\frac{1}{\pi D\sigma c T} + \frac{2}{k_a\bar{c}(1-\bar{n})T}}, \tag{6.5}$$

where k_a is the association rate of ligand-receptor interactions (independently of space and diffusion). This new limit has the advantage of not converging to 0 when the diffusion constant goes to infinity. Unfortunately, as pointed out in [115] it does not agree with the Berg and

Purcell formula in the limit of deterministic binding ($k_a \to \infty$). In [115] Kaizu et al show that a new version of Eq. 6.5 where the first term is replaced with the Berg and Purcell limit can be derived analytically:

$$\frac{\delta c}{c} = \sqrt{\frac{1}{2\pi D\sigma\bar{c}(1-\bar{n})T} + \frac{2}{k_a\bar{c}(1-\bar{n})T}}. \tag{6.6}$$

In Chapter 7 I introduce a new approach to the problem of Bicoid readout accuracy in fly embryos based on effective models of gene switching that are rather agnostic about the details of ligand binding and diffusion. In particular, this approach does not require any assumption about the geometry of the target or the role of diffusion and association. This new approach does not have the constructive advantage of the ones presented above to directly link the result to the microscopic constants of the system. However its main asset is that it is formulated in terms of parameters that can be extracted from available data (as is done also in Chapter 7). I check the consistency of the precision prediction with the experiments.

6.4 Experimental methods

In this section I discuss the different experimental methods to access information about *hunchback* activity. They rely on the expression of fluorescent molecules that accumulate around mRNA produced at active loci. For a complete review of experimental methods to image transcription in living fly embryos see [102].

6.4.1 RNA FISH

Fluorescence in situ hybridization (FISH) is an experimental technique using fluorescent probes to bind specific DNA or RNA targets. In the context of development, it requires the embryo to be "fixed", meaning that its dynamics are stopped and the embryo is killed. The information is collected by fluorescent microscopy.

Until the last three years, gene expression in *Drosophila* embryo was mainly monitored on fixed samples by in situ hybridization or antibody staining. These techniques provided an exhaustive description of the precision and variability of spatial gene expression and helped understand its effect on patterning [116, 117, 118, 119, 120]. FISH has proven a very useful tool for defining the shape of average gene activation along the antero-posterior axis. Apart from problems due to fixation time that can lead to accumulation of non-instantaneous signal in certain areas FISH has almost no experimental problems and can detect single mature mRNA molecules in the embryo both in the cytoplasm and in the nucleus.

However, the development of the embryo happens very fast and the static picture of FISH is not very informative about the temporal dynamics of gene transcription and even less about the variability of gene expression in time. This limitation called for new methods providing live imaging of the embryo through time to study and quantify the transcription processes.

6.4.2 Live fluorescent Imaging

Benefiting from the pioneering work of R. Singer [121] several systems have been developed to fluorescently tag RNA in living cells. The first experiments using such techniques in fly embryos were carried out in 2013 ([18, 19]). In these experiments (and in those presented in Chapter 7)

a fluorescently tagged coat protein (CP) is expressed by the nucleus (MCP in Chapter 7). It binds strongly to a stem loop produced by a transgene inserted next to the *hunchback* gene location (MS2 in Chapter 7). Any time a polymerase transcribes the gene it also transcribes the transgene, producing loops that accumulate fluorescence at the transcription locus. This new method has proven extremely fruitful, giving insights about many processes including the dynamics or RNA synthesis: its initiation, elongation rate, the possibility of splicing or the time spent by the RNA at the locus after the end of transcription.

There are limitations to these approaches. First of all the constant production of fluorescently tagged coat proteins creates a background of fluorescence in the embryo. This background increases the lower bound on the amount of RNA production necessary to distinguish signal from noise. It also makes absolute counts more difficult to estimate. The background effect can be reduced by monitoring the concentration of coat proteins, keeping it as low as possible while high enough to ensure their presence in excess in the loop binding reaction. The last downside of live fluorescent imaging is that the transgene presence can alter the expression of the gene itself for many reasons. This is much harder to control but the insertion of only one transgene copy in the genome strongly reduced the risks of altering gene expression dynamics.

6.4.3 The importance of the construct

As represented in Fig 6.1 B as the transgene is transcribed fluorescence accumulates at the locus. The duration and the shape of this accumulation depend on the location of the probe along the gene. The probe should reasonably only be introduced in non coding regions and should be close enough to the gene to ensure good correlation between probe and gene transcription. This essentially only leaves two possibilities: putting the probe at the $5'$ or at the $3'$ end of the gene. The $5'$ is located where transcription of the gene begins and the $3'$ end is where the termination of transcription happens. The $5'$ options gives a bigger signal and longer accumulation of fluorescence. It increases the signal to noise ratio. Unfortunately it also smoothes out the signature of fluctuation in gene expression (see Fig. 6.2). On the other hand putting the probe at the $3'$ end reduces the length and accumulation of the signal. It makes it more sensitive to the background but it also makes it more sensitive to fast on and off switching that would have been invisible with a $5'$ build. The buffering time t_{buff} is the time spent accumulating signal by a single mRNA transcribing the gene. It is much smaller for $3'$ ($52s$) than for $5'$ ($168s$). The buffering time sets a lower bound on the time scale of dynamics available for direct analysis.

6.4.4 On the dynamics of *hunchback* activation

As mentioned above, gene expression relies on transcription-factor binding and is thus a noisy process contributing to the variability of protein profiles between homogeneous nuclei [122]. Two nuclei with the same genetic material put in the same environment can still generate different levels of RNA. These fluctuations cannot be traced back to any known cause or parameter, making them de facto intrinsically stochastic. Variability in total RNA production over a nuclear cycle can have different sources in temporal profiles: two nuclei producing mRNA at two different time-constant levels can differ just as much as two nuclei expressing levels of mRNA fluctuating with time but with the same mean (see Fig. 6.3 for illustration). Those two scenarios are clearly different from a biological point of view but would lead to similar results in FISH experiments. Live imaging is the only way to explore the temporal variability of signals. To formulate a clear

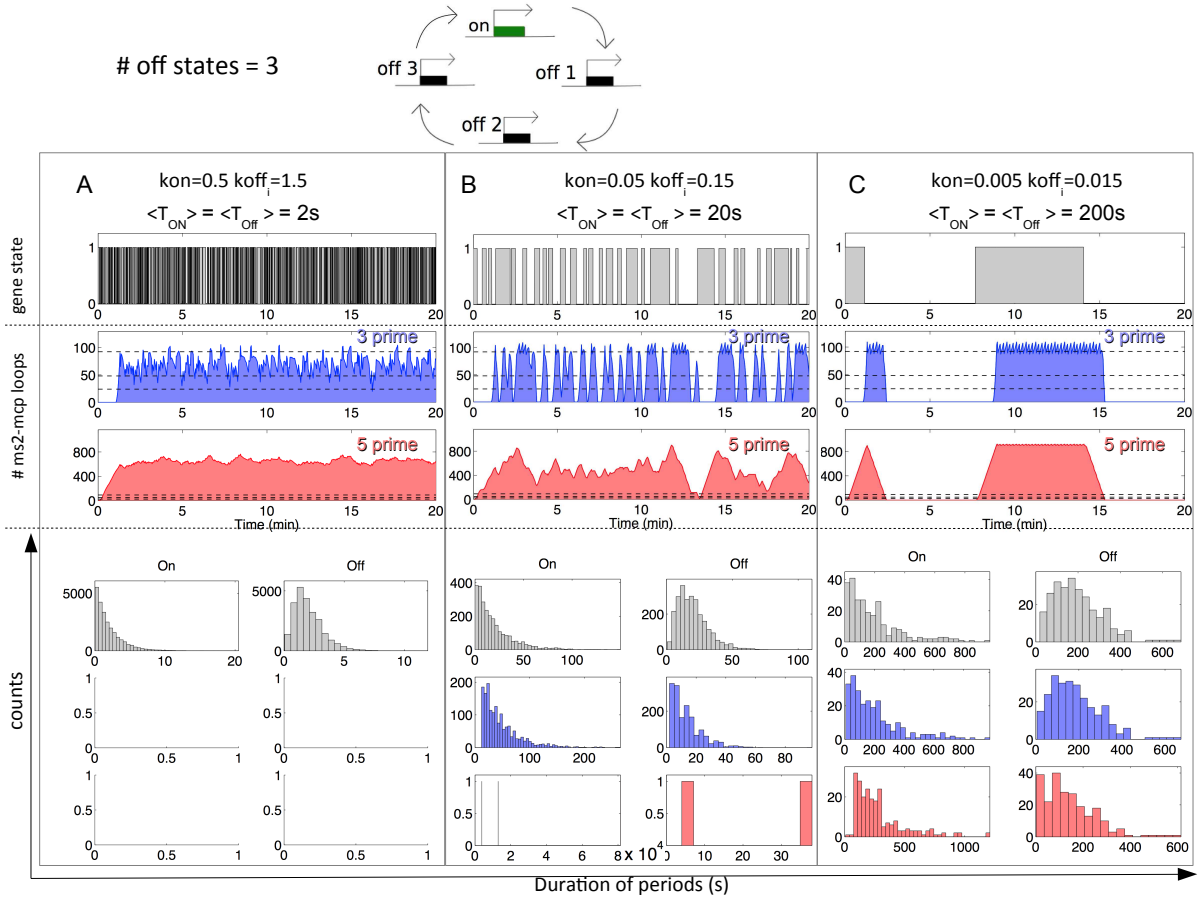

Figure 6.2: A lower sensitivity to noise background with 5' insertions but a better correlation between promoter activity and signal readout with 3' insertions. Simulations were performed assuming an irreversible promoter cycle (with three inactive states) model for transcription activation and deactivation at different frequencies of ON/OFF transitions (high (A), intermediate (B), and low (C)). k_{off} is the parameter of the exponential distribution of waiting times before jumping from the ON state to the first OFF state. k_{on} is the same parameter for each of the jumps from OFF1 to OFF2, OFF2 to OFF3 and OFF3 to ON. For each case, we show: the change in the state of promoter activity with the distributions of ON (expressing) and OFF (not expressing) waiting times as collected at the promoter (top panel), the simulation of the fluorescent signal detected when the loop tagging sequence is inserted in a 3' (middle panel) or 5' (bottom panel) position of the transcribed sequence. The three distributions at the bottom show the distribution of waiting times spent ON and OFF and how well a simple derivative analysis of the signal would do at estimating this distribution for both $3'$ and $5'$ cases. It is assumed that the concentration of PolIIs is not a limiting factor and that they can bind at every moment when the gene is in the ON state. Multiple PolII molecules can constitutively transcribe the gene during one ON event. The horizontal dashed lines in the middle and bottom panels indicate the background levels frequently encountered. The results do not qualitatively depend on the number of inactive states used, and the three models (the reversible two state telegraph, the irreversible three state telegraph model and the Gamma model) of Chapter 7 give qualitatively the same results.

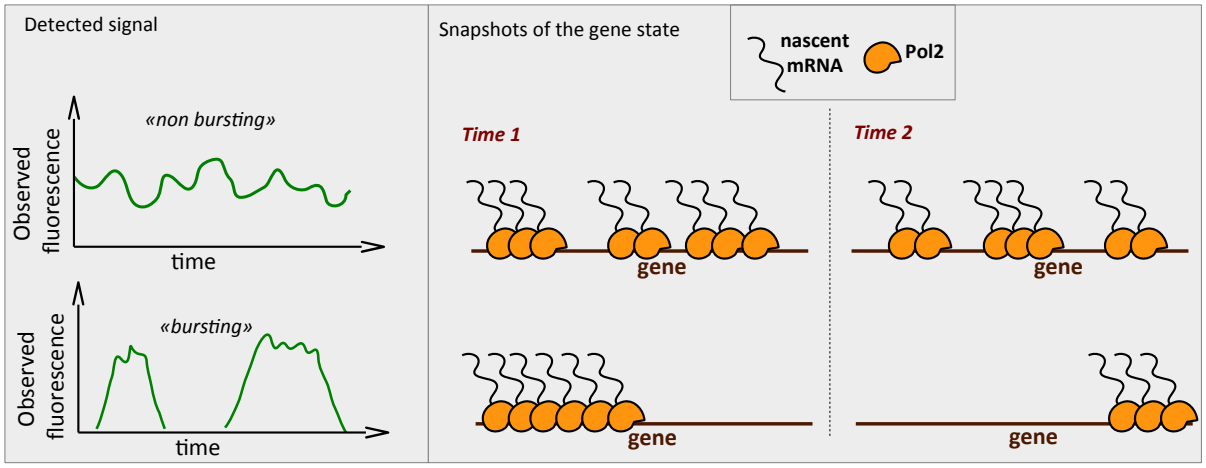

Figure 6.3: Bursting or nonbursting gene activity, the two types of transcription dynamics of a promoter. A continuum between the two extreme situations portrayed here can be found for different genes. (a) The amount of nascent RNA produced at the promoter fluctuates around a given positive value. The RNA PolII initiates transcription at an average constant rate. The promoter is active and nonbursting. (b) The activity of the promoter (measured as the amount of nascent RNA produced at a given time) alternates between periods of strong production (bursts) and periods of inactivity. The promoter is bursting. The characteristics of a bursting promoter include the frequency of the bursts, the intensity of the bursts, and their duration.

question one can ask if the dynamics of gene transcription are bursty (i.e the gene switches from an on to an off transcription state and back several times within one nuclear cycle) or non bursting (small fluctuations around a mean transcription rate). The evidence provided so far hinted at a non bursting situation but was based on 5′ probe experiments where fluctuations are much less visible. While [19] does not exclude bursting dynamics, the model used to represent data is based on a single pulse for each nuclear cycle. I will show that live imaging can go further in complexity and examine the dynamics of gene transcription.

In Chapter 7 I translate the different biological assumptions about the transcription dynamics into mathematical models of gene switching and evaluate their statistical consistency with experimental data. The result hints at bursting dynamics.

6.5 Motivation for autocorrelation method

The goal of the next chapter is to infer the statistics of gene activation from fluorescent live imaging experimental data and find the model that describes best the switching of the gene between expressing and non-expressing states. The gene is assumed to have two possible levels of expression: one where there is no expression (OFF state) and one where there is expression (ON state). Passing from the expressing state to the non-expressing state can take the gene through several changes of configuration and transcription factor binding so different models of gene switching are possible. In Fig. 6.3 I show an example of a model with one ON state and three OFF states.

The goal of the next chapter is to infer the structure of gene switching (number of states) and the parameters of the transition from state to state to quantify directly *hunchback* expression in the embryo as a function of position along the antero-posterior axis. Quantifying the dynamics of gene expression is an essential step in understanding how positional information is encoded in gene regulation networks.

The method developed in the next chapter makes use of the autocorrelation function of the fluorescent signal. Simpler methods do not give reliable results as evaluating on and off switching rates directly from the duration of transcriptional windows is biased by the integration time of the signal: short events are averaged out and blurred by the inertia of gene transcription. Using the derivative to identify increasing and decaying phases of the signal is also unreliable as it strongly enhances noise. Even in the absence of noise some fast OFF events could fail to lead to decrease in total fluorescence as loops would continue to accumulate.

The autocorrelation approach is quite intuitive and rigorous. It still faces a number of challenges, in particular the absence of reliable experimental calibration and the unavoidable short length of time traces. The next chapter shows how to overcome all these difficulties.

Chapter 7

Precision of readout at the hunchback gene

This chapter is submitted for publication and available on the arXiv and bioRxiv [23]

Jonathan Desponds[1,2,3], Huy Tran[1,2,3], Teresa Ferraro[1,2,3], Tanguy Lucas[2,3,4], Carmina Perez Romero[5], Aurelien Guillou[2,3,4], Cecile Fradin[5], Mathieu Coppey[2,3,4], Nathalie Dostatni[2,3,4] and Aleksandra M. Walczak[1,2,3]

[1] Ecole Normale Superieure, PSL Research University, Paris, France
[2] UPMC Univ Paris 06, Sorbonne Universités, Paris, France
[3] UMR3664/UMR168/UMR8549, CNRS, Paris, France
[4] Institut Curie, PSL Research University, Paris, France
[5] McMaster University, Canada

7.1 Abstract

The simultaneous expression of the *hunchback* gene in the multiple nuclei of the developing fly embryo gives us a unique opportunity to study how transcription is regulated in functional organisms. A recently developed MS2-MCP technique for imaging transcription in living *Drosophila* embryos allows us to quantify the dynamics of the developmental transcription process. The initial measurement of the morphogens by the *hunchback* promoter takes place during very short cell cycles, not only giving each nucleus little time for a precise readout, but also resulting in short time traces. Additionally, the relationship between the measured signal and the promoter state depends on the molecular design of the reporting probe. We develop an analysis approach based on tailor made autocorrelation functions that overcomes the short trace problems and quantifies the dynamics of transcription initiation. Based on life imaging data, we identify signatures of bursty transcription initiation from the *hunchback* promoter. We show that the precision of the expression of the *hunchback* gene to measure its position along the anterior-posterior axis is low both at the boundary and in the anterior even at cycle 13, suggesting additional post-translational averaging mechanisms to provide the precision observed in fixed

material.

7.2 Introduction

During development the different identities of cells are determined by sequentially expressing particular subsets of genes in different parts of the embryo. Proper development relies on the correct spatial-temporal assignment of cell types. In the fly embryo, the initial information about the position along the anterior-posterior (AP) axis is encoded in the exponentially decaying Bicoid gradient. The simultaneous expression of the Bicoid target gene *hunchback* in the multiple nuclei of the developing fly embryo gives us a unique opportunity to study how transcription is regulated and controlled in a functional organism [113, 123]. Despite many downstream rescue points where possible mistakes can be corrected [103, 113, 124], the initial mRNA readout of the maternal Bicoid gradient by the *hunchback* gene is remarkably accurate and reproducible between embryos [120, 125]: it is highly expressed in the anterior part of the embryo, quickly decreasing in the middle and not expressed in the boundary part. This precision is even more surprising given the very short duration of the cell cycles (6-15 minutes) during which the initial Bicoid readout takes place and the intrinsic molecular noise in transcription regulation [122, 126, 127].

Even though most of our understanding of transcription regulation in the fly embryo comes from studies of fixed samples, gene expression is a dynamic process. The process involves the assembly of the transcription machinery and depends on the concentrations of the maternal gradients [128]. Recent studies based on single-cell temporal measurements of a short lived luciferase reporter gene under the control of a number of promoters in mouse fibroblast cell cultures [129, 130] and experiments in *E. Coli* and yeast populations [131, 132, 133, 134] have quantitatively confirmed that mRNAs are produced in bursts, which result from periods of activation and inactivation. What are the dynamical properties of transcription initiation that allow for the concentration of the Bicoid gradient and other maternal factors to be measured in these short intervals between mitosis?

In order to quantitatively describe the events involved in transcription initiation, we need to have a signature of this process in the form of time dependent traces of RNA production. Recently, live imaging techniques have been developed to simultaneously track the RNA production in all nuclei throughout the developmental period from nuclear cycle 11 to cycle 14 [18, 19]. In these experiments, an MS2 cassette is placed directly under the control of an additional copy of a proximal *hunchback* promoter. As the gene is transcribed, mRNA loops are expressed that bind fluorescent MCP proteins. Their accumulation at the transcribed locus gives an intense localized signal above the background level of unbound MCP proteins (Fig. 7.1C) [102]. By monitoring the living embryo, we obtain a time dependent fluorescence trace that is indicative of the dynamics of transcription regulation at the *hunchback* promoter (Fig. 7.1B, D and F).

However the fluorescent time traces inevitably provide an indirect observation of the transcription dynamics. The signal is noisy, convoluting both experimental and intrinsic noise with the properties of the probe: the jitter in the signal is not necessary indicative of actual gene switching but could simply result from a momentarily decrease in the recording of the intensity. To obtain a sufficient strong intensity of the signal to overcome background fluorescence, a long probe with a large number of loops is needed, which introduces a minimum buffering time (in the current experiments the minimal buffering time is $\tau_{min}^{buff} = 72$s) and preventing direct

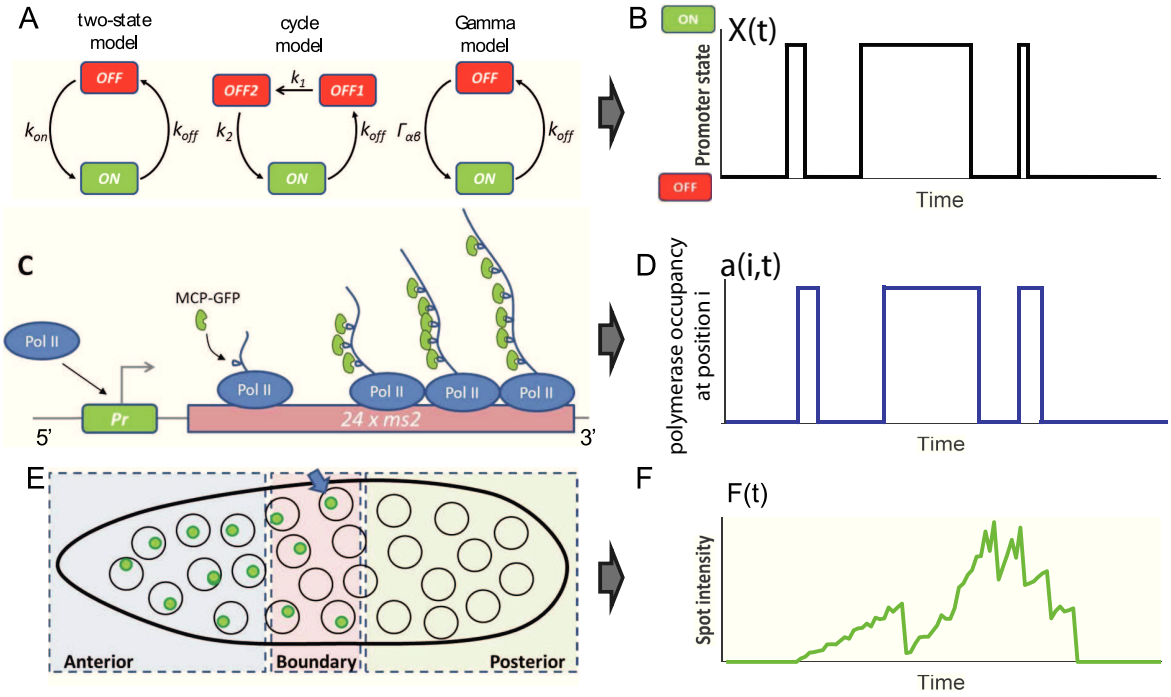

Figure 7.1: **Transcription dynamics in the fly embryo**. (A) The three models of transcription dynamics considered in this paper. From left to right: the two state model, the cycle model and the Gamma model (see SI Sections B, D and E). (B) Example of the promoter state dynamics (either ON or OFF) as a function of time. We assume that the polymerase is abundant and every time the promoter is ON and is not flanked by the previous polymerase a new polymerase will start transcribing. The black lines represent arrival times of the RNA polymerases to the promoter. (C) In the ON state, the promoter (Pr) is accessible to RNA polymerases (Pol II) that initiate the transcription of the target gene and the $24\times$ MS2 loops. As the $24\times$ MS2 mRNA is elongated MCP-GFP fluorescent molecules bind creating a detectable fluorescence signal. (D) The probability that site i on the gene was occupied by a polymerase as a function of time is given by the promoter occupancy in B and the finite size of the polymerase. (E) MCP-GFP molecules labeling several mRNA co-localize at the transcription loci, which appear as green spots under the confocal microscope. The spot intensities are then extracted over time and classified by each nuclei's position in the *Drosophila* embryo as Anterior, Boundary and Posterior. (F) An example of the experimental signal: one spot's intensity a function of time, corresponding to the arrivals of RNA polymerases in (D) and the promoter state in (B).

observation of activation [102].

To understand the details of the regulatory process that controls mRNA expression we need to quantify the statistics of the activation and inactivation times, as has been done in cell cultures [129, 130, 132, 133]. However the very short duration of the cell cycles (5-15 minutes for cell cycles 11-13) in early fly development prevents accumulation of statistics about the inactivation events and interpretation of these distributions. Direct observation of the traces suggests that, contrary to the previous reports[18, 19], transcription regulation is not static but displays bursts of activity and inactivity. However the eye can often be misleading when interpreting stochastic traces. In this paper we develop a statistical analysis of time dependent gene expression traces based on specially designed autocorrelation functions to investigate the dynamics of transcription regulation. This method overcomes the curse of naturally short traces caused by the limited duration of cell cycles that make it impossible to infer the properties of the regulation directly from sampling the activation and inactivation time statistics. Combining our analysis technique with models of transcription initiation and high resolution microscopy imaging of the MS2-MCP transgene under the control of the *hunchback* promoter, we show evidence suggesting that transcription initiation in cell cycles 12-13 is bursty. We focus on characterizing the transcription in the anterior and middle parts of the embryo and find that the dynamics is unchanged between cycle 12 and 13. We use these results to estimate the precision of the transcriptional readout. We show that the readout in each cell cycle is relatively imprecise compared to the precision of the mRNA measurement obtained on fixed samples [125].

7.3 Results

7.3.1 Characterizing the time traces

We study the transcriptional dynamics of *hunchback* by generating embryos that express an MS2-MCP reporter cassette under the control of the proximal *hunchback* promoter (Fig. 7.1C), using previously developed techniques [18, 19], with an improved MS2 reporter [135] (see Materials and Methods for details). The MS2-MCP cassette was placed towards the 3' end of the transcribed sequence and contained 24 MS2 loop motifs. While the gene is being transcribed, each newly synthesized MS2 loop binds a MCP-GFP molecule. In each nucleus, where transcription at this transgene is ongoing, we observe a unique bright fluorescent spot, which corresponds to the accumulation of several MS2-containing mRNAs at the locus (Fig. 7.1C). We assume that the fluorescent signal from a labelled mRNA disappears from the recording spot when the RNAP reaches the end of the transgene. With this setup we image the total signal in four fly embryos using confocal microscopy, simultaneously in all nuclei (Fig. 7.1E) from the beginning of cell cycle (cc) 11 to the end of cell cycle 13. We obtain a signal that corresponds to the temporal dependence of the fluorescence intensity of the transcriptional process in each nucleus, which we refer to as the time trace of each spot. Fig. 7.1F shows a cartoon representation of such a trace resulting from the polymerase activity (Fig. 7.1D) dictated by the promoter dynamics (Fig. 7.1B). We present examples of the traces analyzed in this paper in Fig. B.1 and the signal preprocessing steps in the Materials and Methods and SI Section A.

To characterize the dynamics of the *hunchback* promoter we need to describe its switching rates between ON states, when the gene is transcribed by the polymerase at an enhanced rate and the OFF states when the gene is effectively silent with only a small basal transcriptional

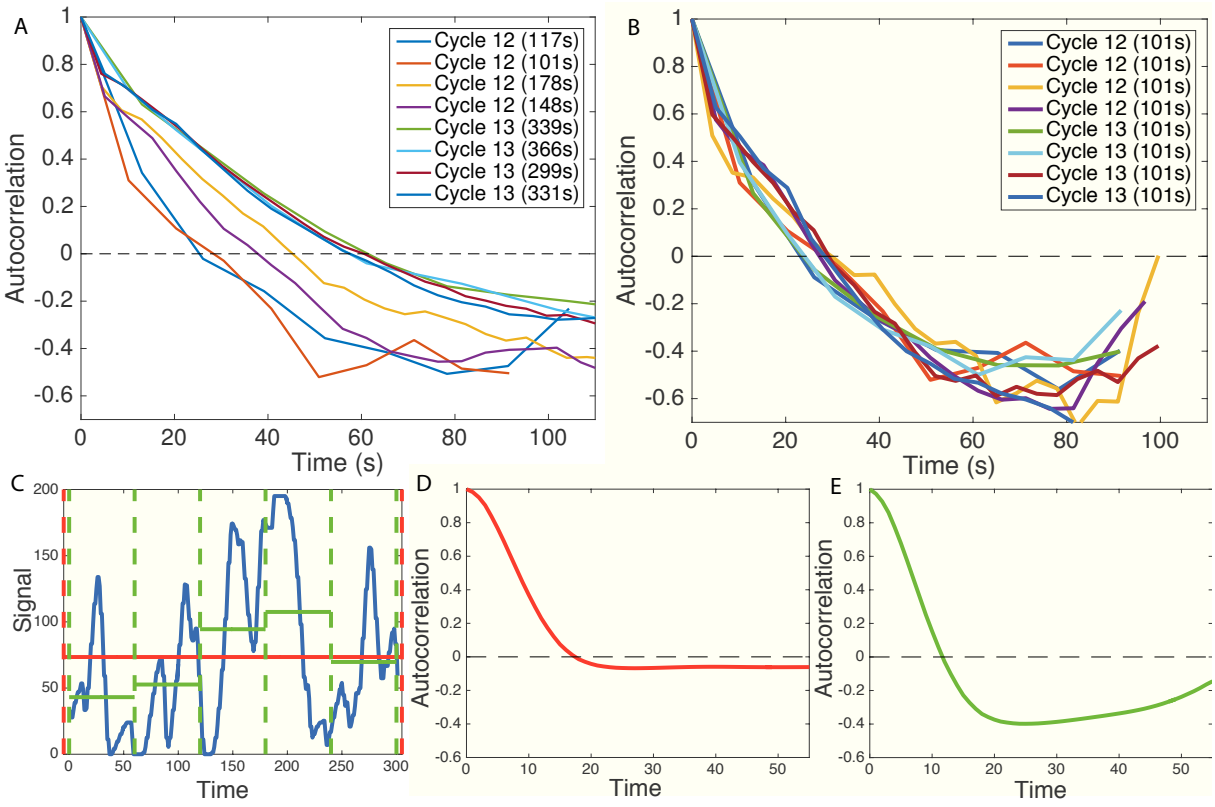

Figure 7.2: **Autocorrelation analysis of fluorescent traces from cell cycles 12-13.** (A) Autocorrelation functions for traces of different length caused by the variable duration of the cell cycle. Reading off the autocorrelation time as the time at which the autocorrelation function decays by a value of e would give different values for each trace. (B) Autocorrelation function calculated for the same traces reduced to have equal trace lengths, all equal to the trace length of the shortest trace, shows that the differences observed in panel A are due to finite size effects. (C) An example of a signal simulated for the process described in Fig.7.1 for a two state model for 300 seconds (blue). Taking the whole 300 second interval (red dashed) gives a good approximation of the average signal (red line) and the effect of finite size on the autocorrelation function is small (D). Reducing the time window to 60 seconds (green dashed line) correlates the average with the signal much more and the effect of finite size on the autocorrelation is strong (E). Parameters for the simulation in (C-E) are: $k_{\text{on}} = k_{\text{off}} = 0.06s^{-1}$, sampling time $dt = 4s$, for the red curve $T = 60s$ and $M = 2000$ nuclei, for the green curve $T = 300s$ and $M = 10000$ nuclei (same total amount of data).

activity (Fig. 7.1A and B). Estimating the ON and OFF rates directly from the traces is problematic due to the high background fluorescence levels coming from the unbound MCF-GFP proteins that make it difficult to distinguish real OFF events from noise. To overcome this problem, we consider the autocorrelation functions of the signal. To avoid biases from differential signal strengths from each nucleus, we first subtract the mean of the fluorescence in each nucleus, $F(t_i) - \langle F(t_i) \rangle$ and then calculate the steady state connected autocorrelation function of the fluorescence signal (equivalent to a normalized auto-covariance), $C(\tau)$, at two time points separated by a delay time τ, $F(t_i)$ and $F(t_i + \tau)$, normalized by the variance of the signal over the traces, according to Eqs. 7.11 and 7.12 in Materials and Methods. We will always work

with the *connected* autocorrelation function, which means the mean of the signal is subtracted from the trace. The autocorrelation function is a powerful approach since it averages out all temporally uncorrelated noise, such as camera shot noise or the instantaneous fluctuations of the fluorescent probe concentrations.

Fig. 7.2A compares the normalized connected autocorrelation functions calculated for the steady state expression in the anterior of the embryo (excluding the initial activation and final deactivation times after and before mitosis) in cell cycles 12 and 13 of varying durations: ~ 3 and ~ 6 minutes. The steady state signal from cell cycle 11 did not have enough time points to gather sufficient statistics. The functions decay as expected, showing a characteristic correlation time, then reaching a plateau at negative values before increasing again. Since the number of data points separated by large intervals is small the uncertainty increases with τ. Autocorrelation functions calculated for very long time traces have neither the negative plateau nor the increase at large τ. For example, the long-time connected autocorrelation functions shown in Fig. 7.2D calculated from the simulated trace of the process described in Fig. 7.1 and shown in Fig. 7.2C differ from the short time connected autocorrelation function in Fig. 7.2E calculated from the same trace (see SI Section G for a description of the simulations). As the traces get longer the connected autocorrelation function approaches the longtime results (Fig. B.4) and the connected autocorrelation function of a finite duration trace of a simple correlated brownian motion (an Ornstein-Uhlenbeck process) displays the same properties (see Fig. B.5). The dip is thus an artifact of the finite size of the trace. We also see that the autocorrelation functions shift to the left for short cell cycles (Fig. 7.2A), resulting in shorter correlation times, defined as the value of τ at which the autocorrelation function decays by e, for earlier cell cycles. However, calculating the autocorrelation functions for time traces of equal lengths for all cell cycles (Fig. 7.2B) shows that the shift was also a bias of the finite trace lengths, and after taking it into account, the transcription process in all the cell cycles has the same dynamics (although we note that the dynamics from this truncated trace is not the true long time dynamics).

This preliminary analysis shows that to extract information about the dynamics of transcription initiation we will need to account for the finite time traces. Additionally, a direct readout of even effective rates from the correlation time is difficult, because the autocorrelation due to the underlying gene regulatory signal (Fig. 7.1B) is obscured by the autocorrelation due to the timescale for the elongation of the sequence to be transcribed after the MS2 cassette (Fig. 7.1D) – the gene buffering time. The observed time traces are a convolution of these inputs (Fig. 7.1F). The form of the autocorrelation function and our ability to distinguish signal from noise also depends on the precise positioning and length of the fluorescent gene [102]. The analysis is thus limited by the buffering time of the signal, given as the length of the transcribed genomic sequence that carries the fluorescing MS2 loops divided by the polymerase velocity, and is only possible if the autocorrelation time of the promoter is larger than the buffering time. A construct with the MS2 transgene placed at the 3' end of the gene (Fig. 7.4B) gives a reliable readout of the promoter activity even for fast switching between the two states but the weak signal is hard to distinguish from background fluorescence levels. Conversely, a 5' positioning of the transgene (Fig. 7.4A) is insensitive to background fluorescence but can only be used to infer very slow switching [102].

7.3.2 Promoter switching models

The promoter activity we are interested in inferring can in principle be described by models of varying complexity (see Fig. 7.1A). In the simplest case, the gene is consecutively yet noisily expressed following a Poisson distribution of punctual ON events – this has previously been called a static promoter (not represented in Fig. 7.1A). Although the promoter dynamics would be uncorrelated in this case, the gene buffering would still produce a finite correlation time (see SI Section F). Alternatively, the promoter could have two well defined expression states: an ON state during which the polymerase is transcribing at an enhanced level and OFF state when it transcribes at a basal level. This situation can be modeled by stochastic switching between the two states with rates k_{on} and k_{off} (left panel in Fig. 7.1A and Materials and Methods). However, as was previously observed in both eukaryotic and prokaryotic cell cultures [129, 130, 132, 133], once the gene is switched off the system may have to progress through a series of OFF states before the gene can be reactivated. Recently these kinds of cycle models have been discussed for the *hunchback* promoter [136]. The intermediate states can correspond to, for example, the assembly of the transcription initiation complex, opening of the chromatin or transcription factor presence. These kinds of situations can either be modeled by a promoter cycle (middle panel in Fig. 7.1A and Materials and Methods), with a number of consecutive OFF states, or by an effective two state model that accounts for the resulting non-exponential, but gamma function distribution of waiting times in the off state (right panel in Fig. 7.1A and Materials and Methods). We present our method for all of these models and consider all but the gamma function distributed switching time model to learn about the dynamics of *hunchback* promoter dynamics.

7.3.3 Autocorrelation approach

To infer the transcription dynamics from the data we built a mathematical model that calculates the autocorrelation functions that account for the experimental details of the probes, incorporating the MS2 loops at various positions along the gene and correcting for the finite length of the signal. The basic idea behind our approach is that while the initiation of transcription is stochastic and involves switching between the ON and possibly a number of OFF states ($X(t)$ in Fig.7.1B denotes the binary gene expression state), the obscuring of the signal by the probe design is completely deterministic [137, 19], resulting in the probability $a(i,t)$ that the polymerase is at position i at time t (Fig.7.1D). The promoter dynamics can thus be learned from the noisy autocorrelation function of the fluorescence intensity $F(t) = \sum_{i=1}^{r} L_i a(i,t)$ (Fig.7.1F), provided the parameters of the probe design encoded in the loop function L_i (positioning of the probe etc.) are known (Fig. 7.1C) and the signal is calibrated to know the fluorescence intensity coming from one loop [19].

Broadly, our model assumes that once the promoter is in an ON state the polymerase binds and deterministically travels along the gene producing MS2 loops containing mRNA that immediately bind MCP and result in a strong localized fluorescence (Fig. 7.3). We count the progression of the polymerase in discrete time steps, where one time step corresponds to the time of it takes the polymerase to cover a distance of 150 base pairs equal to its own length (Fig. 7.3A). The probability that there is a polymerase at position i at time t, $a(i,t)$ is simply a delayed readout of the promoter state at time $t-i$, $a(i,t) = X(t-i)$ where t is measured in polymerase time steps (Fig. 7.1B). We assume that polymerase is abundant and that at every

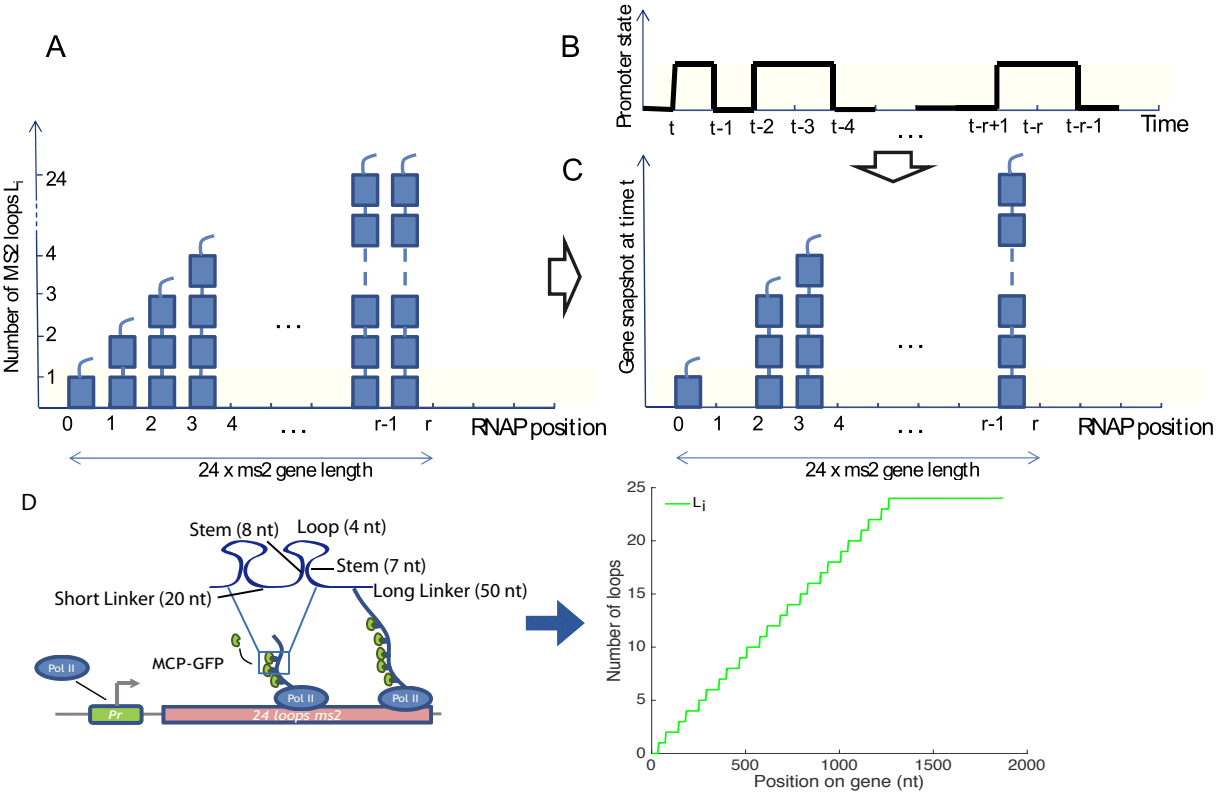

Figure 7.3: **The gene expression model used in the autocorrelation function calculation.** The autocorrelation inference approach is based on the idea that the stochastic transcriptional dynamics can be deconvoluted from the signal coming from the deterministic fluorescent construct, if we know the gene construct design. (A) A concatenation of snapshots of the gene from r consecutive time steps. A polymerase covers a length on the gene corresponding to its own length in one time step, producing one MS2 loop. The gene has total length r and at any position i along the gene $L_i < 24$ loops have been produced. (B) The promoter state as a function of time and (C) an instantaneous snapshot of the gene corresponding to transcription from this promoter. (D) The construct design is encoded in the loop function L_i. As the polymerase moves along the gene it produces MS2 loops. L_i is an average representation in terms of polymerase time steps of how many loops have been produced by a single polymerase. It is based on the experimental design shown on the left of the panel.

time step a new polymerase starts transcribing, provided the gene is in the ON state (Fig. 7.1B and D). The amount of fluorescence produced by the gene at one time point is determined by the number of polymerases on the gene (Fig. 7.3A). The amount of fluorescence from one polymerase that is at position i on the gene depends on the cumulated number of loops that the polymerase has produced L_i, where $1 \leq i \leq r$, r corresponds to the maximum number of polymerases that can transcribe the gene at a given time and $L_i = 1$ corresponds to one loop fluorescing, as depicted in the cartoon in Fig. 7.3D. The known loop function L_i depends on the build and the position of the MS2 cassette on the gene, it is input to the model and does not necessarily take an integer value since the polymerase length and the loop length do not coincide (Fig. 7.3D). Given the steady state probability of the gene to be on P_{on} the average fluorescence in the steady state is:

$$\langle F \rangle = P_{\text{on}} \sum_{i=1}^{r} L_i. \tag{7.1}$$

Since we assume the polymerase moves deterministically along the gene, seeing a fluorescence signal both at time t and position and i and at time s and position j means the gene was ON at time $t - i$ and $s - j$, which is determined by how many loops (i and j) the polymerase has produced. Taking the earlier of these times, we need to calculate the probability that the gene is also ON at the later time. The autocorrelation function of the fluorescence can thus be written as:

$$\begin{aligned} \langle F(t)F(s) \rangle = \; & \textstyle\sum_{i=1}^{r} \sum_{j=1}^{r} L_i L_j P\,(\text{gene was ON at time} \\ & \min(t-i, s-j)) \cdot A(|t-i-s+j|), \end{aligned} \tag{7.2}$$

where $A(n)$ is the probability that the gene is ON at time n given that it was ON at time 0. The precise form of P_{on}, $P(\text{gene was on at time}\, \min(t-i, s-j))$ and $A(|t-i-s+j|)$ depends on the type of the promoter switching model. We assume that the polymerase moves at constant speed along the gene and that there is no splicing throughout the transcription process. We give explicit expressions for all the models used in the Materials and Methods section and the Supplementary Information. Importantly, if we know the design of the construct, and calibrate the signal, we can use Eq. 7.1 to obtain the ratio of switching rates and Eq. 7.2 to obtain their particular values (see Materials and Methods).

To avoid biases coming from nucleus to nucleus variability, we calculated the normalized connected correlation function defined in Eqs. 7.11 and 7.12 in Materials and Methods. The theoretically calculated connected autocorrelation function, C_r (Eq. 7.13 which corresponds to the longtime correlation function in Fig. 7.2C and D) differs from the empirically calculated connected autocorrelation function from the traces, $c(r)$ (Eqs. 7.11 and 7.12 in Materials and Methods, which corresponds to the short time correlation function in Fig. 7.2C and E) due to finite size effects coming from spurious correlations between the empirical mean and the data points. Since by definition the mean of a connected autocorrelation function is zero (see Eqs. 7.11 and 7.12 in Materials and Methods), the area under the autocorrelation function must be zero. For short traces this produces the artificial dip discussed in Fig. 7.2, which for long traces is not visible as it is equally distributed over long times. To compare our theoretical and empirical correlation functions we explicitly calculate the finite size correction and include this correction in our analysis (Materials and Methods and SI Section H and I).

In this paper, we have analyzed data from fly embryos with 3' promoter constructs only, limiting ourselves to the steady state part of the trace. We limit our analysis to the steady

state part of the interphase by taking a window in the middle of the trace to avoid the initial activation and final deactivation of the gene between the cell cycles (see Materials and Methods). However the method can also be applied to non-steady state systems (see SI Section C) and other constructs, including cross-correlation functions calculated from signals of different colors inserted at different positions along the gene (see SI Section J), which we discuss using simulated data.

7.3.4 Simulated data

We first tested the autocorrelation based inference on simulated short-trace data with underlying molecular models with different levels of complexity for a construct with the MS2 probe in the 3' end of the gene (Fig. 7.4B). In Fig. 7.4D we compare autocorrelation functions for the three state model for constructs with the MS2 loops positioned at the beginning of the transcribed region (5', Fig. 7.4A) and at the end of the transcribed region (3', Fig. 7.4B), and the cross-correlation function calculated from a two-colored probe construct (Fig. 7.4E). The analytical model correctly calculates the short trace autocorrelation function approach and is able to infer the dynamics of promoter switching for all models. It can also be adapted to infer the promoter switching parameters for any intermediate MS2 construct position, given of the limitations of each of the constructs discussed above [102].

The autocorrelation function based inference reproduces the underlying parameters of the dynamics with great accuracy for switching timescales smaller than the gene buffering time that obscures the signal (Fig. 7.4F). In Fig. 7.4F we show the results of the inference for the 3' two state model for difference values of the ON and OFF rates, k_{on} and k_{off}. For switching rates faster than the gene buffering time, the autocorrelation function coming from the length of the construct dominates the signal and the precision of the inference goes down. For very fast switching rates ($> 0.12s^{-1}$), increasing the length of the traces or the number of nuclei (red vs blue curve above $k_{\text{on}} + k_{\text{off}} = 0.1s^{-1}$ in Fig. 7.4F) does not help estimate the properties of transcription. For intermediate switching rates ($0.07 - 0.12s^{-1}$), increasing the trace length or increasing the number of nuclei extends the inference range (black and green dashed lines vs blue solid line Fig. 7.4F) and in all cases increasing the number of nuclei decreases the uncertainty as can be seen from the smaller error bars (shown only for the red and blue lines for figure clarity).

Using two colored probes attached at different positions along the gene gives two measurements of transcription allows for an independent measurement of the speed of the polymerase - one of the parameters of the model that currently must be taken from other experiments. While the estimates of polymerase speed in the fly embryo are reliable [19], it has been pointed out as a confounding factor in other correlation analysis [138].

The autocorrelation approach also correctly infers the parameters of transcriptional processes when applied to traces that are out of steady state (see SI Section C). However, since the process is no longer translationally invariant more traces are needed to accumulate sufficient statistics. For this reason, in the current analysis of fly embryos we do not analyze the transient dynamics at the beginning and end of each cycle and we restrict ourselves to the middle of the interphase assuming steady state is reached.

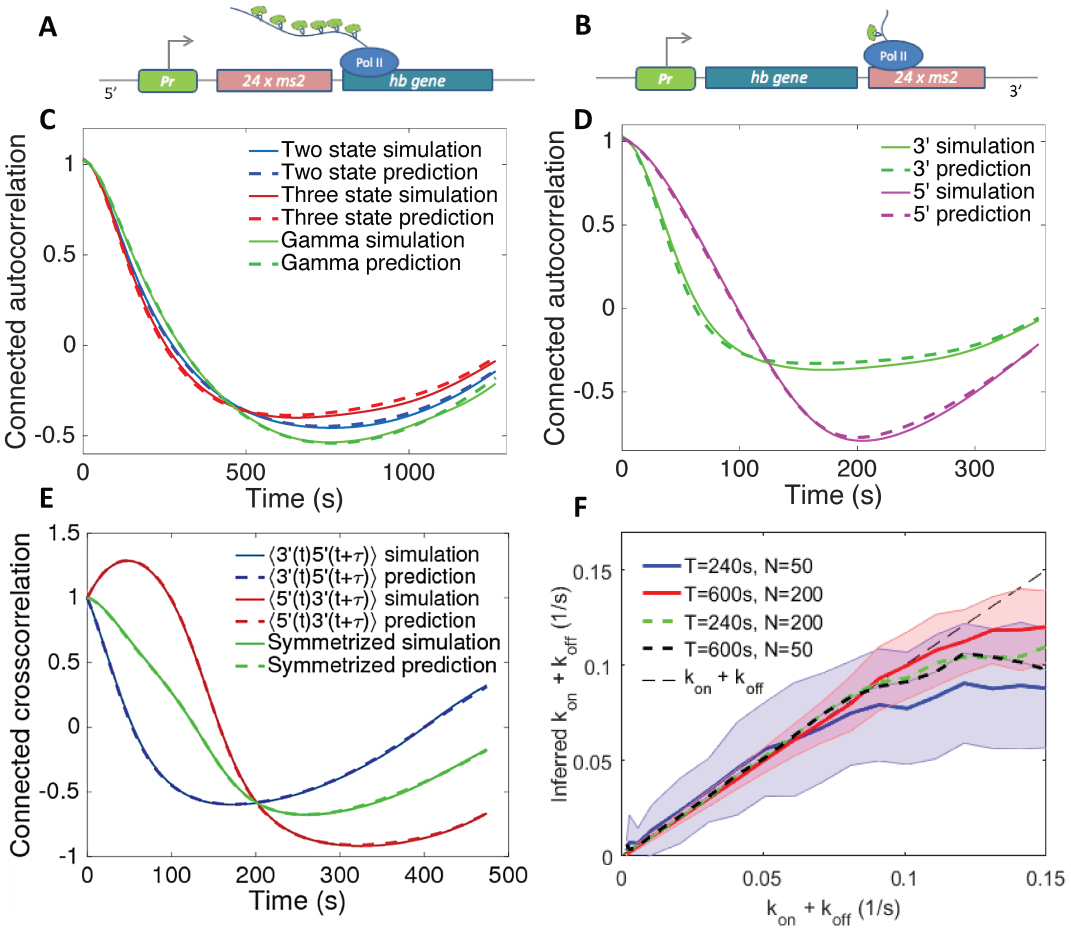

Figure 7.4: **The autocorrelation based inference analysis performed on short trace simulated data for models of various complexity and positioning of the MS2 probe.** Examples of the inferred autocorrelation functions fit to ones calculated from simulated traces (according to Gillespie simulations described in SI Section G) show perfect agreement for 3' MS2 insertions assuming a two state (telegraph) model, three state model and gamma function bursty model (A), as well as 3', 5' for the two state model (B). C. The cross-correlation between the signal coming from two different colored fluorescent probes positioned at the 3' and 5' ends. D. The inference procedure for the two state model correctly finds the parameters of transcription initiation in a wide parameter range. The inference range grows with trace length and the number of nuclei. Error bars shown only for $T = 240s$, $N = 50$ nuclei (blue line) and $T = 600s$, $N = 200$ nuclei (red line) for clarity of presentation. Parameters for the simulations and predictions are: (C) For two state $k_{\text{on}} = 0.005$ s^{-1}, $k_{\text{off}} = 0.01$ s^{-1}, sampling time $dt = 6$ s, $T = 360$ s and number of cells $M = 20000$, for three state same parameters with $k_{\text{off}} = 0.01$ s^{-1}, $k_1 = 0.01$ s^{-1} and $k_2 = 0.02$ s^{-1}, for Γ model same parameters with $k_{\text{off}} = 0.005$ s^{-1} and $\alpha = 2$ and $\beta = 0.01$ s^{-1}. (D) $k_{\text{on}} = 0.02$ s^{-1}, $k_{\text{off}} = 0.01$ s^{-1}, sampling time $dt = 6$ s, $T = 600$ s and number of cells $M = 20000$. (E). $k_{\text{on}} = 0.01$ s^{-1}, $k_{\text{off}} = 0.01$ s^{-1}, $dt = 6$ s, $T = 480$ s and $M = 20000$. The 5' construct is modeled as having 20 more fluorescent polymerase sites than the 3′ construct. F. $P_{\text{on}} = 0.1$

7.3.5 Fly trace data analysis

We divided the embryo into the anterior region, defined as the region between 0% and 35% of the egg length (the position at 50% of the egg length marks the embryo midpoint), where *hunchback* expression is high, and the boundary region, defined as the region between 45% and 55% egg length, where *hunchback* expression decreases. The mean probability for the gene to be ON during a given cell cycle P_{on} (restricted to the times excluding the initial activation and deactivation of the gene, which we will call the steady state regime), given by Eq. 7.1, is reproducible between the four embryos in cell cycle 12 and 13, both in the anterior region and at the boundary (Fig. 7.5A). The probability for the gene to be ON is over three fold higher in the anterior region than in the boundary and does not change with the cell cycle. $P_{\text{on}} \sim 0.5$ in the anterior indicates that in each nucleus the polymerase spends about half the steady state expression time transcribing the observed gene. At the boundary the gene is transcribed on average during about 10% of the steady state part of the cell cycle. The estimates for P_{on} in the earlier cell cycles were not reproducible between the four embryos, likely because the time traces were too short to gather sufficient statistics for this kind of analysis. We concentrated on cell cycle 12 and 13 for the remainder of the analysis.

Based on the different behavior at the boundary and in the anterior, we separately inferred the transcriptional dynamics parameters in the two regimes, using the autocorrelation approach that corrects for finite time traces. The Poisson random firing model, the two and three state cycle models all provide reasonably good fits to the all the traces in both regions (see Fig. 7.5B for an example and Fig. B.3 for the fits in both regions in all embryos). However, the fit of the Poisson random firing model (red line) only captures the short time behavior of the measured autocorrelation function. The two and three state model fits are indistinguishable and the two state fit is reproducible between cell cycles and embryos (Fig. 7.5B). The variability of the two-state inferred parameters is given in Fig. B.9. The three state fit is reproducible at the level of the sum of the effective ON and OFF rates (same fit as shown for the two state model in Fig. 7.5C), but gives fluctuating values for k_1/k_2, the parameter determining how well it is approximated by a two state model (see Fig. B.6, $k_1/k_2 < 1$ describes one fast reaction between the OFF states, effectively giving a two state model, while $k_1/k_2 = 1$ gives equal weights to the two reactions, clearly distinguishing two OFF states). Since the two state model is reproducible and has lesser complexity we will further consider the two state model.

The inference procedure independently fits the characteristic timescale of the process, defined as the inverse of the sum of two rates, $k_{\text{on}} + k_{\text{off}}$ (Fig. 7.5C), and then uses an independent fit of the probability of the gene to be ON, P_{on} (Fig. 7.5A), to disentangle the two rates (Fig. 7.5D). Examples of the promoter state over time with the rates' inferred values are shown in Fig. 7.5E (for the anterior region) and Fig. 7.5F (for the posterior region). Assuming the two state model we find that the characteristic timescale in most embryos is slighter shorter at the boundary ($\sim 25s$) than in the anterior region ($\sim 33s$) and the variability between the two cell cycles is comparable to the embryo to embryo variability (Fig. 7.5C). Both timescales are much larger than the $6s$ buffering time during which a second polymerase cannot bind because the first one has not cleared the binding site (shown as the gray dashed line in Fig. 7.5D), which sets a natural scale for the timescales we can infer. We find that in the anterior region of the embryo the two switching rates k_{on} and k_{off} show variability from embryo to embryo (between $0.009s^{-1}$ to $0.078s^{-1}$ – see Table I and II in the SI) but always scale together, which gives the

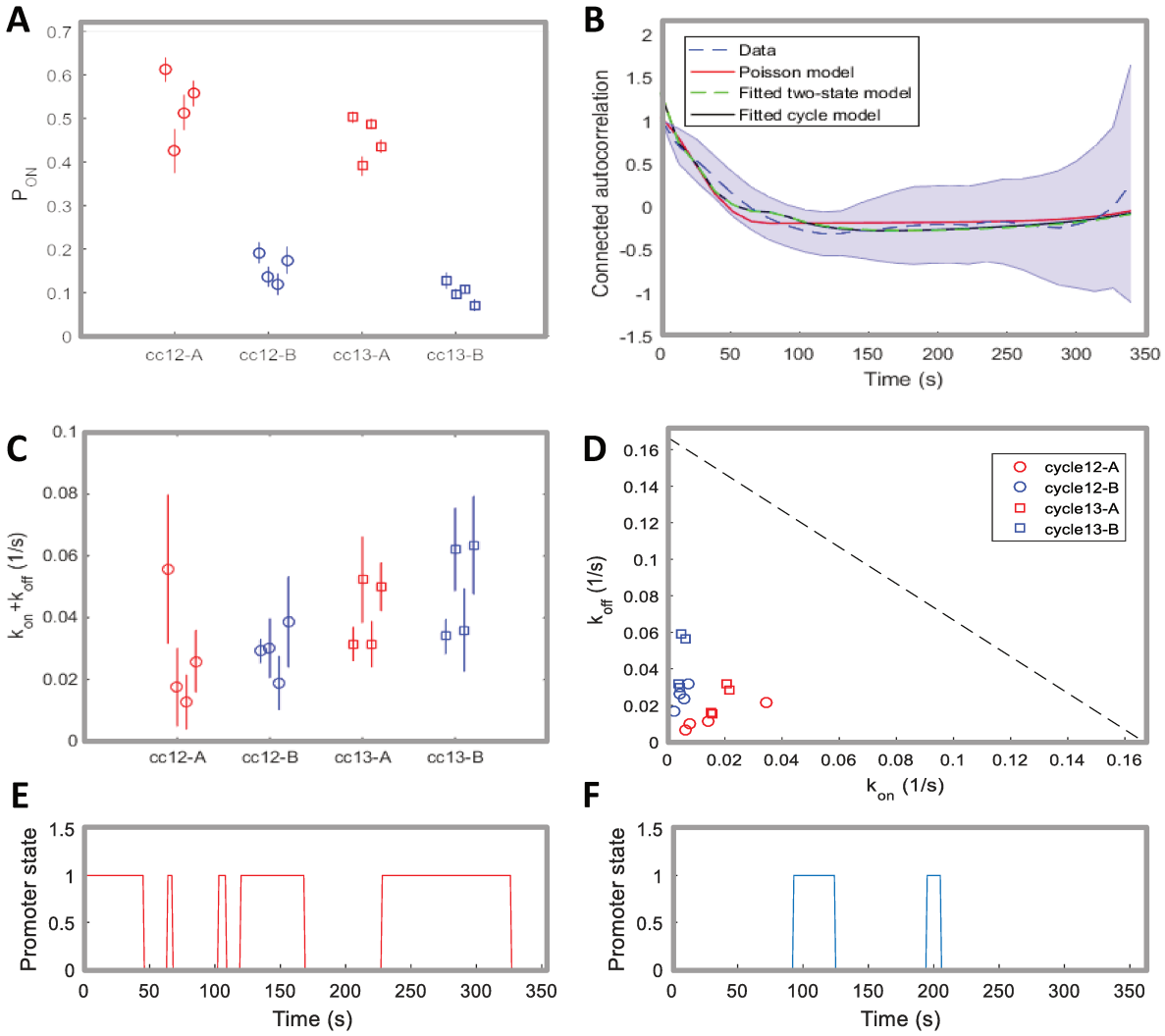

Figure 7.5: **Inference results for fly data**. (A) Inferred values of P_{ON} for different nuclei positions (A-Anterior, B-Bounary) and cell cycles. (B) Example of the mean connected auto-correlation function of the traces in cell cycle 13 (dashed blue line, with shaded error region) and of the fitted Poisson (red), two-state (green) and cycle (black) models. The fitted curves generated from the two-state and three state cycle model are almost superimposed. (C) Inferred values of $k_{\text{on}} + k_{\text{off}}$ using the two-state model. In (A) and (C), the standard error bars are calculated by performing the inference on 20 random subsets that take 60% of the original data. (D) Inferred values of k_{on} and k_{off} in the Anterior (red) and Boundary (blue), in cell cycle 12 (circle) and cell cycle 13 (square). For each condition, 4 inferred values for 4 movies are shown. (E-F) Two trajectories of the promoter state with the inferred parameters in the Anterior (red) and Boundary (blue).

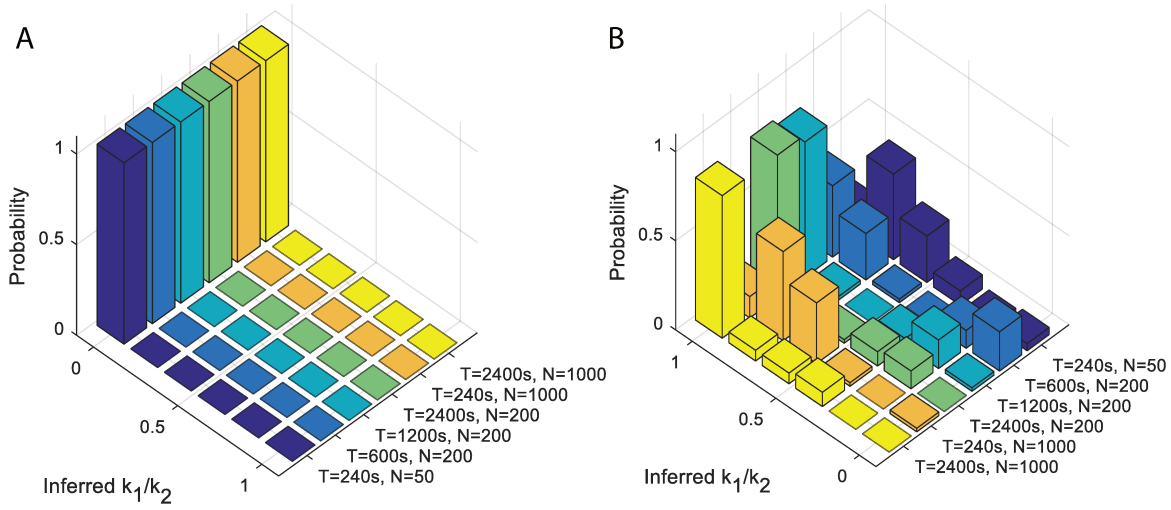

Figure 7.6: **Longer time traces help distinguish between two state and three state cycle models.** A. Inference on data generated by a two state model, which corresponds to $k_1/k_2 = 0$, from traces of different lengths T and using different numbers of nuclei N shows that longer traces help increase the probability to correctly learn the model type. Increasing the number of nuclei for short traces shows little improvement. The inference is repeated 50 times per condition. The experimental conditions studied in this paper are closest to the $T = 240s$ and $N = 50$ nuclei panel. B. The same numerical experiment but assuming a three state cycle model, which corresponds to $k_1/k_2 = 1$. Parameters of the simulations: $P_{\text{on}} = 0.1$, $k_{\text{off}} + 1/(1/k_1 + 1/k_2) = 0.02s^{-1}$ and $k_1/k_2 = 0$ in A, and $k_1/k_2 = 1$ in B.

observed one-half probability of the gene to be ON in a given nuclei during the steady state part of the interphase. Since the polymerase in the anterior on average spends half the steady state interphase window transcribing the gene, this suggests a clear bursting behavior of the transcription process, with switching between an identifiable active and inactive state of the promoter.

k_{on} is much smaller at the boundary with very little embryo to embryo variability, while k_{off} has a similar range as in the anterior. This behavior is expected since high Bicoid concentrations in the anterior upregulate the transgene whereas lower concentrations at the boundary result in smaller activation rates. The ratio of the average k_{on} rates in the boundary and anterior is ~ 5, which can be compared to the 4 fold decrease expected from pure Bicoid activation, assuming the Bicoid gradient decays with a length scale of $100\mu m$ [139] and comparing the activation probabilities in the middle of the anterior and boundary regions. Given the crudeness of the this argument stemming from the variability of the Bicoid gradient in the boundary region and the uncertainty of the inferred rates, these ratios are in good agreement and suggest that a big part of the difference in the transcriptional process between the anterior and boundary is due to the change in Bicoid concentration. Of course other factors, such as maternal Hunchback, could also affect the promoter, leading to discrepancies between the two estimates.

The current data coming from four embryos and ~ 50 nuclei in each region with trace lengths of $\sim 300s$ does not make it possible to distinguish between the two and three state models. We asked whether having longer traces or more nuclei could help us better characterize the bursty

properties. We performed simulations with characteristic times similar to those inferred from the data ($k_{on} + k_{off} = 0.01$) assuming a two (Fig. 7.6A) and three state model (Fig. 7.6A). We then inferred the sum of the ON and OFF rates ($k_{on} + k_{off}$) and the ratio of the two OFF rates (k_1/k_2). If the two OFF rates are similar ($k_1/k_2 \sim 1$) we infer a three state model. If one of the rates is much faster ($k_1/k_2 \sim 0$), we infer a two state model. We find that having more nuclei, which corresponds to collecting more embryos, would not significantly help our inference. However looking at longer traces would allow us to disambiguate the two scenarios, if the traces were 4 times longer, or ~ 20 minutes long. Since cell cycle 14 lasts for ~ 45 minutes, analyzing these traces could inform us about the effective structure of the OFF states. However in cell cycle 14, other genes get turned on after 15 minutes, so additional regulatory elements could be responsible for the observed transcriptional dynamics than in cell cycle 12 and 13. Our results suggest that with our current trace length we should be able to identify a two state model with large certainty, but we could not clearly identify a three state model. Our data may thus point towards a more complex model than two state, but a different kind of multistate model or a two state model obscured by other biases cannot be ruled out.

The error bars for the autocorrelation functions describe the variability between nuclei coming from both natural variability and measurement imprecision. While the autocorrelation function is insensitive to white noise, it does depend on correlated noise. The noise increases for large time differences τ, as the number of pairs of nuclei decreases and in our inference we reweigh the points according to their sampling so that the noise does not impair the precision of our inference. The error bars on the inferred parameter are due to variability between nuclei and are obtained from sampling different subsets of the data in each region and cell cycle. Additionally to the inter-nuclei and experimental noise there is natural variability between embryos. Since each nucleus transcribes independently and we assume similar Bicoid concentrations in each of the regions, the inter-embryo variability is of a similar scale as the inter-nuclei variability (Fig. 7.5C), as one expects given that the Bicoid gradient is incredibly reproducible between embryos [139].

7.3.6 Accuracy of the transcriptional process

At the boundary, neighboring nuclei have dramatically different expression levels of the Hunchback protein. From measurements of the Bicoid gradient, Gregor and collaborators estimated that for two neighboring nuclei to make different readouts, they must be able to distinguish Bicoid concentrations that differ by 10% [116]. Following the Berg and Purcell [11] argument for receptor accuracy, and using measurements of diffusion constants for Bicoid proteins from cell cycle 14, the authors showed that, based on protein concentrations, the *hunchback* gene is not able to read-out the differences in the concentrations of Bicoid proteins to the required 10% accuracy in the time that cell cycle 14 lasts. The authors invoked spatial averaging of Hunchback proteins as a possible mechanism that achieves this precision. Spatial averaging can increase precision, but it can also smear the boundary. Erdmann et al calculated the optimal diffusion constant Hunchback proteins must have for the averaging argument to work [140] and showed it is similar to experimental observations [120, 139]. However precision can already be established at the mRNA level and using static measurements Little and co-workers found that the relative variability of the mRNA transcribed from a *hunchback* locus in one nucleus is $\sim 50\%$ [125]. However measurements of cytoplasmic mRNA reduced this variability to $\sim 10\%$ [125].

Here we go one step further and use our direct measurements of transcription from the *hunchback* gene to directly estimate the precision with which the *hunchback* promoter makes a readout of its regulatory environment in a given cell cycle, $\delta P_{\rm on}/P_{\rm on}$. $\delta P_{\rm on}/P_{\rm on}$ is the relative error of the probability of the gene to be ON averaged over the steady state part of a cell cycle. Since the total number of mRNA molecules produced in a given cycle is proportional to $P_{\rm on}$ (shown in Fig. B.7E as a function of embryo length), the precision at the level of *produced* mRNA in a given cycle is equal to the precision in the expression of the gene, $\delta{\rm mRNA}/{\rm mRNA} = \delta P_{\rm on}/P_{\rm on}$. The accuracy of transcription activation is encoded in the stochasticity of gene activation. The gene randomly switches between two states: active and inactive, making a measurement about the regulatory factors in its environment and indirectly inferring the position of its nucleus. Since no additional information is provided by a measurement that is strongly correlated to the previous one, the cell can only base its positional readout on a series of independent measurements. Two measurements are statistically independent if they are separated by at least the expectation value of the time τ_i it takes the system to reset itself:

$$\tau_i \sim \frac{1}{k_{\rm on}^{\rm eff} + k_{\rm off}^{\rm eff}}, \tag{7.3}$$

where in a two state model $k_{\rm on}^{\rm eff} = k_{\rm on}$ and $k_{\rm off}^{\rm eff} = k_{\rm off}$. A more detailed estimate obtained by computing the variance of the time spent ON by the gene during the interphase (see SI Section K) shows that Eq. 7.3 underestimates the time needed to perform independent measurements. We find that for a two state model the accuracy of the readout of the total mRNA produced is limited by the variability of a two state variable divided by the estimated number of independent measurements within one cell cycle:

$$\frac{\delta{\rm mRNA}}{\rm mRNA} = \sqrt{2\frac{\tau_i(1-P_{\rm on})}{TP_{\rm on}}}, \tag{7.4}$$

where T is the duration of the cell cycle and the factor $\sqrt{2}$ is a prefactor correction to the naive estimate. Eq. 7.4 is valid in the limit of $T >> \tau_i$ (the exact result if given in SI Section K). Using the rates inferred from the autocorrelation analysis (Fig. 7.5D) we see that the precision of the gene readout is much lower at the boundary than in the anterior, does not change with the cell cycle and is reproducible between embryos (ordinate in Fig. 7.7A). In the anterior part of the embryo it reaches $\sim 50\%$, while at the boundary, it is very large, $\sim 150\%$, even at cell cycle 13.

We can compare these theoretical estimates with direct estimates of the relative error of the total mRNA produced during a cell cycle , δmRNA/mRNA, from the data. We divide the embryo into anterior and boundary strips, as we did for the inference procedure and calculate the mean and variance of $P_{\rm on}$. These empirical estimates of the precision of the gene measurement calculated agree with the theoretical estimates (Fig. 7.7A). We verified that our conclusions about the scale of our empirical estimates do not depend on the definition of the boundary and anterior regions (Fig. B.7B). To see whether integrating the mRNA produced can increase precision we compared the empirical estimate of the steady state mRNA production (red line in Fig. 7.7B) to the relative error of the total mRNA produced in cell cycle 13 (blue line in Fig. 7.7B) and the total mRNA produced from cell cycle 10 to 13 (green line in Fig. 7.7B) averaged over embryos. We assumed that each nuclei has the total mRNA produced in cell cycle 13, 1/2 of the total mRNA produced by its mother in cell cycle 12, 1/4 of the mRNA produced

by its grand-mother in cell cycle 12 etc. While we see about a 1/3 increase in the precision at the boundary from integrating the mRNA produced in different cell cycles, the estimate in the anterior region is not helped by integration over the cell cycles.

Since we are not able to rule out the three state cycle model as an accurate description of the transcriptional dynamics, we calculated the relative error assuming the same $k_{on} + k_{off}^{eff}$ for a three state cycle ($k_{off}^{eff} = k_1 + k_2$) as for a two state model ($k_{off}^{eff} = k_{off}$) for different values of k_{on} and k_{off}^{eff} (Fig. 7.7C). We found that the relative error is always lower for the three state cycle model and the error decreases, regardless of the duration of the cell cycle, and as expected from Eq. 7.4 as the relative error is decreased by increasing k_{on} and decreasing k_{off}. However the increase in precision from a three state cycle model in the parameter regime we inferred from the fly embryo is relatively modest.

Many previous analysis of precision from static images calculated the relative error of the distribution of a binary variable, which in each nucleus was 1 if the nucleus expressed mRNA in the snapshot, and 0 if it did not express[22, 119]. We analyzed our data using this definition of activity (see Fig. B.7D for mean activity as a function of position) and found that for most embryos the relative error in the anterior drops to zero (Fig. B.7C), indicating that all nuclei in a given region show the same expression state, but at the boundary the precision is still $\sim 50\%$, in agreement with previous reports about the total mRNA in the nucleus [125]. This provides additional evidence for the bursty nature of transcription in the anterior of the embryo.

7.4 Discussion

Contrary to initial reports [18, 19] about the static nature of transcription initiation controlled by the *hunchback* promoter in fly development we show that the promoter is bursty with distinct periods of enhanced polymerase transcription followed by identifiable periods of basal polymerase activity. Our conclusions are based on a new autocorrelation based analysis approached applied to live imaging MS2-MCP data. The data we used in this paper was generated with a modified MS2 cassette [135] compared to the previously published data [18]. However the difference in our conclusions mainly comes from a detailed analysis of the traces.

Quantification of transcription from time dependent fluorescent traces in prokaryotes and mammalian cell cultures has shown that the promoter states cycle through at least three states [129, 130]. In one of these states the polymerase transcribes at enhanced levels, while in most of the remaining states the transcription machinery gets reassembled or the chromatin remodels. We find that in the anterior part of a living developing fly embryo, the *hunchback* promoter also cycles through at least two states, although we cannot conclusively rule out the possibility of more states when the gene is inactive. The main impediment to distinguishing different types of transcriptional cycles comes from the very short durations of the interphase in the early cell cycles when the *hunchback* gene is expressed. We showed that increasing the number embryonic samples would not help us distinguish between two and three state models, however looking at longer time traces would be informative (Fig. 7.6). Since cell cycle 14 lasts about 45 minutes, our analysis shows that the steady state part of the interphase provides enough time to gather statistics that can inform us about the detailed nature of the bursts. Unfortunately, other transcription factors such as the other gap genes regulate *hunchback* expression in cell cycle 14, possibly changing the nature of the transcriptional dynamics in a time dependent manner. We showed that the transcriptional dynamics is constant and reproducible in the earlier cell cycles

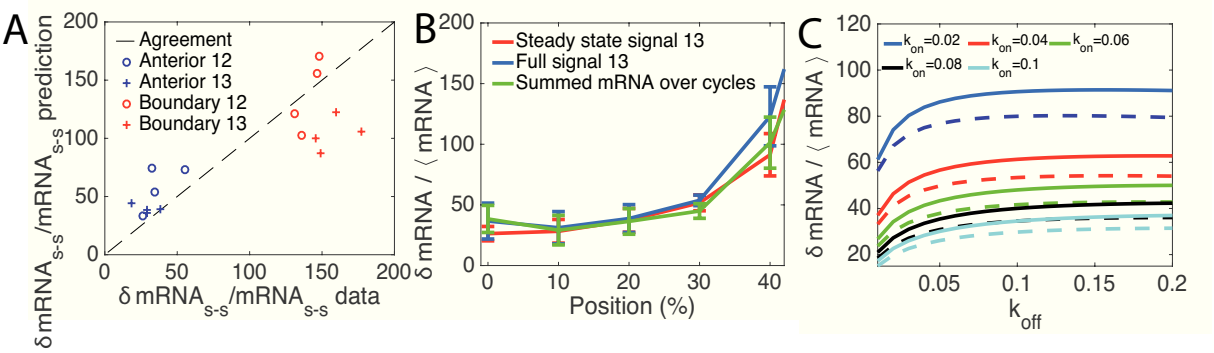

Figure 7.7: **Precision of the *hunchback* gene transcription readout.** A. Comparison of the relative error in the mRNA produced during the steady state of the interphase estimated empirically from data (abscissa) and from theoretical arguments in Eq. 7.4 using the inferred parameters in Fig. 7.5C (ordinate), in the anterior (blue) and the boundary (red) regions, show very good agreement. B. The relative error in the total mRNA produced in cell cycle 13 directly estimated from the data as the variance over the mean of the steady state mRNA production (red line, same data as in A), sum of the intensity over the whole duration of the interphase (blue line) and the total mRNA produced during cell cycles 11 to 13 (green line) for equal width bins equal to 10% embryo length at different positions along the AP axis. Each line describes a an average over four embryos (see Fig. B.7C for the same data plotted separately for each embryo) and the error bars describe the variance. To calculated the total mRNA produced over the cell cycles, we take all the nuclei within a strip at cell cycle 13 and trace back their lineage through cycle 12 to cycle 11. We then sum the total intensity of each nuclei in cell cycle 13 and half the total intensity of its mother and 1/4 of its grandmother. C. Comparison of the relative error in the mRNA produced during the steady state for a two state, $k_1/k_2 = 0$, (solid lines) and three state cycle model, $k_1/k_2 = 1$, (dashed lines) with the same $k_{\text{on}} + k_{\text{off}}^{\text{eff}}$ for different values of k_{on} and k_{off} shows that the three state cycles system allows for greater readout precision.

(12-13) (Fig. 7.2), so independently of the question of the nature of the bursts it would be very interesting to see whether and how it changes when the nature of regulation changes.

Alternatively to looking at longer traces, a construct with two sets of MS2 loops placed at the two ends of the gene that bind different colored probes could be used to learn more about transcription dynamics [141]. We do not have access to data coming from such a promoter , but our analysis approach can be extended to calculate the cross-correlation function between the intensities of the two colored probes. Such cross-correlation analysis have previously been used to study transcription in cell cultures [142], transcriptional noise [143] and regulation in bacteria [144, 145] . Our theoretical prediction for such a cross-correlation function agrees with simulation results (Fig. 7.4C). Unfortunately, the cross-correlation function with one set of probes inserted at the 5' end and the other at the 3' shares the same problems of a 5' construct. For fast switching rates, such a cross-correlation function suffers from the large buffering time ($\sim 300s$ in[19]) drawback of the 5' design and can only be used for inferring large switching rates [146]. However, it does gives us access into dynamical parameters of transcription such as the speed of polymerase and it is able to characterize whether mRNA transcription is in fact deterministic and identify potential introns. Possibly, cross-correlations from two colored probes both inserted closer to the 3' end could be optimal designs.

We assumed an effective model that describes the transcription state of the whole gene and

does not explicitly take into account the individual binding sites. As a result all the parameters we learn are effective and describe the overall change in the expression state of the gene and not the binding and unbinding of Bicoid to the individual binding sites. For concreteness we presented our model assuming a change in the promoter state and constitutive polymerase binding, but our current model does not discriminated between situations where the transcriptional kinetics are driven by polymerase binding and unbinding and promoter kinetics. The presented formalism can be extended to more complex scenarios that describe the kinetics of the individual binding sites and random polymerase arrival times. Since we already have little resolution power to discriminate between these effective models, we chose to interpret the results of only these effective models. The exact contribution of the individual transcription binding sites could be inferred from the activity of promoters with mutated binding sites.

The time traces we had to analyze are very short and finite size effects are pronounced. Unlike in cell culture studies, where long time traces are available, we could not collect enough ON and OFF time statistics to characterize the promoter dynamics from the waiting time distributions. In this paper we show that simple statistics, the auto- and cross-correlation functions are powerful general tools that can be used in these kinds of challenging circumstances.

The approach we propose is a general method that can be used for any type of time trace analysis. However it becomes very useful in studying in vivo biological process where the biology naturally limits the available statistics. In our case the number of ON and OFF events is naturally limited by the short duration of the cell cycles. Our method explicitly calculates correlation functions for short traces, correcting for the finite size effects, and can be also used without making steady state assumptions about the dynamics (although this requires collecting sufficient statistics about two time points, which may be hard for short traces). With these corrections we see that while an effective two state model of the underlying dynamics of transcription regulation holds in the anterior and boundary regions of the embryo in all of the early cell cycles, the rates are different in the boundary and anterior regions, showing a strong dependence on position dependent factors such as Bicoid or maternal Hunchback concentrations. More statistics will make it possible to build more explicit models of Bicoid dependent activation.

In all cases, the rates that we can infer from time dependent traces are naturally limited by the timescales at which the polymerase leaves the promoter, which in our case is estimated to be $\sim 6s$. If the switching rates are faster than this scale, even a perfect, noiseless and infinitely accurate sampling of the dynamics will not be able to overcome this natural limit.

Our method requires knowing the design of the experimental system (number and position of the loops), the speed of polymerase as input and calibrating the maximal fluorescence from one gene. Measurements using two colored probes positioned at a distance on the same gene combined with a cross-correlation function analyses could access parameters such as the speed of polymerase and verify assumptions about the monotonous progression of the polymerase. Such effects can easily be incorporated into the model. While the polymerase speed is an important parameter and erroneous assumption could influence the inference, we have shown that our inference is relatively insensitive to polymerase speeds (see Fig. B.8). In the current experiments we do not have an independent calibration of the maximal fluorescence coming from one gene, which could introduce potential errors in our analysis. However the reproducibility of our results suggests that these potential errors are small.

The presented analysis is an investigation of transcription dynamics from time dependent traces in living functioning organisms. It shows that the functional promoter that controls the

first regulatory steps in fly development is bursty, even in the region with the highest activator concentration. The inferred rates are reproducible between nuclei and embryos and the inter embyo variability is similar to the inner embryo variability (Fig. 7.5A, C and D).

We used the obtained results to estimate the precision of the transcriptional process from the *hunchback* promoter. We found that even in the boundary region the variability in the mRNA produced in steady state by the different nuclei is large, with a relative error of about 50% (Fig. 7.7A). This variability further increases to 150% of the mean mRNA produced at the boundary. These empirical estimates are completely explained by theoretical arguments that treat the gene as an independent measuring device that samples the environment, correcting for the number of independent measurements during a cell cycle. In both cases, the precision at the level of the gene readout is not sufficient to form the precise Hunchback boundary up to half a nuclear width [147]. However, although we can extend our argument to the total mRNA produced in the early cell cycles (Fig. 7.7B), we do not know the amount of maternal *hunchback* mRNA in the nuclei. Having an irreversible promoter cycle could increase the theoretical precision, but only slightly in the parameter regime we have inferred and it would not change the quantitative conclusions about low precision backed by the empirical results.

In the same spirit, the construct we used here was limited to the 500 bp of the proximal *hunchback* promoter, which recapitulates the formation of a sharp boundary at later cell cycles in Fluorescent In Situ Hybridization (FISH) [135]. It is possible that the boundary phenotype is recovered thanks to averaging of mRNAs and proteins produced by the real gene or the transgenes in other nuclei. In the latter case, this would point towards a robust "safety" averaging mechanism that relies on the population. Alternatively, we have to be aware that the sharp boundaries were only detected on fixed samples and that having access to the dynamics of the transcription process likely provides a more accurate view on the process. We calculated and estimated from the data the precision of the gene readout based on the variability of the transcription process between nuclei. We find that the transcriptional process at a given position is quite noisy. Previous estimates of precision were based on static data and did not consider the probability of the gene to be ON, but assumed a binary representation where each nuclei is either active or inactive. By analyzing the full dynamic process we show that the gene is bursty and the transcriptional process itself is much more variable. Reducing the information contained in our traces to binary states, we find precise expression in the anterior, but still large variability at the boundary, similarly to previous results from Fluorescent In Situ Hybridization (FISH)[125].

Assuming that the precision in determining the position of the nuclei is encoded in the precision of the gene readout, a gene with the dynamics characterized in this paper needs to measure the signal ~ 200 times longer at the boundary to achieve the observed $\sim 10\%$ precision. A gene in the anterior would need to integrate only ~ 25 times longer. These results again suggest that the precision in determining the position of the nuclei is not only encoded in the time averaged gene readout, but probably relies either on spatial averaging mechanisms [116, 148, 140] or more detailed temporal information.

In summary, the early developing fly embryo provides a natural system where we can investigate in a functional setting the dynamics of transcription in a living organism. In our data analysis we are confronted by the same limitations that natural genes face: an estimate of the environmental conditions must be made in a very short time. Analysis of dynamical traces suggests that transcription is a bursty process with relatively large inter-nuclei variability, sug-

gesting that simply the templated one to one time-averaged readout of the Bicoid gradient is unlikely. Comparison of mutant experiments can shed light on exactly how is the decision to form the sharp *hunchback* mRNA and protein boundary made.

7.5 Materials and Methods

7.5.1 Constructs

For live monitoring of *hb* transcription activity in Drosophila embryos, we used the MS2-MCP system which allows fluorescent labeling of RNAs as they are being transcribed [121, 18, 146]. To implement the reporter system in embryos, we generated flies transgenic for single insertions of a P-element carrying *hb* proximal promoter upstream of an iRFP-MS2 cassette carrying 24 MS2 repeats [149, 18]. The flies also carry the P{mRFP-Nup107.K} [150] transgenic insertion on the 2^{nd} chromosome and the Pw[+mC]=Hsp83-MCP-GFP transgenic insertion on the 3rd chromosome. These allow the expression of the Nucleoporin-mRFP (mRFP-Nup) for the labeling of the nuclear envelopes and the MCP-GFP required for labeling of nascent RNAs [121]. All stocks were maintained at 25°C.

7.5.2 Live Imaging

Embryo collection, dechorionation and imaging have been done as described in [18]. Image stacks (∼19Z × 0.5μm, 2μm pinhole) were collected continuously at 0.197μm XY resolution, 8bits per pixels, 1200x1200 pixels per frame. A total of 4 movies capturing 4 embryos from nuclear cycle 10 to nuclear cycle 13 were taken. Each movie, due to having different scanned field along the embryos' width, has a different time resolution: 13.1 s, 10.2 s, 5.1 s and 4.3 s.

7.5.3 Image analysis

Nuclei segmentation, tracking and MS2-MCP loci analysis were performed as in [18] and recapitulated here. All steps were inspected visually and manually corrected when necessary. Nuclei segmentation and tracking were done by analyzing, frame by frame, the maximal Z- projection of the movies' mRFP-Nup channel. Each image was fitted with a set of nuclei templates, disks of adjustable radius and brightness comparable with raw nuclei's, from which the nuclei positions are extracted. During the cycle's interphase, each nucleus was tracked over time with a simple minimal distance criterion. For the analysis of MS2-MCP loci detection and fluorescent intensity quantification, the 3D GFP channel (MS2-MCP) were masked with the segmented nuclei images obtained in the previous step. This procedure also helps associating spots to nuclei. We then applied a threshold equal to 2 times the background signal to the masked images and selected only the connected regions with an area larger than 10 pixels. The spot positions are set as the position of the centroids of the connected regions. The intensity of each spot was calculated by summing up all the pixel intensity in the vicinity of the centroids (region of 1.5μm x 1.5μm x 1μm) subtracted to the background intensity extracted from the region around and excluding the spots. In the (rare) case of multiple spots detected per nucleus, the biggest spot was selected.

For each nucleus, we collected the nucleus' position and the spot intensity over time (here referred as "traces"). The traces were then classified according to their respective embryos (out

of 4 embryos), cell cycle (10 to 13) and position along AP axis (either Anterior or Boundary). See SI Section A and Fig. B.1 for examples of traces.

7.5.4 Trace preprocessing

Before the autocorrelation function can be calculated the traces need to be preprocessed. To ensure that the data captures the dynamics of gene expression in its steady state, for each embryos and each cell cycle, we observed the spot intensity only in a specific time window. The beginning and the end of this window is determined as the moment the mean spot intensity over time of all traces (both at the anterior and the boundary) reaches and leaves its an expression plateau (see example in Fig. B.2).

7.5.5 The two state model

The detailed form of the autocorrelation function in Eq. 7.2 depends on the underlying gene promoter switching model. For the two state – telegraph switching model (left panel in Fig. 7.1A) the jumping times between the two states are both exponential and the dynamics is Markovian. The mean steady state probability for the promoter to be ON is $P_{\text{on}} = k_{\text{on}}/(k_{\text{on}} + k_{\text{off}})$, which combined with Eq. 7.1 gives the form of the mean fluorescence $\langle F \rangle$. The probability that the gene is ON at time n given that it was on at time 0 is $A_n = P_{\text{on}} + e^{(\delta-1)n}(1 - P_{\text{on}})$, where $\delta = 1 - k_{\text{on}} - k_{\text{off}}$. The steady state connected correlation function depends only on the time difference (see SI Section B):

$$\langle F(t)F(t+\tau)\rangle - \langle F(t)\rangle^2 = \sum_{i,j} L_i L_j P_{\text{on}}(1 - P_{\text{on}}) e^{(\delta-1)|\tau-j+i|}. \tag{7.5}$$

7.5.6 The cycle model

In the cycle model (center panel in Fig. 7.1A) the OFF period is divided into different sub-steps that correspond to K intermediate states with exponentially distributed jumping times from one to the next. The transition matrix T encodes the rates of this irreversible chain. The probability of the promoter to be in the ON state is:

$$P_{\text{on}} = \frac{k_{\text{on}}}{k_{\text{on}} + \sum_{i=1}^{K-1} k_i}, \tag{7.6}$$

and that the steady state connected autocorrelation at is (see SI Section D):

$$\langle F(t)F(t+\tau)\rangle - \langle F(t)\rangle^2 = \sum_{i=1}^{r}\sum_{j=1}^{r} L(i)L(j)P_{\text{on}} \left[\begin{pmatrix} 1 & 0 \end{pmatrix} e^{(T-\mathbb{1})*|i-j-\tau|} \begin{pmatrix} 1 \\ 0 \end{pmatrix} - P_{\text{on}} \right], \tag{7.7}$$

where τ is counted in polymerase steps. In the simple case of a two state model Eq. 7.7 reduces to Eq. 7.5.

7.5.7 The γ waiting time model

An alternative description of a promoter cycle relies on a reduced description to an effective two state model where we use the fact that the transitions between the states are irreversible. The distribution of times spent in the effective OFF state τ, is no longer exponential, as it was in

the two state model, but it has a peak at nonzero waiting times, which can be approximated by a Gamma distribution

$$\Gamma(\tau) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \tag{7.8}$$

with mean α/β, where β is the scale parameter, α is the shape parameter and $\Gamma(\alpha)$ is the gamma function. The true distribution of waiting times in a cycle model approaches the γ distribution if the OFF rates are all the same and $k_{\text{off}} << 1$. In this limit $\beta \approx k_{\text{off}}$, and α describes the number of intermediate OFF states. In the more general case it correctly captures the effective properties of the process. The mean probability of the promoter to be in ON state in the γ waiting time model is given by

$$P_{\text{on}} = (1 + \frac{\alpha k_{\text{off}}}{\beta})^{-1}. \tag{7.9}$$

The autocorrelation function cannot be computed directly analytically. The steady state Fourier transform of the steady state autocorrelation is (see SI Section E):

$$\mathcal{F}(\langle F(t)F(t+\tau)\rangle - \langle F(t)\rangle^2)(\xi) = \int_{-\infty}^{+\infty} d\tau e^{-2i\pi\tau}(\langle F(t)F(t+\tau)\rangle - \langle F(t)^2\rangle) \tag{7.10}$$

$$= \sum_{k,j} L_k L_j P_{\text{on}} 2\Re\Big[e^{-2i\pi(i-j)}\big[\Big(k_{\text{off}} + 2i\pi\xi - k_{\text{off}}(1 + \frac{2i\pi\xi\beta}{\alpha k_{\text{off}}})^{-\alpha}\Big)^{-1} - \frac{P_{\text{on}}}{2i\pi\xi}\big]\Big].$$

7.5.8 Finite cell cycle length correction to the connected autocorrelation function

Due to the short duration of the cell cycle, the theoretical connected correlation functions need to be corrected for finite size effects when comparing them to the empirically calculated correlation functions. When analyzing the data we calculate the autocorrelation function from M traces $\{\boldsymbol{v_\alpha}\}_{1\le\alpha\le M}$ of the same length K, $\boldsymbol{v_\alpha} = \{v_{\alpha j}\}_{1\le j\le K}$. We calculate the connected autocorrelation function for each trace and normalize it to 1 at time $t = 0$ to avoid spurious nucleus to nucleus variability:

$$c_\alpha(r) = \frac{\sum\limits_{(i,j),|i-j|=r} \left\{\left(v_{\alpha i} - \frac{1}{K}\sum\limits_{l=1}^{K} v_{\alpha l}\right)\left(v_{\alpha j} - \frac{1}{K}\sum\limits_{l=1}^{K} v_{\alpha l}\right)\right\}}{\frac{K-r}{K}\sum\limits_{j=1}^{K}\left(v_{\alpha j} - \frac{1}{K}\sum\limits_{l=1}^{K} v_{\alpha l}\right)^2}, \tag{7.11}$$

and then average over all M traces to obtain the final connected autocorrelation function:

$$c(r) = \frac{1}{M}\sum_{\alpha=1}^{M} c_\alpha(r). \tag{7.12}$$

For $\bar{v} = \langle v_i\rangle$ – the steady state true theoretical average of the random fluorescence intensity over random realization of the process, and $\bar{v^2} = \langle v_i^2\rangle$ – the true theoretical second moment of the fluorescence signal, when $K \to \infty$ the average over time points is equal to the theoretical average, $1/K\sum\limits_{i=1}^{K} v_{\alpha i} = \bar{v}$ and the using time invariance in steady state the autocorrelation function becomes:

$$C_r = \frac{\langle v_i v_{i+r}\rangle - \bar{v}^2}{\bar{v^2} - \bar{v}^2}, \tag{7.13}$$

where $\langle\cdot\rangle$ is an average over random realizations of the process. Eq. 7.13 corresponds to the limit we calculated in the theoretical model. To account for the finite size effects that arise due to short time traces we need to correct for the fact that for short traces $\frac{1}{K}\sum_{i=1}^{K} v_{\alpha i} \neq \bar{v}$ and $\frac{1}{K}\sum_{i=1}^{K} v_{\alpha i}^2 \neq \bar{v^2}$ but both the mean and the variance are functions of K. We note that for short traces the definitions of autocorrelation and autocovariance differ:

$$\sum_{(i,j),|i-j|=r} \left\{ \left(v_{\alpha i} - \frac{1}{K}\sum_{l=1}^{K} v_{\alpha l} \right) \left(v_{\alpha j} - \frac{1}{K}\sum_{l=1}^{K} v_{\alpha l} \right) \right\} \neq \sum_{(i,j),|i-j|=r} \left(v_{\alpha i} v_{\alpha j} - \frac{1}{K^2}\sum_{l=1}^{K} v_{\alpha l} \sum_{m=1}^{K} v_{\alpha m} \right) \quad (7.14)$$

In practice for the analyzed dataset we found that the finite size effects for the variance can be neglected, however the mean over time points is a bad approximation to the ensemble mean. We present the finite size correction to the mean below. For completeness we include the finite size correction for the variance in SI Section I, although we do not use it in the analysis due to its numerical complexity and small effect.

If the variance of the normalized fluorescence intensity over random realizations of the process is well approximated by the average over the K time points, we can replace the denominator in Eq. 7.11 by $\bar{v^2} - \bar{v}^2$ and in steady state evaluate the mean connected autocorrelation function (see SI Section H for details):

$$\begin{aligned} c(r) \;=\; & \frac{1}{\bar{v^2} - \bar{v}^2}\Big[\tilde{C}_r + \frac{1}{K}\left(\frac{1}{K} - \frac{2}{(K-r)}\right)\left(K\tilde{C}_0 + \sum_{k=1}^{K-1} 2(K-k)\tilde{C}_k\right) \\ & + \frac{2}{K(K-r)}\left(r\tilde{C}_0 + \sum_{k=1}^{r-1} 2(r-k)\tilde{C}_k + \sum_{m=1}^{K-1} \tilde{C}_m[\min(m+r,K) - \max(r,m)]\right)\Big] \end{aligned} \quad (7.15)$$

where $\tilde{C}_k = \langle v_i v_{i+k}\rangle$ is the theoretical steady state non-connected correlation function of the process and the average is over random realizations of the process. If $v_i = X(i)$ then C_k is proportional to $A(k)$.

7.5.9 Inference

The inference proceeds in three steps:

Step 1. Signal calibration. The intensity of the measured signal depends on a constant trace dependent offset value I_0, $I(t) = \sum_{i=1}^{r} I_0 a_i L_i$. To calibrate this offset we take the maximum expression to be the mean of the maximun expression over all traces in a given region $I_{\max} = \langle \max_t I(t)\rangle = I_0 \sum_{i=1}^{r} L_i$. The calibrated fluorescence signal used in the analysis is then $F(t) = I(t)/I_0 = \sum_{i=1}^{r} a_i L_i$. P_{on} is directly calculated using Eq. 7.1.

Step 2. Estimating parameter ratios. The ratios of the rates can be estimated directly from the steady state mean fluorescence values using Eqs. 7.6 and 7.9.

Step 3. Estimating parameters. Using the estimate for the ratio of the rates, the ON and OFF rates are found by minimizing the mean squared error between the data and the model.

Chapter 8

Conclusions

8.1 About models in biophysics

I have presented two topics that are very different in biological content: immunology and development. Both of them require the use of tools from statistical mechanics to be modeled and include random networks at their core. In practice the methods used to tackle the challenges they bring are quite different: continuous population dynamics equations in immunology and discrete Markov chains in regulatory networks.

What do physicists bring to the table in biology? At the end of the analysis what stands out is an intuitive but rigorous formalism and the ability to write minimal models capturing the essence of the dynamics. Looking for general laws and simple equations in biological systems goes against the flow of large models including vast numbers of parameters that account for every constituent of the system. Such models do not bring much to the field of biophysics unless particularly smart or powerful ways of implementing them are developed.

Apart from having very few parameters, the models developed also have in common that they are specifically designed to answer a set of questions. It is very important to remember that there is no absolute right description of a system but that the approximations, the choice of scales, the variables have to be adapted to the questions asked. The issue with this vision of modeling is that it is dangerous to rely on a model that can only reproduce one specific feature of the data. In that sense, large models with many predictions that can be compared with experimental results are safer. In particular, in the two examples presented in this Dissertation, different models reproduce the observed data for a specific set of parameters (antigen-based and cytokine-based for immune repertoires, and Poisson or bursty for the autocorrelation functions). While there are several hints that one model is more convincing than the others, a rigorous process of model selection based on one feature is difficult, even when knowledge of the biological system limits the possibilities. For the morphogenesis models, it is eventually the comparison of the precision of the *hunchback* activation boundary that allows to favour the bursty dynamics (updated publication in preparation).

8.2 Future work

The immune models can open up a set of interesting new experiments. To test the validity of the power law result, an experiment could be designed where different mice live in environments with

different levels of pathogenic stimulation (such as fully sterilized, normal laboratory conditions and wildlife conditions). The distribution of clone sizes should vary with the level of pathogenic threats the environment would pose as it would affect both the amplitude and the correlation time of antigenic fluctuations.

In immunology, in terms of model, I would like to extend the analysis to the hypermutation phase of the immune response. A first attempt at this type of models is shown in Appendix A and already hints at a strong effect of hypermutation rates on the distribution of clone sizes. It seems that a lot could be learned from comparing these predictions with observed distribution of antibodies during infections.

In development it seems that much could be learned from a two-colored fluorescent experiment where the crosscorrelation of a $3'$ probe and a $5'$ probe could be computed. This would not only provide another set of data for analysis of gene switching parameters, it would also open new insights in elongation speed, splicing, and variability of polymerase reading paths. Applying the method to other development networks (such as the famous pair rule genes) would also produce fast results about gap genes and morphogen interactions.

From the point of view of modeling, an exciting prospect is to mix the two frameworks in a dynamical analysis of correlation of populations in immune repertoires. This would give much more information than the static picture of clone size distributions and help understand the fine structure of immune dynamics considered in Chapter 5. Another interesting direction to explore is to expand the analysis started in Appendix A.11 to divide clones between naive and memory. In this framework it is possible to investigate the effect of a less abundant pathogenic contribution to the antigenic pool (compared to self antigen) on the dynamics of the naive repertoire. One could then test how sanitized environments affect peripheral immune dynamics and discuss potential links to auto-immune disorders. The cytokine model of Chapter 4 and the models of Chapter 5 both give potential descriptions of the naive repertoire. They could be a way to test what are the potential sources of auto-immune disorders in the dynamics of self-antigen-receptor interactions in the periphery as opposed for instance to problems in positive and negative selection of incoming clones. A theoretical analysis would make a strong case with predictions for clonal populations would help call for a deep sequencing of auto-immune disorder patients repertoires.

More broadly, a lot of the analysis done above calls for a general theoretical framework defining a complex cell state with many variables and limited cell-to-cell fluctuations. This would constitute a solid basis for general work on variability inside and across cell lineages. It would be exciting to see what type of prediction can be made on a system based on assumptions such as noise sources rather than precise descriptions of the dynamics in a general framework (as is done in Chapter 4 in a specific case).

Appendix A

Fitness shapes clone size distributions of immune repertoires: supplementary information

A.1 Simple birth-death process with no fitness fluctuations, and its continuous limit

In this Appendix we derive the steady-state clone size distribution for a system that does not experience any environmental stimulation or noise, but is governed by a birth death process. We will show that the small number fluctuations arising from the discrete nature of birth and death are not sufficient to explain the observed distributions. We also show that our choice of a continuous birth death process is equivalent to its discrete version.

The multiplicative birth–death process corresponds to the following discrete dynamics:

$$\begin{cases} P(n \to n+1) = \mu n dt \\ P(n \to n-1) = \nu n dt, \end{cases} \tag{A.1}$$

where μ is the division rate, ν the death rate. We assume that the population of cells of size n is maintained out of equilibrium by a source of new cells. The steady state solution for cell numbers above the value of the source satisfies detailed balance

$$P(n)\mu n = P(n+1)\nu(n+1) \tag{A.2}$$

and, assuming the death rate is larger than the birth rate, takes the form

$$P(n) \sim \frac{K}{n} e^{-n \log \nu/\mu}. \tag{A.3}$$

The continuous counterpart of this discrete stochastic process corresponds to the following linear-noise approximation:

$$\partial_t C_i = f_0 C_i + \sqrt{(\mu+\nu)C_i}\xi, \tag{A.4}$$

where $\langle \xi_i(t)\xi_i(t')\rangle = \delta(t-t')$ and $f_0 = \mu - \nu < 0$ (and we use the Îto convention). In terms of $x = \log C$ the Langevin equation is

$$\partial_t x = f_0 + \sqrt{\mu+\nu} e^{-x/2}\xi - e^{-x}\frac{(\mu+\nu)}{2}, \tag{A.5}$$

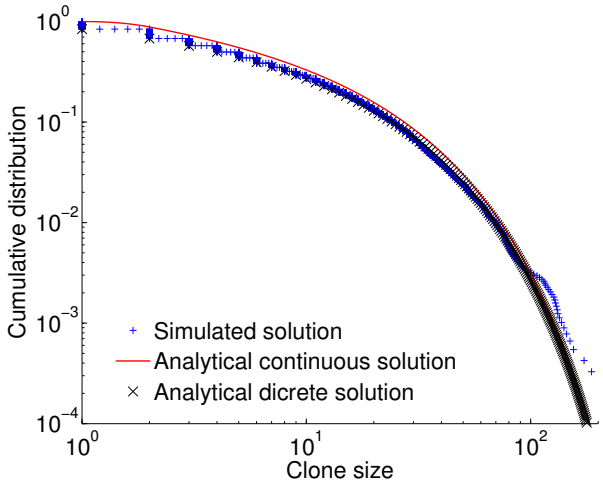

Figure S1: We compare results from a full Gillespie simulation (blue crosses) of a system with only birth-death dynamics with analytical prediction for a discrete system (black crosses, Eq. A.3) and a continuous system (red curve, Eq. A.12). The prediction with discrete variables is more accurate for small clones but the behaviour of all systems is the same for large populations. The parameters are $\nu = 1.45$ day^{-1}, $\mu = 1.5$ day^{-1}, $C_0 = 2$ and we introduce 2000 new clones per day.

and the corresponding Fokker-Planck equation reads

$$\partial_t \rho = \partial_x(-f_0\rho) + \partial_x^2\left(\frac{\mu+\nu}{2}e^{-x}\rho\right) + \partial_x\left(e^{-x}\rho\frac{\mu+\nu}{2}\right) + s(x), \tag{A.6}$$

where $s(x)$ is the distribution of sizes of newly arriving clones. At steady state, we find

$$K - s_C\theta(x-x_0) = -f_0\rho + \frac{\mu+\nu}{2}e^{-x}\rho', \tag{A.7}$$

where K is an integration constant. Defining

$$C_m = (\mu+\nu)/(2|f_0|) \tag{A.8}$$

for $x < x_0$ we obtain

$$\rho(x) = e^{-e^x/C_m}K\int_0^x e^x e^{e^x/C_m} = KC_m(1-e^{-(e^x-1)/C_m}) \tag{A.9}$$

and for $x > x_0$

$$\begin{aligned}\rho(x) \quad = \quad & e^{-e^x/C_m}C_m\Big[Ke^{e^x/C_m} - Ke^{1/C_m} \\ & -\frac{s_C}{|f_0|C_m}e^{e^x/C_m} + \frac{s_C}{|f_0|C_m}e^{e^{x_0}/C_m}\Big]\end{aligned} \tag{A.10}$$

To ensure convergence we set $K = s_C/(|f_0|C_m)$ and the steady solution of the Fokker-Planck equation is

$$\rho(x) = \begin{cases} \frac{s_C}{|f_0|}(1-e^{-(e^x-1)/C_m}) \text{ , if } x < x_0 \\ \frac{s_C}{|f_0|}(e^{e^{x_0}/C_m} - e^{C_m^{-1}})e^{-e^x/C_m}, \text{ if } x > x_0 \end{cases} \tag{A.11}$$

or in terms of the clone size

$$\rho(C) = \begin{cases} \frac{1}{C}(1 - e^{-(C-1)/C_m}), \text{ if } C < C_0 \\ (e^{C_0/C_m} - e^{C_m^{-1}})\frac{e^{-C/C_m}}{C}, \text{ if } C > C_0 \end{cases} \tag{A.12}$$

This result is exactly equivalent to that of Eq. A.3 when $\nu - \mu = |f_0| \ll \mu, \nu$. The accuracy of the approximation is verified in Fig. S1. Even for very large exponential cutoff values, C_m, the apparent exponent is $\alpha = 0$, corresponding to a flat cumulative distribution. This distribution is inconsistent with experiments, regardless of sequencing depth and we conclude that pure birth-death noise is not sufficient to explain the observed distributions.

A.2 Effects of explicit global homeostasis

In the simulations of clone dynamics in a fluctuating environment presented in the "Clone dynamics in a fluctuating antigenic landscape" Results section of the main text, we did not explicitly include a homeostatic control term, but tuned the division and death rates to achieve a given repertoire size. Here we add an explicit homeostatic term to the growth and degradation terms in the Langevin simulations described by Eq. 1 of the main text

$$-h\left[\frac{\sum_i C_i}{N}\right]^r, \tag{A.13}$$

where N is a carrying capacity, h is the homeostatic constant multiplicator and r is the exponent of homeostatic response that described the sharpness of the response when approaching then carrying capacity limit. Comparing in Fig. S2 the resulting clone size distribution obtained with the explicit homeostatic term to the distribution from the simulations in the main text, we see that the explicit homeostatic term does not have an effect on the form of the distribution. It does have an effect on the trajectory of certain clones, and in particular on the response of the system to a very large invasion, making it an important feature of the dynamics of the immune system. However, as shown by the results in Fig. S2 its net effect on the clone size distribution can be taken into account by tuning division and death. When considering specific trajectories in the mean field approximation homeostatic control will add a systematic negative drift to the clonal population and can be accounted for by an additional contribution to f_0.

A.3 Details of noise partition do not influence the clone size distribution function

In the simulation of the dynamics of receptors experiencing a clone-specific fitness presented in the "Clone dynamics in a fluctuating antigenic landscape" Results section of the main text we distributed the noise between the different random distributions: the poisson distributed number of new antigens (s_A), the variance of the initial concentrations ($a_{j,0}$) and the variance of the binding probability (the values of K_{ij}). We made specific choices for this reparation by picking specific parameters of the random processes. Here we show that these specific choices of repartitioning the contributions to the noise do not influence the clone size distributions. Fig. S3 compares clone size distributions obtained with different values of the poisson distributed number of newly arriving antigen N_a and the variance of the Gaussian distributed binding probabilities K_{ij}, reproducing the same distributions in both cases.

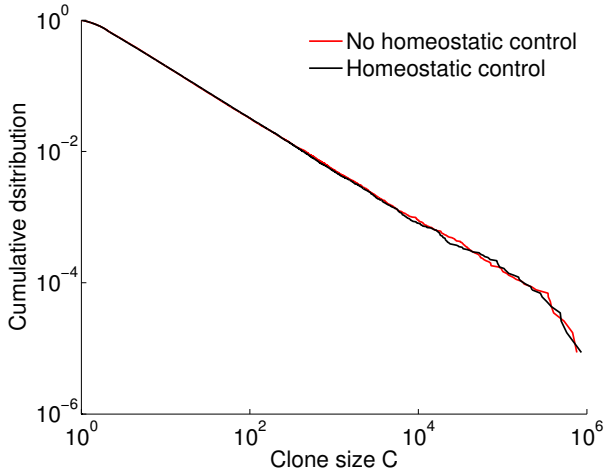

Figure S2: Adding an explicit homeostatic control term does not affect the clone size distribution compared to tuning the degradation and death rates to obtain a given repertoire size as is done in the main text. Comparison of the clone size distribution with an explicit homeostatic control term given by Eq. A.13 (black line) to the distribution presented in the main text (red line). We simulate the Langevin equation for a division rate $\nu = 0.2$ days^{-1}, death rate $\mu = 0.4$ days^{-1}, introduction size $C_0 = 2$, environmental correlation time of $\lambda^{-1} = 0.5$ days and an amplitude of variations of the environment $A = 1.41$ days^{-1} without any homeostatic control for the red curve and with carrying capacity $N = 4 \cdot 10^{10}$ ($h = 1$) and a homeostatic exponent $r = 3$ for the black curve.

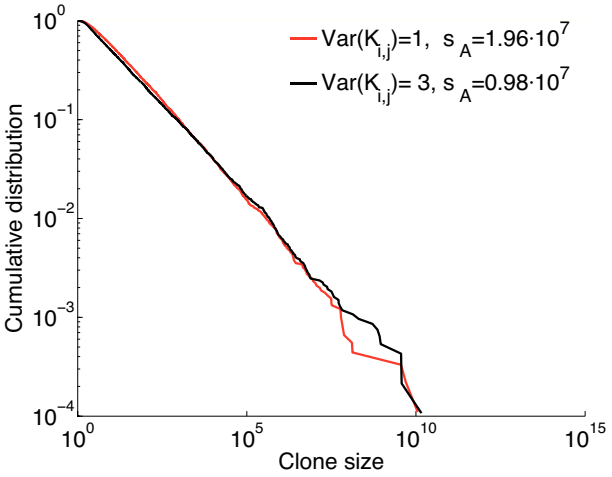

Figure S3: Repartitioning the sources of stochasticity between the number of new antigens per time unit or the variability of binding probabilities does not influence the clone size distributions. We compare simulations of the full system dynamics defined by Eq. 1 of the main text with two sets of values s_A of the poisson distributed number of newly arriving antigen N_a and the variance of the Gaussian distributed binding probabilities K_{ij} that give the same total environmental noise $A^2 = s_\mathrm{A} p a_0^2 \langle K^2 \rangle \lambda^{-1}$. The parameters were taken to be (as in Fig. 1) $s_\mathrm{C} = 2000$ day^{-1} , $C_0 = 2$, day^{-1}, $a_{j,0} = a_0 = 1$, $\lambda = 2$ day^{-1}, $p = 10^{-7}$, $\nu = 0.98$ day^{-1}, $\mu = 1.18$ day^{-1}. For the red curve the variance of the entries of K_{ij} is 1, so that $\langle K^2 \rangle = 2$ and $s_\mathrm{A} = 1.96 \cdot 10^7$ while for the black curve the variance of the entries of K_{ij} is 3, so that $\langle K^2 \rangle = 4$, and $s_\mathrm{A} = 0.98 \cdot 10^7$.

A.4 Model of temporally correlated clone-specific fitness fluctuations

In the "Simplified models and the origin of the power law" Results section of the main text we make a series of approximations to effectively describe the dynamics of immune cells: we first approximate the antigenic environment by a random process with time correlated (colored) noise and we later neglect these temporal correlations. In this section and Appendix A.6 we give the details that lead to the specific forms of the effective equations. In this Appendix we derive the Fokker-Planck equations for the time correlated noise model. In Appendix A.6 we will consider the limit of an infinitely quickly changing environment.

The Langevin equations describing the dynamics of cells experiencing clone specific fitness fluctuations with a finite correlation time are

$$\frac{dC_i}{dt} = [f_0 + f_i(t)]C_i(t) + \sqrt{(\nu+\mu)C_i(t)}\xi_i(t), \tag{A.14}$$

$$\frac{df_i}{dt} = -\lambda f_i(t) + \sqrt{2}\gamma\eta_i(t), \tag{A.15}$$

where $\langle \xi_i(t)\xi_i(t')\rangle = \delta(t-t')$ represents birth death noise in the linear-noise approximation (with the Îto convention) and $\langle \eta_i(t)\eta_i(t')\rangle = \delta(t-t')$ is the noise of antigenic environment. The autocorrelation function of this Ornstein-Uhlenbeck process is

$$\langle f_i(t)f_i(t')\rangle = e^{-\lambda(t+t')}\left(\langle f_i(0)^2\rangle - \frac{\gamma^2}{\lambda}\right) + \frac{\gamma^2}{\lambda}e^{-\lambda|t-t'|}. \tag{A.16}$$

We pick the steady-state value of the initial fitness distribution to cancel the first in Eq. A.16, $\langle f_i(0)^2\rangle = \gamma^2/\lambda$ and obtain

$$\langle f_i(t)f_i(t')\rangle = \frac{\gamma^2}{\lambda}e^{-\lambda|t-t'|}, \tag{A.17}$$

(conditioned on the integral of the net growth rate $f + f_0$ being positive so that the clone does not go extinct). Setting $x = \log C$, we obtain a new set of Langevin equations

$$\partial_t x_i = f_0 + f_i + \sqrt{\mu+\nu}e^{-x_i/2}\xi_i - e^{-x_i}\frac{(\mu+\nu)}{2}, \tag{A.18}$$

$$\frac{df_i}{dt} = -\lambda f_i + \sqrt{2}\gamma\eta_i, \tag{A.19}$$

where the birth-death noise is now treated in the Îto convention. The corresponding Fokker-Planck equation for the distribution of fitness and clone size at time t, $\rho(x, f, t)$, verifies

$$\begin{aligned}\partial_t\rho \;=\; & \partial_x(-f_0\rho) + \partial_f(\lambda f\rho) + \partial_f^2(\gamma^2\rho) + \\ & \partial_x^2\left(\frac{\mu+\nu}{2}e^{-x}\rho\right) + \partial_x\left(e^{-x}\rho\frac{\mu+\nu}{2}\right) \\ & + s(x,f),\end{aligned} \tag{A.20}$$

where $s(x, f)$ is the source of new clones. We solve this equation numerically using finite element methods to obtain clone size distributions for the clone-specific fitness model.

A.5 The Ornstein Uhlenbeck process and maximum entropy

In this Appendix we show that the maximum entropy or maximum caliber process with autocorrelation function $\langle x(t)x(t+s)\rangle = A^2 e^{-\lambda|s|}$ corresponds to the Ornstein-Uhlenbeck process. We consider this continuous maximum entropy process as the continuous limit of a simpler maximum entropy system in discrete time. Burg's maximum entropy theorem [151] states that the maximum entropy process in discrete time that constrains $\langle X_n(t)^2\rangle = A^2$ and $\langle X_n(t)X_{n+1}(t)\rangle = A^2 e^{-\lambda\tau}$ corresponds to the following Markovian dynamics:

$$X_{n+1} = e^{-\lambda\tau}X_n + \sqrt{1-e^{-2\lambda\tau}}A\eta, \tag{A.21}$$

where η is Gaussian white noise. In the limit of $\tau \to 0$ we recover the constrained autocorrelation function in the vicinity of $s = 0^+$: $\langle x(t)^2\rangle = A^2$, $(d/ds)\langle x(t)x(t+s)\rangle|_{s=0^+} = -\lambda A^2$, and Eq. A.21 converges to an Ornstein-Uhlenbeck process.

A.6 Model solution for white-noise clone-specific fitness fluctuations

In the limit of infinitely quickly fluctuating environments, $\gamma \to +\infty$ and $\lambda \to +\infty$ while keeping their ratio $\sigma = \gamma/\lambda$ constant, the autocorrelation of the fitness noise approaches a Dirac delta function, and the fluctuating part of the growth rate $f_i(t)$ converges to Gaussian white noise, $\langle f_i(t)f_i(t')\rangle = 2\sigma^2\delta(t-t')$. Effectively the immune cell dynamics are now described by a one dimensional Langevin equation for the clone size

$$\partial_t C_i = f_0 C_i + \sqrt{2}\sigma C_i\eta_i + \sqrt{(\nu+\mu)C_i(t)}\xi_i, \tag{A.22}$$

where $\langle\eta_i(t)\eta_i(t')\rangle = \delta(t-t')$ follows the Stratanovich convention and ξ_i is as before. The equation for the logarithm of the clone size $x = \log C$ is

$$\partial_t x_i = f_0 + \sqrt{2}\sigma\eta_i + \sqrt{\mu+\nu}e^{-x_i/2}\xi_i - e^{-x_i}\frac{(\mu+\nu)}{2}. \tag{A.23}$$

We explicitly checked that the numerical solution to the clone specific fitness model in Eqs. A.14 and A.15 converged to the dynamics described by Eq. A.22, as demonstrated in Fig. S4.

We now solve this equation analytically, starting with the case of no birth-death noise: Eq. A.22 simplifies to

$$\partial_t C_i = f_0 C_i + \sqrt{2}\sigma C_i\eta_i \tag{A.24}$$

The equation for $x = \log C$ (using the Stratanovich convention) is

$$\partial_t x_i = f_0 + \sqrt{2}\sigma\eta_i, \tag{A.25}$$

with the corresponding Fokker Planck equation

$$\partial_t\rho(x,t) = \partial_x(-f_0\rho) + \frac{1}{2}\partial_x[2\sigma^2\partial_x\rho] + s(x), \tag{A.26}$$

where $s(x)$ is the source term describing the size of newly introduced clones. Assuming a constant initial clone size, $s(x) = s_C\delta(x-x_0)$, the steady state solution is

$$\rho(x) = e^{-\alpha x}\frac{1}{\alpha}\left[Ke^{\alpha x} - K - s_C\sigma^2 e^{\alpha x} + s_C\sigma^2 e^{x_0}\right], \tag{A.27}$$

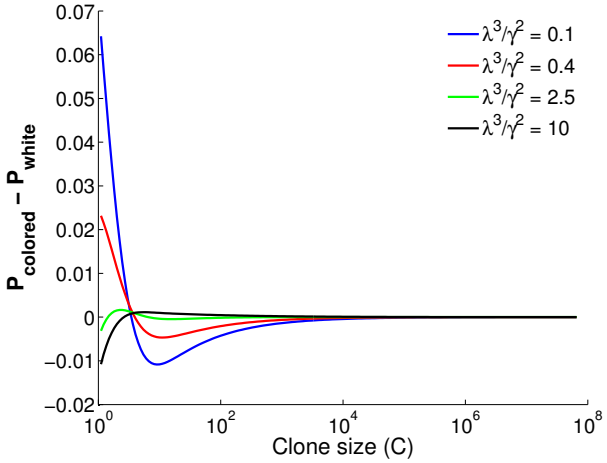

Figure S4: Comparison between clone size distribution obtained as solutions of the time-correlated and time-uncorrelated noise models (without birth death noise). As the values of the dimensionless parameter related to the effective strength of antigen fluctuations relative to their characteristic lifetime λ^3/γ^2 grow the time correlated noise prediction converges to the exact power-law solution of the white-noise model. The cut-off value of the power law decreases with λ^3/γ^2. All simulations performed at a constant value of $\alpha = |f_0|\lambda^2/\gamma^2$ set to 0.5. The value of f_0 is kept fixed to -0.5 days^{-1} for all solutions.

where we have defined

$$\alpha = |f_0|/\sigma^2, \tag{A.28}$$

and K is an integration constant. Imposing that ρ vanishes at infinity sets $K = s_C\sigma^2$ and the final form of the steady state clone size distribution is

$$\rho(x) \quad = \quad \begin{cases} \frac{s_C}{|f_0|}\left(1 - e^{-\alpha x}\right) & \text{if } x < x_0 \\ \frac{s_C}{|f_0|}e^{-\alpha x}\left(e^{x_0} - 1\right) & \text{if } x > x_0, \end{cases} \tag{A.29}$$

or in terms of clone size $C = e^x$,

$$\rho(C) \quad = \quad \begin{cases} \frac{s_C}{|f_0|C}\left(1 - \frac{1}{C^\alpha}\right) & \text{if } C < C_0 \\ \frac{s_C}{|f_0|}\frac{1}{C^{\alpha+1}}\left(\frac{1}{C^{x_0}} - 1\right) & \text{if } C > C_0. \end{cases} \tag{A.30}$$

In all simulations and solutions we find that for large clones, the model of temporally correlated fitness fluctuations behaves as the its white noise limit. This behaviour can be explained by the fact that large clones need a long time to become large. At these long timescales, the characteristic time of noise correlation is negligible and the noise may be approximated as white. For this reason, the exponent α of the power law computed assuming a white noise for the fitness fluctuations is still valid even when that noise is actually correlated in time.

Next, we re-introduce the birth-death noise and solve the general equation. The Langevin equation for $x = \log C$,

$$\partial_t x = f_0 + \sqrt{2}\sigma\eta + \sqrt{\mu+\nu}e^{-x/2}\xi - e^{-x}\frac{(\mu+\nu)}{2} \tag{A.31}$$

Figure S5: We compare simulations of the Langevin dynamics with time correlated antigenic noise with birth-death noise (black line) to the same dynamics without the birth-death noise (red line). All other parameters are kept fixed.We find similar values of the power law exponents but different small clone behaviours. The parameters are $\nu = 0.2$ day^{-1}, $\mu = 0.4$ day^{-1} (for red curve simply $f_0 = -0.2$ day^{-1}) , $C_0 = 2$, $\lambda = 2$ day^{-1} and $\gamma = 1$ day$^{-3/2}$

results in the Fokker-Planck equation for the distribution of clone sizes

$$\partial_t \rho = \partial_x(-f_0\rho) + \frac{1}{2}\partial_x[2\sigma^2\partial_x\rho] + \partial_x^2\left(\frac{\mu+\nu}{2}e^{-x}\rho\right) + \partial_x\left(e^{-x}\rho\frac{\mu+\nu}{2}\right) + s(x). \tag{A.32}$$

Assuming that the initial size is constant, the steady state solution is given by the solution of the inhomogeneous linear equation:

$$K - s_C\theta(x - x_0) = -f_0\rho + \sigma^2\rho' + e^{-x}\frac{\mu+\nu}{2}\rho'. \tag{A.33}$$

The full solution is the sum $\rho = \rho_0 + \rho_1$ of the particular solution,

$$\rho_0(x) = \begin{cases} \frac{K}{|f_0|} \text{ for } x < x_0, \\ \frac{K - s_C}{|f_0|} \text{ for } x > x_0, \end{cases} \tag{A.34}$$

and the solution ρ_1 to the homogeneous equation

$$f_0\rho_1 = \sigma^2\rho_1' + e^{-x}\frac{\mu+\nu}{2}\rho_1' \tag{A.35}$$

of solution:

$$\rho_1(x) = K'\left[\frac{e^x + \frac{(\mu+\nu)}{2\sigma^2}}{1 + \frac{(\mu+\nu)}{2\sigma^2}}\right]^{-\alpha}, \tag{A.36}$$

with $\alpha = |f_0|/\sigma^2$. Therefore, for $x > x_0$

$$\rho(x) = K'\left[\frac{e^x + \frac{(\mu+\nu)}{2\sigma^2}}{1 + \frac{(\mu+\nu)}{2\sigma^2}}\right]^{-\alpha} + \frac{K - s}{|f_0|} \tag{A.37}$$

we set $K = s$ for convergence and obtain the steady state clone size distribution for large x

$$\rho(x) = \left[e^x + \frac{\mu + \nu}{2\sigma^2}\right]^{-\alpha}, \tag{A.38}$$

or in terms of the clone size

$$\rho(C) = \frac{1}{C\left(C + \frac{\mu+\nu}{2\sigma^2}\right)^{\alpha}}. \tag{A.39}$$

We see that the white noise solution with birth–death noise has the same large clone power law behaviour as without birth–death noise. Fig. S5 illustrates how birth death noise in the clone-specific fitness models with time correlated noise also does not affect the power law exponent but only the cut off of the power law.

A.7 Data analysis

In the main text we report values of the power law exponents and power law cut off values obtained from the high throughput sequencing repertoire study of clone size distributions of zebrafish B-cell heavy chain receptors of Weinstein et al. [46]. We extracted the power law exponent and the best fit for the starting point of the power law, defined as its lower bound cutoff, from the discrete clone size distributions plotted in Fig. 1 of the main text using the methods discussed by Clauset and Newman [81]. Specifically, for each point of the cumulative clone size distribution we compute an estimate of the power law exponent with that point as cutoff (i.e the best fit of the power law including only the values of the distribution above that point) using

$$\alpha(C_{\min}) = 1 + n\left[\sum_{i=1}^{n} \log\left(\frac{C_{\mathrm{i}}}{C_{\min}}\right)\right], \tag{A.40}$$

where $C_{\min}$ is the cut off and n is the number of points with y-axis values above $C_{\min}$. For each of these cut-off values we compute the Kolmogorov-Smirnov distance between the data and the estimated power law distribution:

$$d(C_{\min}) = \max_{C>C_{\min}} |F_d(C) - F_e(C; C_{\min})| \tag{A.41}$$

where the maximum is taken over all values above the cut off $C_{\min}$, F_d is the cumulative distribution function (CDF) of the data and $F_e(C; C_{\min})$ is the CDF of the estimated power law distribution with $C_{\min}$ as a cutoff, using Eq. A.40. The the cut off is taken to be the minimum of this distance over all possible cut off values and the exponent is the exponent found for this value.

The obtained power law parameters are presented in Table A.1. The power law exponent gives reproducible values for different individuals and agrees with values of the same exponent obtained from human data [82]. We note that the power law exponent of the cumulative distribution function is α for a power law distribution with exponent $1 + \alpha$. As discussed in detail in the main text, the reliability of the cutoff estimate C^* is sensitive to experimental precision of capturing the rare clones. In the presented dataset the reads were not barcoded and the counts had to be renormalized by a known PCR amplification factor. Therefore, these normalized counts could not to used as normal counts, making the definition of a cut-off clone size problematic. To overcome this problem, we estimate the power law cut-off from the value of the

Fish	$1+\alpha$	C^*	$\log(1-\mathrm{CDF}(C^*))$
A	2.0591	32.6445	- 3.1389
B	2.0214	10.7231	-1.8644
C	2.0708	16.7386	-2.4655
D	2.0670	14.9313	-2.1492
E	2.0529	8.2685	-1.8332
F	2.0006	5.8972	-1.6161
G	1.9867	52.2909	-2.7329
H	2.2242	32.1719	-2.6877
I	2.0835	18.4385	-2.2757
J	1.6907	44.4885	-2.2877
K	1.7641	3.6030	-0.9907
L	1.9417	18.5298	-2.2730
M	1.9901	18.5531	-2.2031
N	1.8877	108.4732	-2.7984

Table A.1: Fit of the power law exponent of the clone size distribution $1+\alpha$ and power law cut-off value C^* for zebrafish B-cell heavy chain D segment data from Weinstein et al [46] presented in Fig. 1. The fit for 14 fish (named A to N) shows a similar fit of the power law exponent.

cumulative distribution function at the cut-off clone size (instead of the cut-off clone size itself). That value is invariant under rescaling of absolute clone size values, unlike C^*.

We notice that the steady state solution is invariant under a full rescaling of time in the equations of the dynamics. This means that the system can be described by two dimensionless parameters, $\alpha = f_0\lambda^2/\gamma^2$ and λ^3/γ^2, and the introduction size C_0. Fitting α to data and assuming value for C_0, we can compare the value of the power law cut-off in data and in simulations to fit the remaining dimensionless parameter, λ^3/γ^2. Estimating f_0 based on thymic output we can predict the order of magnitude of λ and γ.

A.8 Cell specific simulations

In the "A model of fluctuating phenotypic fitness" Results section of the main text, we present results of Fokker-Planck simulations for the cells dynamics. Here we verify that the stochastic dynamics of cells subject to a fluctuating cell-specific fitness are well approximated at the population level by a Fokker-Planck equation with a source term accounting for the import of new clones by comparing its numerical steady-state solution obtained by a finite elements method to explicit Gillespie simulations. We simulated the dynamics of clones using a Gillespie algorithm where cell division and death are accounted for explicitly and depend linearly on a fitness $f_c(t)$ fluctuating according to Eq. 7. The death rate is kept constant (above the average birth rate) and the fluctuations of the fitness only affect the birth rate (with the constraint that the birth rate is always positive). The agreement between the results of this detailed simulation and the Fokker-Planck solution, shown in Fig. S6, validates the linear-noise approximation for the birth-death noise as well as the averaging argument leading to Eq. 8 and 9. This allows us to rely on the Fokker-Planck solution to explore parameter space.

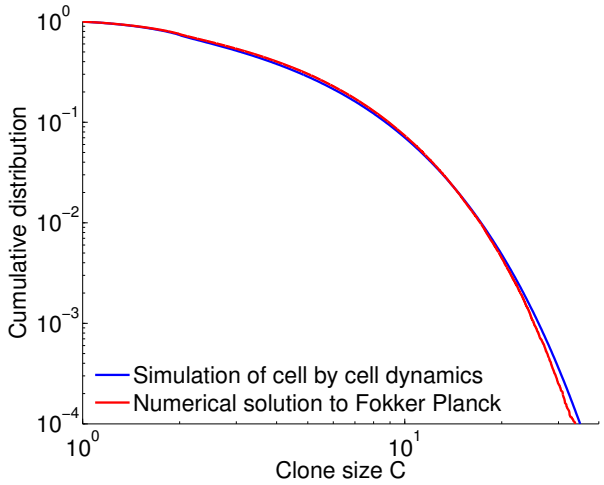

Figure S6: Comparison of the Fokker-Planck solution (red line) and explicit Gillespie simulations of the dynamics (blue line) for the cell specific fitness model discussed in the "A model of fluctuating phenotypic fitness" Results section of the main text, show good agreement allowing us to use the population level Fokker-Planck solution to explore parameter space. Parameters were taken to be $\nu = 0.5$ day^{-1}, $\mu = 0.8$ day^{-1}, $C_0 = 2$, $\lambda_c = 4$ days^{-1} and $\gamma_c = 4$ day$^{-3/2}$.

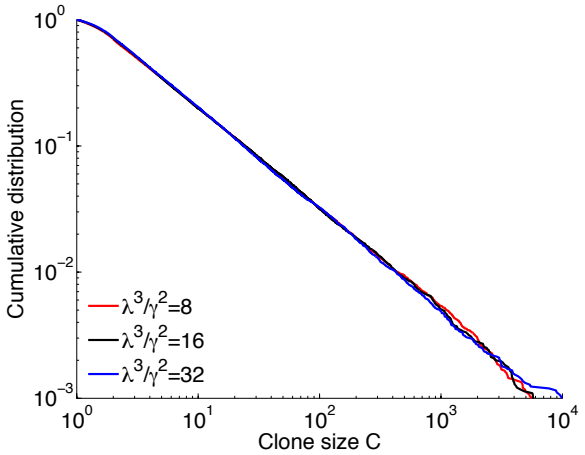

Figure S7: Varying the dimensionless parameter related to the effective strength of antigen fluctuations relative to their characteristic lifetime λ^3/γ^2 does not affect the exponent of the power law if the ratio between exponential decay λ and standard deviation of the variation γ is kept constant. For all three curves the exponent is $\alpha = 0.8$ and $\nu = 0.5$ days^{-1}, $\mu = 0.8$ days^{-1}, $C_0 = 2$ while λ and γ vary.

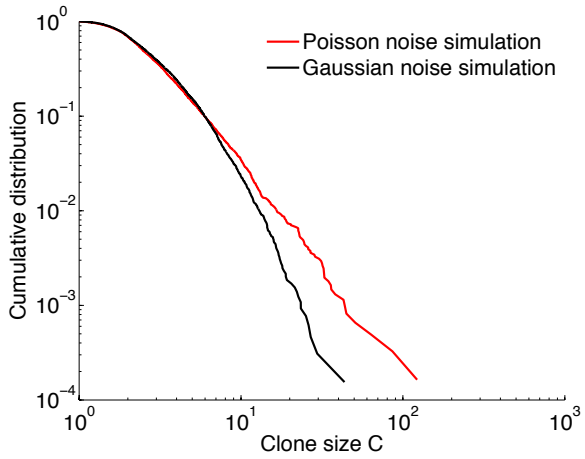

Figure S8: Large deviations can influence the effect of Poisson noise on the simulated clone size distributions and create a discrepancy between Poisson noise (red line) and the Gaussian approximations (black line) we assume in the main text. The discrepancy is most apparent for small clones. We simulated the Langevin dynamics of the Gaussian model with $\nu = 0.5$ day^{-1}, $\mu = 1$ day^{-1}, $C_0 = 2$, $\lambda = 3$ day^{-1} and $\gamma = 1$ day$^{-3/2}$ and the same dynamics with Poisson noise and $\nu = 0.5$ day^{-1}, $\mu = 1$ day^{-1}, $C_0 = 2$, $\lambda = 3$ day^{-1} and $s_{\mathrm{A}} = 10^7$ day^{-1}. In both cases we introduce $s_C = 2000$ new clones per day.

A.9 Model of cell-specific fitness fluctuations, and its limit of no heritability

The cell specific fitness model described in the "A model of fluctuating phenotypic fitness" Results section of the main text arises as a description of a population where each cell experiences its own growth fluctuations but cells deriving from the same lineage remain correlated. In this Appendix we derive the equations that describe the dynamics of clones in this system.

Each cell c experiences a time-correlated multiplicative noise from environmental growth factors. For cells j in a given cell lineage (or clone) i, each individual cell's fitness follows the stochastic dynamics:

$$\partial_t f_c(t) = -\lambda_c f_c + \sqrt{2}\gamma_c \eta_c \tag{A.42}$$

where $\langle \eta_c(t)\eta_c(t')\rangle = \delta(t-t')$. Averaging over all cells in the clone, we obtain

$$\begin{cases} \partial_t C_i = f_0 C_i + f_i C_i + \sqrt{(\mu+\nu)C_i}\xi_i \\ \partial_t f_i = -\lambda_c f_i + \sqrt{\dfrac{2}{C_i}}\gamma_c \eta_i, \end{cases} \tag{A.43}$$

where f_i is the average fitness in clone i

$$f_i(t) = \frac{1}{C_i}\sum_{c\in i} f_c(t), \tag{A.44}$$

and where we have added a birth-death noise term $\sqrt{(\mu+\nu)C_i}\xi_i$. We use the Îto convention for the birth-death noise, $\langle \xi_i(t)\xi_i(t')\rangle = \delta(t-t')$ and the Stratanovich one for the environmental noise $\langle \eta_i(t)\eta_i(t')\rangle = \delta(t-t')$. The equivalent equations for $x = \log C$ are

$$\partial_t x_i = f_0 + f_i + \sqrt{\mu+\nu}e^{-x_i/2}\xi - e^{-x_i}\frac{\mu+\nu}{2} \tag{A.45}$$

$$\partial_t f_i = -\lambda_c f_i + \sqrt{2}e^{-x_i/2}\gamma_c\eta_i \tag{A.46}$$

and the Fokker-Planck equation is

$$\begin{aligned}\partial_t\rho(t,x,f) = &-(f_0+f)\partial_x\rho + \lambda_c\partial_f(f\rho) + e^{-x}\gamma_c\partial_f^2\rho \\ &+ \frac{\mu+\nu}{2}\partial_x(e^{-x}\rho) + \frac{\mu+\nu}{2}\partial_x^2(e^{-x}\rho) \\ &+ s(x,f),\end{aligned} \tag{A.47}$$

where $s(x,f)$ is the joint distribution of size and fitness or newly arriving clones (from thymic or bone marrow output). This is the full Fokker-Planck equation that is solved numerically in the main text using the finite elements method.

Because of the $1/\sqrt{C_i}$ prefactor in front of the noise term, we could expect fitness fluctuations to behave like a birth-death noise in the limit of low heritability ($\lambda_c \to \infty$). In the remainder of this Appendix we show that this is not the case, and we show how to take the limit of no heritability properly.

Consider the limit of $\lambda_c \to \infty$ and $\gamma_c \to \infty$, keeping the ratio γ_c/λ_c constant, so that f does not become infinitesimally small. The equation for the environmental stimulation f in $x = \log C$ space is given by (in Stratanovich convention)

$$\partial_t f = -\lambda_c f + \sqrt{2}\gamma_c e^{-x/2}\eta. \tag{A.48}$$

Direct integration gives

$$f(t) = \sqrt{2}\gamma_c\int_0^t e^{-\lambda_c u}e^{-x(t-u)/2}\eta(t-u)du \tag{A.49}$$

and we divide the integral into two sub-integrals for $k > 0$

$$\begin{aligned}f(t) = &\sqrt{2}\gamma_c\int_{k/\lambda_c}^t e^{-\lambda_c u}e^{-x(t-u)/2}\eta(t-u)du \\ &+\sqrt{2}\gamma_c\int_0^{k/\lambda_c} e^{-\lambda_c u}e^{-x(t-u)/2}\eta(t-u)du.\end{aligned} \tag{A.50}$$

With infinite precision, for any value of t, we set the integral of η to be bounded and obtain the first integral is with probability $1-\epsilon$ smaller in norm than

$$\sqrt{2}\gamma_c\sqrt{t}K(\epsilon)e^{-k}, \tag{A.51}$$

where $K(\epsilon)$ is a constant to control the variations of the integral of ξ with probability ϵ (the time factor for the control of the integral is in the $\sqrt{t}$).

The second sub-integral is

$$\begin{aligned}\sqrt{2}\gamma_c \int_0^{k/\lambda_c} e^{-\lambda_c u} e^{-x(t-u)/2}\eta(t-u)du \\ \approx e^{-x(t^-)/2}\eta(t)\sqrt{2}\frac{\gamma_c}{\lambda_c}(1-e^{-k}).\end{aligned} \tag{A.52}$$

We choose $k = \sqrt{\lambda_c}$ and in the limit of $\lambda_c \to \infty$ and $\gamma_c \to \infty$ keeping $\gamma_c/\lambda_c = \text{const}$ we obtain the final form of environmental fluctuations

$$f(t) \longrightarrow \sqrt{2\frac{\gamma_c}{\lambda_c}} e^{-x(t^-)}\eta(t), \tag{A.53}$$

where t^- means the left-hand limit. $f(t)$ depends only on the past, which means that in $x = \log C$ space the noise is similar to a birth-death noise in the Îto convention. Yet in terms of clone sizes C additional Îto terms make the effect of environmental fluctuations different from classical birth-death dynamics.

A.10 Model solutions for cell-specific fitness fluctuations in the limit of no heritability

In this Appendix we solve the model of cell-specific fitness fluctuations in the limit where trait heritability is low. In this limit, the dynamics is described by a model with an instantaneous random fitness that is uncorrelated for cells in the same clone. The resulting Langevin equation reads:

$$\frac{dC_i}{dt} = f_0 C_i + \sqrt{2C_i}\frac{\gamma_c}{\lambda_c}\eta_i + \frac{\gamma_c^2}{\lambda_c^2} + \sqrt{(\mu+\nu)C_i}\xi_i \tag{A.54}$$

where all noise is treated in the Îto convention, and where the extra term γ_c^2/λ_c^2 comes the converting back the low-heritability limit of the fitness fluctuations, given by Eq. A.53, into $C = e^x$ space. We note that although the fitness and birth-death noise have very similar forms, the birth-death noise is self-generated and intrinsic, while the fitness noise is environmental and extrinsic. This small difference greatly affects the steady-state clone size distribution.

To see this, we first consider the case of no birth-death noise. In the cell-specific fitness model consider the following equations with the Stratanovich rule:

$$\begin{cases}\partial_t C_i = f_0 C_i + f C_i, \\ \partial_t f_i = -\lambda_c f_i + \sqrt{\frac{2}{C_i}}\gamma_c \eta_i,\end{cases} \tag{A.55}$$

and its equivalent for $x = \log(C)$

$$\begin{cases}\partial_t x_i = f_0 + f_i, \\ \partial_t f_i = -\lambda_c f_i + e^{-x_i/2}\gamma_c \eta_i\end{cases} \tag{A.56}$$

In Appendix A.9 we have shown that in the limit of $\lambda_c \to \infty$ and $\gamma_c \to \infty$, the system reduces to the one dimensional equation

$$\partial_t x_i = f_0 + e^{-x_i/2}\sqrt{2}\frac{\gamma_c}{\lambda_c}\eta_i \tag{A.57}$$

with the Îto rule for the white noise η_i. The corresponding Fokker-Planck equation is

$$\partial_t \rho = \partial_x(-f_0\rho) + \frac{1}{2}\partial_x^2\left[\frac{2\gamma_c^2}{\lambda_c^2}e^{-x}\rho\right] + s(x). \tag{A.58}$$

Assuming a deterministic introduction size $s(x) = s_C\delta(x - x_0)$, at steady-sate we get

$$K - s_C\theta(x - x_0) = -f_0\rho + e^{-x}\frac{\gamma_c^2}{\lambda_c^2}\rho' - \frac{\gamma_c^2}{\lambda_c^2}\rho e^{-x}, \tag{A.59}$$

which for $x > x_0$ is solved by

$$\rho(x) = e^{-e^x/C_m+x}\Big[KEi(e^x/C_m) - KEi(C_m^{-1}) \tag{A.60}$$

$$-\frac{s_C\lambda_c^2}{\gamma_c^2}Ei(\frac{e^x}{C_m}) + \frac{s_C\lambda_c^2}{\gamma_c^2}Ei(\frac{e^{x_0}}{C_m})\Big], \tag{A.61}$$

where K is an integration constant, Ei is the exponential integral function and

$$C_m = \frac{\gamma_c^2}{|f_0|\lambda_c^2}. \tag{A.62}$$

The divergence of Ei at infinity sets $K = s_C\lambda_c^2/(\gamma_c^2)$ and the clone size distribution is

$$\rho(x) = \begin{cases} \left(Ei(e^x/C_m) - Ei(C_m^{-1})\right)e^{-e^x/C_m+x} & \text{for } x < x_0 \\ \left(Ei(e^{x_0}/C_m) - Ei(C_m^{-1})\right)e^{-e^xC_m+x} & \text{for } x > x_0 \end{cases} \tag{A.63}$$

or in terms of $x = \log C$

$$\rho(C) = \begin{cases} e^{-C/C_m}\left(Ei(C/C_m) - Ei(C_m^{-1})\right) & \text{for } C < C_0 \\ e^{-C/C_m}\left(Ei(e^{x_0}/C_m) - Ei(C_m^{-1})\right) & \text{for } C > C_0 \end{cases} \tag{A.64}$$

The validity of this solution is checked in Fig. S9 and the convergence of the full solution of Eq. A.47 (with no birth-death noise) to the analytical solution in the limit of no heritability ($\lambda_c \to \infty$) is show in Fig. S10.

For comparison, in a pure birth-death process (no fitness fluctuations) the clone-size distribution is, for C large enough, $\rho(C) \sim e^{-C/C_m}/C$ where $C_m = (\mu + \nu)/(2(\mu - \nu))$, as shown in Appendix A.1. These two solutions both have an exponential cutoff, but have very different power-law exponents, corresponding to $\alpha = 0$ and $\alpha = -1$, respectively.

We now add the birth-death noise, *i.e.* consider both types of noise, still in the limit of no heritability. The corresponding Langevin equation reads:

$$\partial_t x_i = f_0 + \sqrt{\mu + \nu}e^{-x_i/2}\xi - e^{-x_i}\frac{\mu + \nu}{2} + e^{-x_i/2}\frac{\sqrt{2}\gamma_c}{\lambda_c}\eta \tag{A.65}$$

where all noise is in the Îto convention. Integrating the Fokker Planck associated to this equation gives at steady state condition

$$K - s_C\theta(x - x_0) = -f_0\rho + \left[\frac{\mu + \nu}{2} + \frac{\gamma_c^2}{\lambda_c^2}\right]e^{-x}\rho' - \frac{\gamma_c^2}{\lambda_c^2}e^{-x}\rho. \tag{A.66}$$

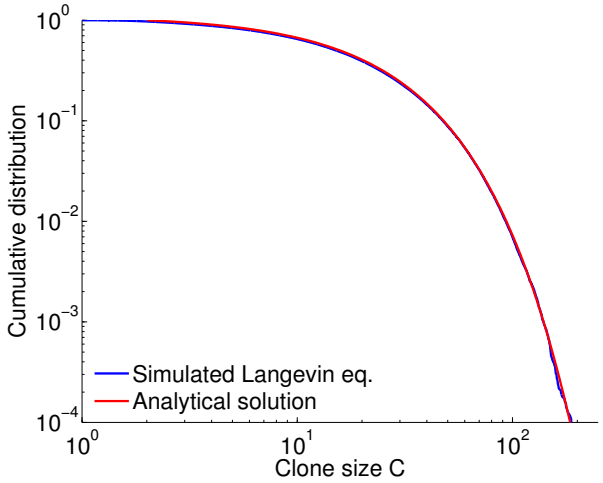

Figure S9: The result of a simulation of the Langevin equation of the white noise cell-specific fitness model (blue line) compared to the analytical prediction of Eq. A.64 (red line) show very good agreement. The parameters are $\nu = 0.2$ day^{-1}, $\mu = 0.4$ day^{-1}, $C_0 = 2$, $\lambda_c = 4$ day^{-1} and $\gamma_c = 8$ day$^{-3/2}$.

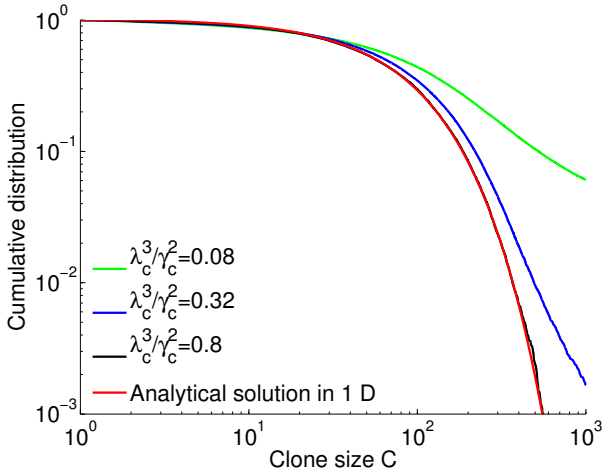

Figure S10: Convergence of the cell-specific fitness models (Eq. A.47) without birth-death noise to Eq. A.64 in the limit of no heritability ($\lambda_c \to \infty$). For all four curves $\alpha = 0.2$. Parameters used: $\nu = 0.2$ day^{-1}, $\mu = 0.25$ day^{-1}, $C_0 = 2$ and 1000 new clones introduced each day.

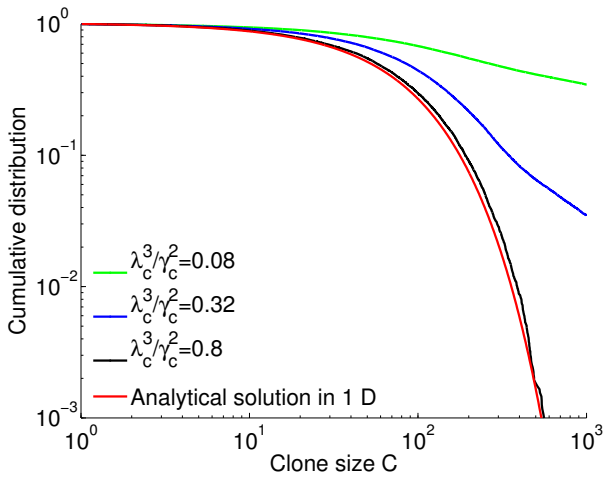

Figure S11: Convergence of the cell-specific models (Eq. A.47) with birth-death noise to the analytical result of Eq. A.68 (red line). Keeping constant α while $\lambda_c \to \infty$ and $\gamma_c \to \infty$ we recover the solution of Eq. A.68. Parameters are the same as in Fig. S10

In order for ρ to be well defined we set $K = s_C$. For $x > x_0$ the equation is homogeneous and solved by separation of variables:

$$\frac{d\rho}{\rho} e^{-x} \left[\frac{\mu + \nu}{2} + \frac{\gamma_c^2}{\lambda_c^2} \right] = \left(f_0 + \frac{\gamma_c^2}{\lambda_c^2} e^{-x} \right) \rho, \tag{A.67}$$

and gives the solution:

$$\rho(C) = \frac{K e^{-C/C_m}}{C^{1+\alpha}}, \tag{A.68}$$

with

$$\alpha = -\left(1 + \frac{(\mu + \nu)\lambda_c^2}{2\gamma_c^2} \right)^{-1}, \tag{A.69}$$

which is a power-law with an exponent $0 \leq 1 + \alpha \leq 1$ and an exponential cutoff

$$C_m = (\mu - \nu)^{-1} \left(\frac{\mu + \nu}{2} + \frac{\gamma_c^2}{\lambda_c^2} \right). \tag{A.70}$$

The convergence of the solution of the full system, Eq. A.47, to this solution is checked in Fig. S11.

A.11 Dynamics of naive and memory cells

In this section we present our results on the division of the population between naive and memory cells and its impact on the distribution of clone sizes. In our simulations and analysis so far we have always considered the system to be uniform, because most of the data available at this time is not sorted into naive and memory/effector cells and because the main difference between naive and memory cells (higher stimulation of memory cells by binding events) is already included in our models.

In principle, memory and naive cells could have a completely different set of parameters. None of the values of these parameters are known with high, accuracy although it emerges from

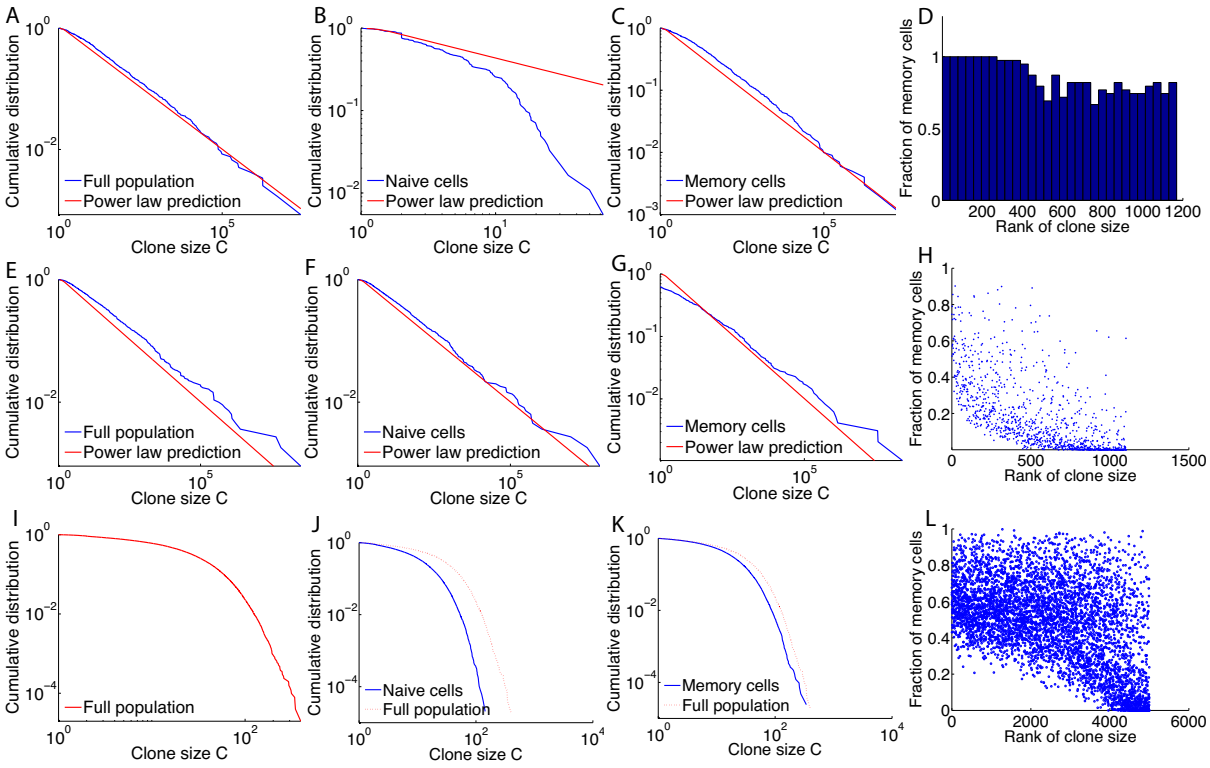

Figure S12: Simulation results for clone and cell specific model with two cell compartments for naive and memory. Panels A to D are results from clone-specific fitness model with a switching rate θ from naive to memory taken to be infinite (the whole clone switches instantly to memory when above a fitness threshold) and fitness threshold $f_{\text{mem}} = 1$ day^{-1}. Panels E to H are results for a model with clone-specific fitness with a finite switching rate $\theta = 0.05$ days^{-1} and fitness threshold $f_{\text{mem}} = 1$ day^{-1}. For both clone-specific simulations the parameters are: $s_{\text{C}} = 200$ day^{-1} , $C_0 = 2$, $s_{\text{A}} = 1.96 \cdot 10^7$ day^{-1}, $\langle a_{j,0} \rangle = 1$, $\text{Var}(a_{j,0}) = 1$, $\lambda = 2$ day^{-1}, $p = 10^{-7}$, $\nu = 0.98$ day^{-1}, $\mu = 1.18$ day^{-1}. Panels I to L are results from simulations of a model with cell-specific fitness with a switching rate $\theta = 0.25$ and threshold $f_{\text{mem}} = 0.5$. The other parameters are: $s_{\text{C}} = 10^4$ day^{-1}, $C_0 = 2$, $\lambda_c = 2$ day^{-1}, $\gamma_c = 4$ day$^{-3/2}$, $\nu = 0.5$ day^{-1}, $\mu = 0.7$ day^{-1}. Panels A, E and I show the clone size distribution of the whole population adding memory and naive contributions to each clone and the power law prediction from the white noise model for clone-specific fitness. Panels B, F and J show the clone size distributions of the naive pool of cells compared to the white noise prediction for the clone-specific fitness (B, F) and the full population distribution for the cell-specific dynamics (J). Panels C, G and K show the clone size distributions of the memory pools (same comparisons as for naive). Panels D, H and L show the fraction of memory cells in clones as a function of their rank (biggest clones have smallest ranks) as a histogram for an infinite switching rate (because clones are either all naive or all memory) and as scatter plots for the two other types of dynamics.

all studies that memory cells have a higher turnover rate (or death rate μ) than naive cells. However, our estimate of f_0 (which is the average division rate minus the death rate) cannot be performed for separate groups of naive and memory cells without knowledge of their total population and the rate of conversion from naive to memory cells. For these reasons we keep the same effective f_0 for the whole population.

We model the immune system with two pools of cells: naive and memory/effector for both the clone-specific and cell-specific fitness models. Clones from the naive pool with fitness over a given threshold f_{mem} turn irreversibly into memory cells at a certain rate θ per day. In both cases the two pools have the same dynamics but memory cells have a higher turnover: the death rate μ and the basal birth rate ν are higher in the memory pool but their difference f_0 is unchanged. This means the birth-death noise is higher in the memory pool. We find that in the clone-specific fitness model it does not affect the power-law exponent of the clone-size distribution, but it does affect strongly the distribution (and more specifically the cutoff value C_m) in the cell-specific fitness model, as birth-death noise is of the same order of magnitude as the environmental noise (Fig. S12).

In the clone-specific fitness model, we find that the distribution still displays power-law behavior with the expected exponent (Fig. S12A and E). For very high rates of conversion from naive to memory we see that naive cell distributions drop exponentially above a threshold, as all high fitness clones are completely converted into memory (Fig. S12B). For lower rates of conversion both memory and naive pools have heavy tails and the memory pool has a higher power law cutoff for small values (Fig. S12F and G). For the cell-specific fitness model we find that the memory pool can have significantly heavier tails (as its dynamics is much faster) and a higher cutoff C_m (a power-law like behavior in a wider range) than the naive pool (Fig. S12A-B-C). In all cases we recover that naive clones are smaller than memory clones, or in other words large clones are mostly made up of memory cells (Fig. S12D-H-L).

A.12 Effects of hypermutations

In this section we show that including the effect of somatic hypermutations in the clone-specific fitness dynamics does not change the power law behavior of the distribution. We model the somatic hypermutations by replacing a small fraction of the offspring of the fastest expanding clones by new clones with binding affinities close to the ones of their parents. For each clone such that $f_i > f_{\text{hyp}}$, offspring with hypermutated receptors are being produced with rate r_{hyp}. A large fraction r_{del} of those are assumed to have acquired deleterious mutations and are removed from the pool. The rest (fraction $1 - r_{\text{del}}$) form new clones of size 1 (in our definition, which differs from the usual convention for B cells, a clone is a subset of cells with the exact same receptor sequence). The interaction matrix $K_{i',j}$ of each new, hypermutated clone i' is formed from the interaction matrix $K_{i,j}$ of its progenitor i by changing each non-zero entry of $K_{i,j}$ to:

$$K_{i',j} = \begin{cases} 0 \quad \text{with probability } 1 - p_{\text{hyp}} \\ \psi K_{i,j} + (1 - \psi) + \sigma_{\text{hyp}} \zeta \quad \text{otherwise,} \end{cases} \tag{A.71}$$

where ψ is a parameter controling the heritability of the values of the K entries, and p_{hyp} the probability that the specificity to a given antigen is passed on to the hypermutated offspring; ζ is a Gaussian variable of mean 0 and variance 1. To compensate the loss of specificity, zero

entries of $K_{i,j}$ are assigned new, non-zero values of binding affinities with probability $(1-p_{\rm hyp})p$ (where we recall that p is the probability for a given clone to be specific to a given antigen), so that the number of non-zero values of K remains the same on average. The value of these new binding affinities are drawn completely at random, as before (no inheritance).

A small part of the hypermutated clones branch out and undergo affinity maturation, meaning that they are selected generation after generation. Their fitness increases until the environment varies enough for their branch to be obsolete and decay back to low fitnesses. The effect of hypermutations on the distribution depends on the ratio between the speed at which hypermutated lineages drift in fitness space and the time scale for variations of the environment (λ^{-1}).

Somatic hypermutations add a source of stochasticity in fitness and increase the number of large clones. Accordingly, simulations of the model with hypermutations (see Fig. S13) show that the clone size distribution still exhibits power law behavior, but with a lower exponent (heavier tails) due to the extra stochasticity induced by hypermutations.

A.13 Time dependent source terms and aging

In this section we investigate the effect of a decaying thymic output on the distribution of clones for the antigen recognition based model. In all our simulations we assume that the source of new clones (thymic output) produces a number of clones that is on average constant with time. It is an approximation since in humans or in mice thymic output is high at birth and during growth and slowly decreases during adult life. This decrease is very slow compared to the time scales involved in this analysis [38] and so within the time frames considered it can be considered constant. In this section we look at the effect of this decrease over long time scales.

We model the decrease of thymic output with an exponentially decaying (with time) source term. In real organisms, homeostatic control ensures that the total number of cells in the body is conserved during this reduction of thymic output. We do not model this homeostatic control explicitly, but rather tune the difference between birth and death rates f_0 to keep the total population constant on average, which we showed was equivalent (see Fig. S2). Simple averaging of the dynamics shows that

$$\frac{d\langle N\rangle}{dt} = f_0 N + n_C\langle f_i C_i\rangle + s_C \tag{A.72}$$

where n_C is the number of clones in the system and N is the total number of cells. Since our source term is a function of time, to have on average a constant total population size we need to define :

$$f_0(t) = -\frac{n_C(t)\langle f_i C_i\rangle + s_C(t)}{N}. \tag{A.73}$$

We show the results of a simulation in Fig. S14 with $s_C = s_{C,0}e^{-t/\tau}$, $\tau = 8.3$ yr. We recover results known in humans and get predictions for the behavior of the exponent of the power law at different ages. We find that, with the decrease of thymic output, the number of clones is decreasing (Fig. S14C), meaning that clones become on average fitter (*i.e.* better at recognizing antigens), but at the expense of repertoire diversity. Keeping the population constant (Fig. S14D) slowly decreases the decaying rate of clones $|f_0|$ and so is expected to decrease the exponent, which behaves as $\alpha = \lambda|f_0|/A^2$. Accordingly, simulations show a clear power-law

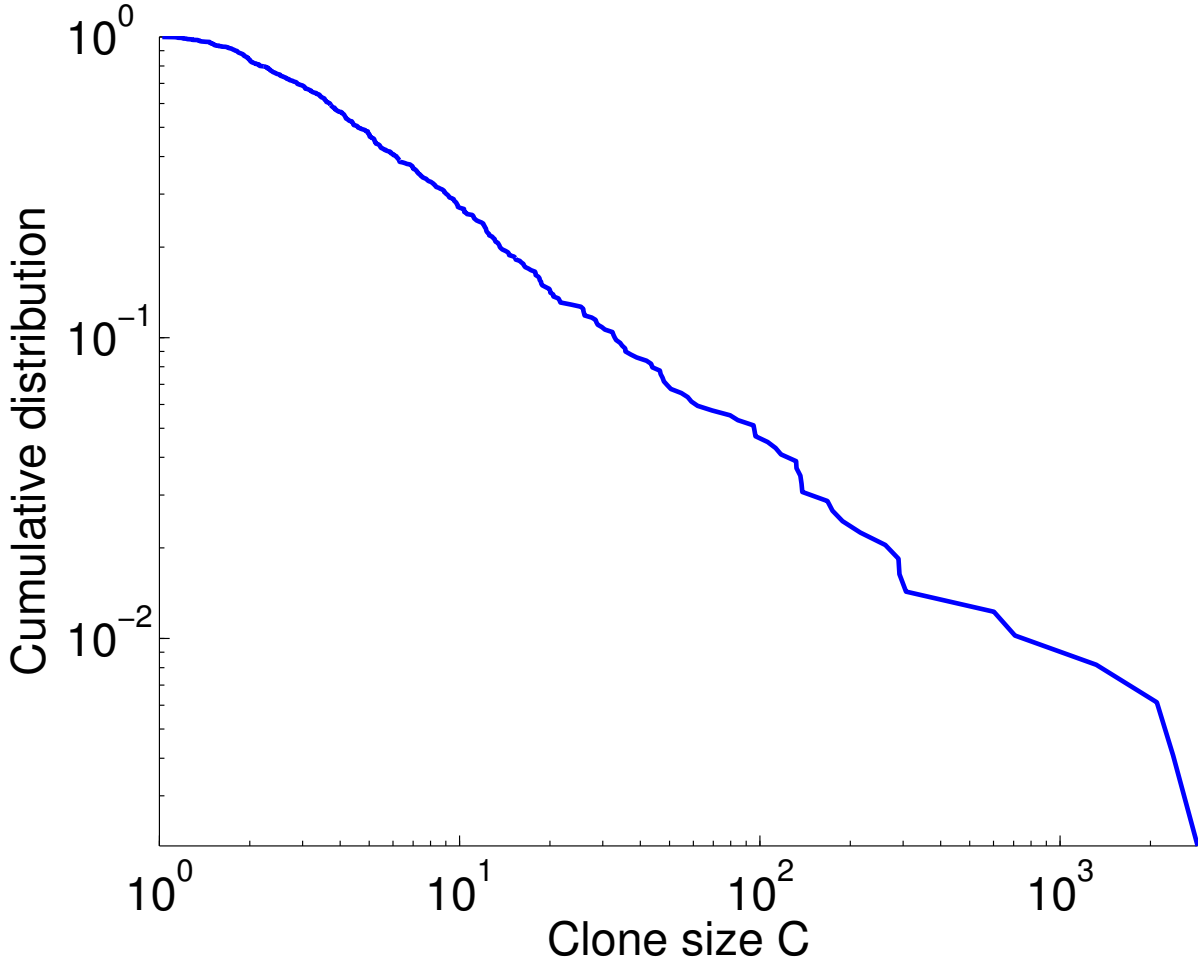

Figure S13: We show the clone size distribution that results from simulating a model of clone-specific fitness with somatic hypermuations as described in Appendix A.12 and Eq. A.71. The distribution exhibits clear power law behavior. Hypermutation parameters are: $f_{\text{hyp}} = 4$ days^{-1}, $r_{\text{hyp}} = 0.01$ days^{-1}, $r_{\text{del}} = 0.01$, $p_{\text{hyp}} = 0.5$, $\psi = 0.7$ and $\sigma_{hyp} = 0.05$. Other parameters are: $s_{\text{C}} = 200$ day^{-1} , $C_0 = 2$, $s_{\text{A}} = 1.5 \cdot 10^7$ day^{-1}, $\langle a_{j,0} \rangle = 1$, $\text{Var}(a_{j,0}) = 1$, $\lambda = 2$ day^{-1}, $p = 10^{-3}$, $\nu = 0.75$ day^{-1}, $\mu = 1.15$ day^{-1}. Non zero $K_{i,j}$ entries from thymic output have mean 1 and standard deviation 0.3.

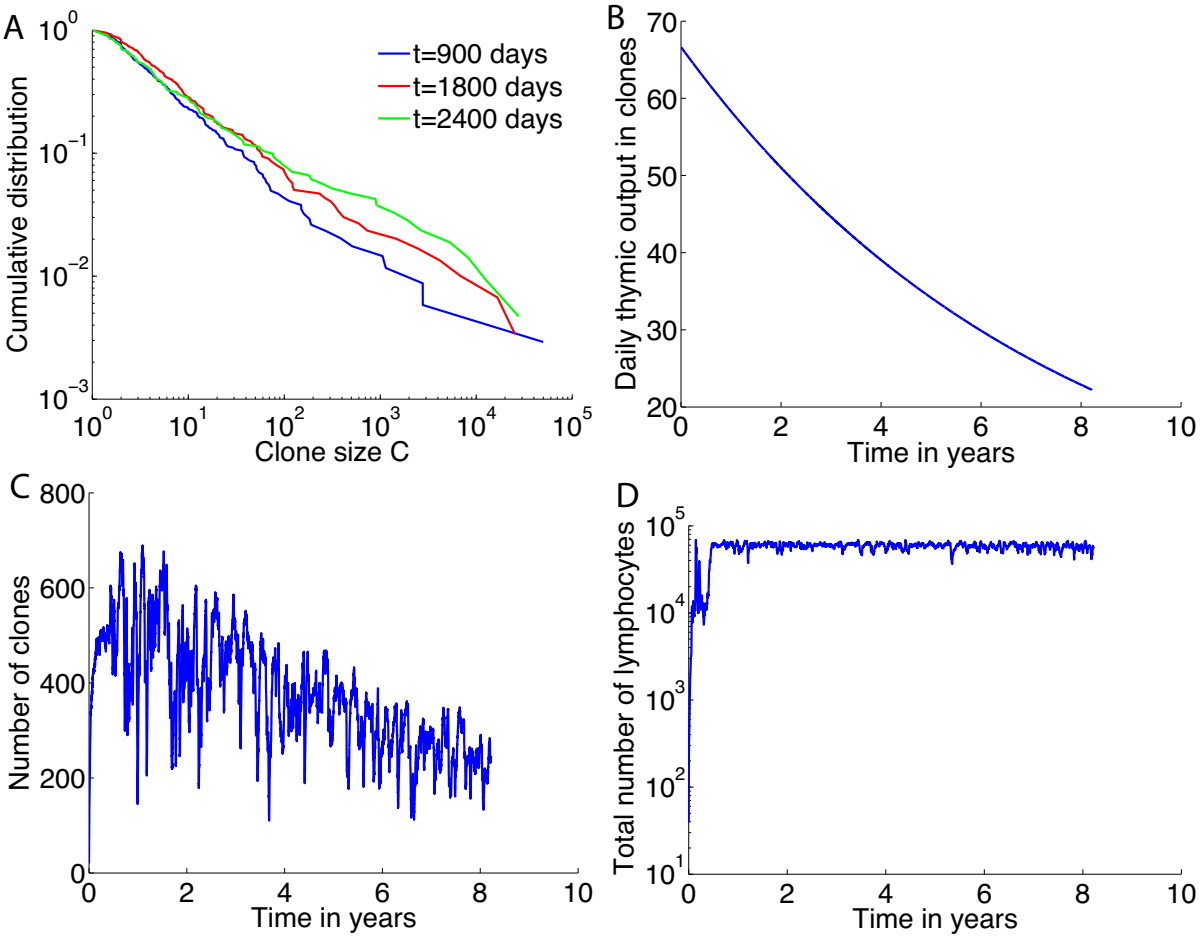

Figure S14: Results of a simulation of a model of clone-specific fitness with a decaying source term and balancing decrease of $|f_0|$ to keep the population size constant. A. The clone size distributions at different time points maintains a power law behavior with an exponent α that decreases with time. B. Decay of the thymic output with time. C. Total number of clones is decreasing with time. D. Total number of cells is maintained by tuning the rate f_0. Parameters used are: source decay timescale $tau = 8.3$ yr, $s_{\mathrm{C},0} = 200$ day^{-1} , $C_0 = 2$, $s_{\mathrm{A}} = 1.5 \cdot 10^7$ day^{-1}, $\langle a_{j,0} \rangle = 1$, $\mathrm{Var}(a_{j,0}) = 1$, $\lambda_c = 2$ day^{-1}, $p = 10^{-7}$, $\nu + \mu = 1.9$ day^{-1}, $f_0 = -0.4$ day^{-1} at time $t = 0$.

behavior in the clone-size distribution (Fig. S14A), with the tail of the distribution becoming heavier with age. We thus expect older organisms with lower thymic output to have a larger tail in their clone-size distribution. We predict thymectomy to lead to distributions with very fat tails.

Appendix B

Precision of readout at the hunchback gene: supplementary information

B.1 Basic setup and data preprocessing

The raw data produced experimentally is a fluorescent signal $I(t)$ measured at discrete times corresponding to the sampling time frame of the movie (see SIFig. B.1 for examples of traces). At each locus and at each time point it is the sum of the background signal and a number of fluorescent molecules attached to loops formed by the mRNA. Each loop contributes to the signal by a constant I_0. This constant is unknown and can vary from trace to trace due to noise in the experimental setup and the variability in the locations of the nuclei in the embryo. All models are written for the renormalized signal $F(t) = I(t)/I_0$.

Because the fluorescent signal is produced by discrete polymerases that travel down the gene, we divide the gene into chunks of 150 base pairs, a length that corresponds to the irreducible space occupied by a polymerase on the gene (Fig. 3 in the main text). The positions the polymerase can occupy on the gene are labeled by an index $1 \leq i \leq r$. The number of MS2 loops that have been formed by a polymerase that has reached a given position depends only on the MS2 gene construct and we define a deterministic function L_i for the whole length of the gene that describes the number of MS2 loops that have been produced by a polymerase at position i. In practice the exact number of loops is not an integer and varies from base pair to base pair so we take L_i as the average number of loops at this polymerase position (see Fig. 3 in the main text).

When the gene is fully loaded with polymerases (the number of polymerases is equal to the length of the gene divided by 150 bp), the fluorescence intensity is $I(t) = I_0 \sum_i^r L_i$. Assuming that the maximum of the signal over the whole trace is a good approximation for the fully loaded value we can determine I_0 and renormalize the data. In practice, since we see variability in the expressed signal in different nuclei at the same position, we are not sure the fully loaded polymerase scenario occurs in each nuclei, so we take the mean of the maximum intensity values in the anterior. We use this renormalized fluorescence signal to infer the parameters of the dynamics.

The experimental data is analyzed assuming the system is in steady state and does not take into account the initial activation period after mitosis. and the end of the trace when the gene is

kon($1/s$)	mov1	mov2	mov3	mov4
12A	0.078	0.056	0.009	0.023
12B	0.004	0.005	0.003	0.011
13A	0.017	0.020	0.014	0.021
13B	0.004	0.006	0.004	0.005

Table B.1: The inferred k_{on} rates from the autocorrelation approach assuming a two state model for the four embryos and cell cycle 12 and 13, in the anterior and boundary.

deactivated before mitosis. We take only the middle window of the trace as shown in SIFig. B.2.

In all models based on a stochastic gene switching (so all models except the Poisson model) we assume that the gene can be in several states with only two effective transcription rates: a non zero transcription rate in the ON state and an basal production rate equal to zero in the OFF state. When the gene is ON the polymerase loads at a maximal rate set by clearing of the binding site by the previous polymerase, which is one polymerase every 6 seconds (calculated as the irreducible polymerase length along the gene 150 bp divided by the polymerase speed, $v = 25bp/s$). The state of the gene is described by a stochastic process $X(t)$ that is equal to 1 when the gene loads polymerase (i.e is ON) and 0 when the gene is OFF (see Fig. 1B in the main text). Once the polymerase is loaded its path is assumed to be deterministic with constant speed.

The gene can be described by the locations where there is a polymerase: we define $a(i,t)$ as a function of time t and position $1 \leq i \leq r$ that is equal to 1 if there is polymerase at position i at time t and 0 otherwise (see Fig. 1D in the main text). The fluorescence signal is then a convolution of the polymerase position, $a(i,t)$, and the details of the loop design of the MS2 construct, L_i:

$$F(t) = \sum_{i=1}^{r} L_i a(i,t), \tag{B.1}$$

and the polymerase position can easily be translated back to the gene state through the deterministic relation, $a(i,t) = X(t-i)$ (see Fig. 3D in the main text for the form of L_i). This disruption is exact for a system with a discrete regulatory process and a discrete time step equal to the polymerase time step. Unfortunately, the moments in time when the gene switches are not necessarily multiples of the natural coarse graining steps of the system (the polymerase time step and its equivalent length) so it is necessary to introduce a continuous time in the system. We will present results for both the discrete and continuous time models. The continuous description is valid in the limit where the typical time spend by the gene in each state is long compared to the polymerase step or equivalently the gene switching constants are small compared to 1/6 s^{-1}. See SI Section B.2 for a more detailed argument.

B.2 The two state model

In this section we derive the equations required for the inference of the dynamics under the assumption that the gene can be in two states: ON or OFF represented by a two dimensional vector $x(t) = [x_{\text{on}}(t), x_{\text{off}}(t)]$. $x_{\text{on}}(t)$ is the probability of the gene to be ON and $x_{\text{off}}(t)$ is the probability for the gene to be OFF. $x_{\text{on}}(t)$ is the average over traces of the random variable $X(t)$

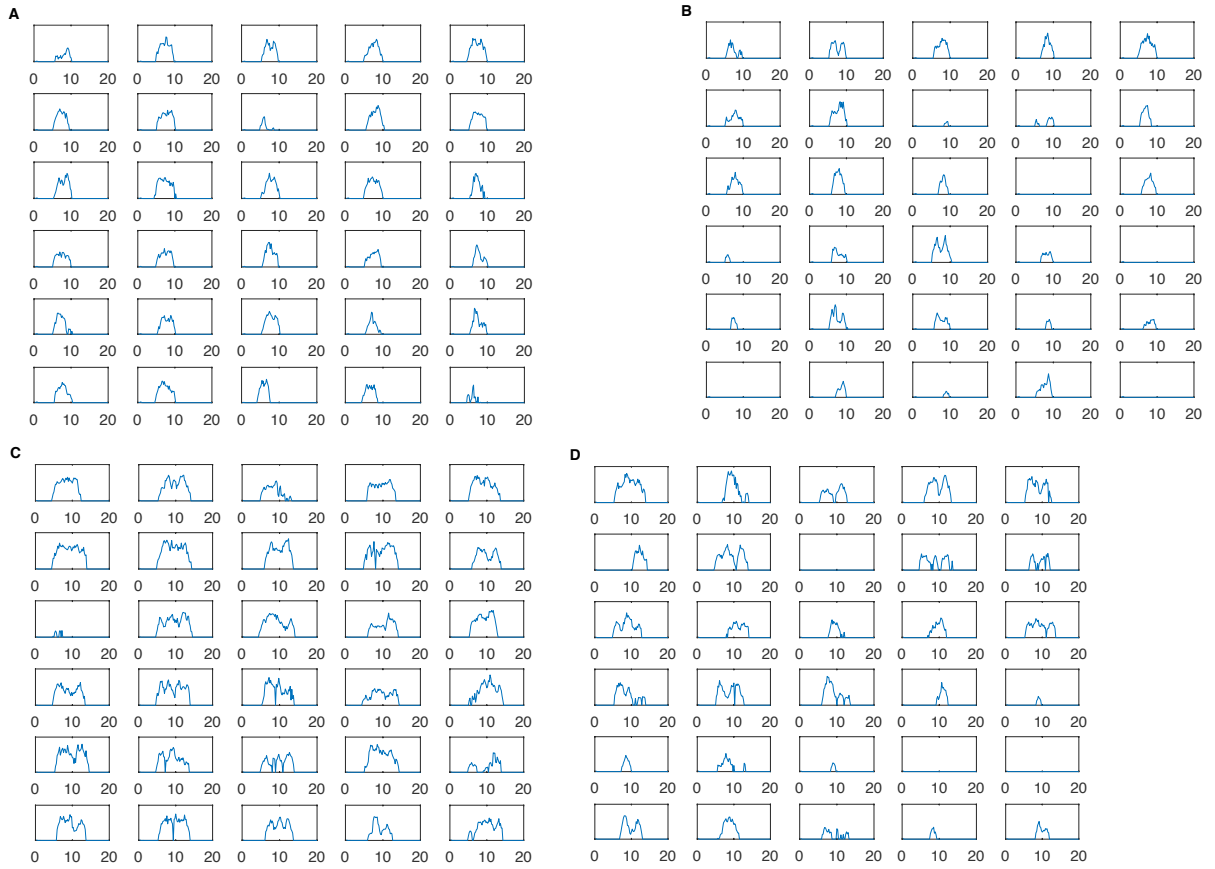

Figure B.1: **Examples of individual spot intensity over time**. Consecutively shown are the traces in (A) Cycle 12, Anterior, (B) Cycle 12, Boundary (C) Cycle 13, Anterior, (D) Cycle 13, Boundary. The x axis is time in minute and y axis is the spot intensity in AU.

koff$(1/s)$	mov1	mov2	mov3	mov4
12A	0.060	0.088	0.008	0.019
12B	0.020	0.034	0.021	0.051
13A	0.018	0.031	0.016	0.027
13B	0.031	0.054	0.031	0.064

Table B.2: The inferred k_{off} rates from the autocorrelation approach assuming a two state model for the four embryos and cell cycle 12 and 13, in the anterior and boundary.

depicted in Fig. 1B of the main text. We assume that the switching times between the two are exponentially distributed:

$$\partial_t \begin{pmatrix} x_{\text{on}} \\ x_{\text{off}} \end{pmatrix} = \begin{pmatrix} -k_{\text{off}} & k_{\text{on}} \\ k_{\text{off}} & -k_{\text{on}} \end{pmatrix} \begin{pmatrix} x_{\text{on}} \\ x_{\text{off}} \end{pmatrix}. \tag{B.2}$$

The steady state probability to be ON is $P_{\text{on}} = x_{\text{on}}(t = \infty) = 1/T \sum_t x_{\text{on}}(t)$, where T is the duration of the steady state window in Fig. B.2, and is:

$$\frac{k_{\text{on}}}{k_{\text{off}}} = \frac{P_{\text{on}}}{1 - P_{\text{on}}}. \tag{B.3}$$

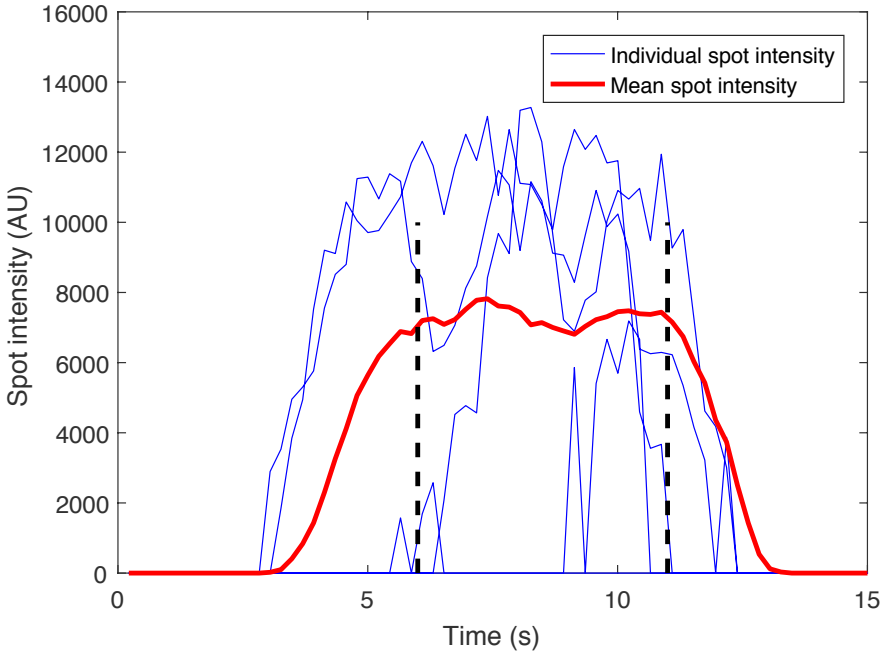

Figure B.2: **Data calibration**. Shown are examples of 5 (out of 154) individual traces (blue) taken from embryo 1, cycle 13. Also shown is the mean spot intensity over time of all traces (red). The steady state window is chosen to be from the 6^{th} minute to the 11^{th} minute (dashed lines).

We learn P_{on} from Eq. 1 in the main text:

$$\langle F \rangle = P_{\text{on}} \sum_{i=1}^{r} L_i. \tag{B.4}$$

and use it to obtain the ratio of the switching rate from Eq. B.3.

The autocorrelation function is:

$$\langle F(t)F(s) \rangle = \sum_{i=1}^{r} \sum_{j=1}^{r} L_i L_j \langle a(i,t)a(j,s) \rangle, \tag{B.5}$$

where the brackets are an average over traces (different realizations of the random process). We define $A(t-i, s-j) = 1/x_{\text{on}}(s-j)\langle a(i,t)a(j,s) \rangle$ – the probability that the polymerase is at position i and time t given that there was a polymerase at position j at time s (here we assume that $t-i \geq s-j$). Using the deterministic relation between the polymerase position at a given time $a(i,t)$ and the probability to be on at an earlier time $X(t-i)$, $A(t-i, s-j)$ is equivalent to the probability that the gene in ON at time $t-i$ given that is was ON at time $s-j$:

$$A(t-i, s-j) = x_{\text{on}}(t-i| \text{ ON at time } s-j). \tag{B.6}$$

Plugging the expression into Eq. B.5 we obtain Eq. 2 in the main text:

$$\langle F(t)F(s) \rangle = \sum_{i=1}^{r} \sum_{j=1}^{r} L_i L_j x_{\text{on}}(s-j) A(t-i, s-j). \tag{B.7}$$

In steady state the system is translationally invariant $A(t-i, s-j) = A(|t-i-s-j|)$ and for brevity we will denote is as $A(n)$ - the probability that the gene is ON at time n, given that

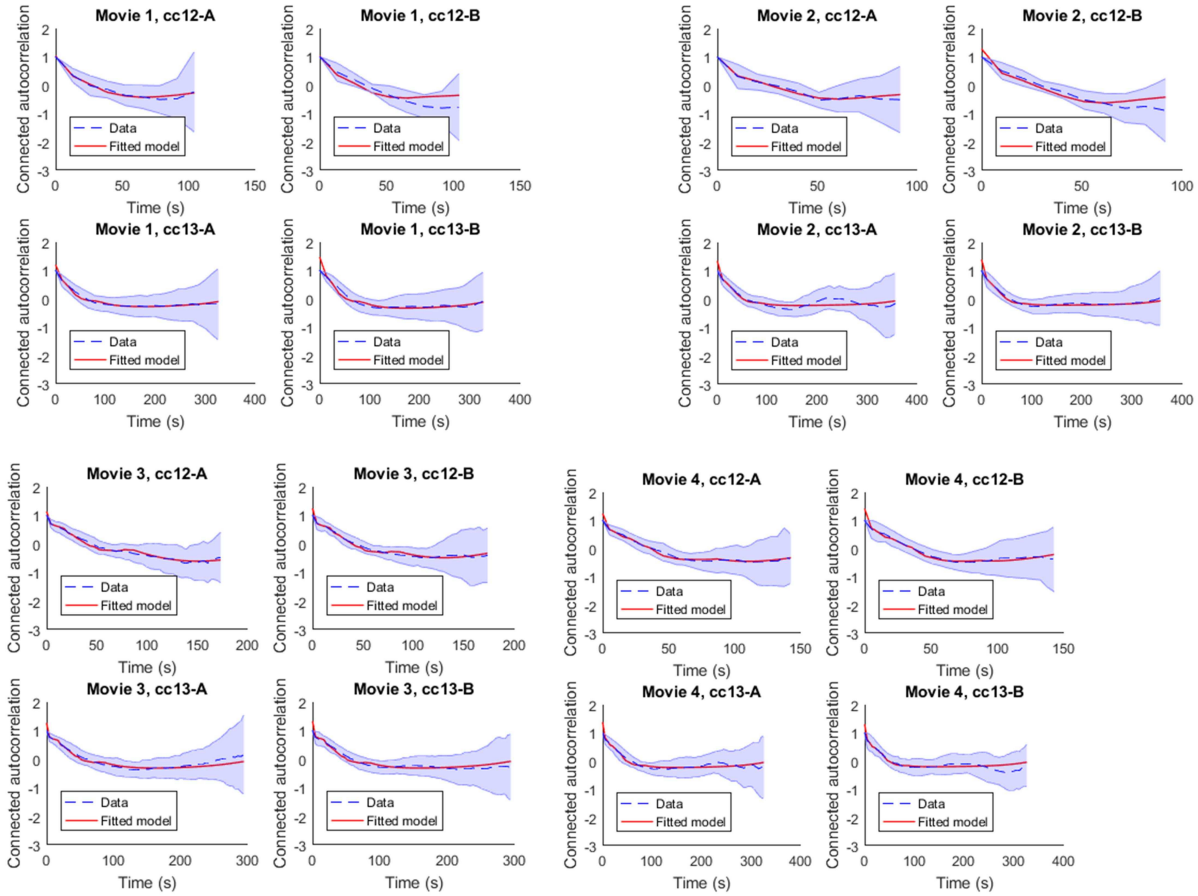

Figure B.3: **Fits of the autocorrelation function**. The empirical autocorrelation function for both the anterior and boundary regions in all four embryos is fit using the autocorrelation function with the finite size corrections for the two state model.

it was ON at time 0. To find A_n we need to solve for $x(t)$:

$$\partial_t x(t) = (T - \mathbb{1})x(t), \tag{B.8}$$

where $T - \mathbb{1}$ is given by Eq. B.2 and calculate the expectation value that the gene is ON at time t given in was ON initially:

$$A_n = \begin{pmatrix} 1 & 0 \end{pmatrix} e^{n(T-\mathbb{1})} \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{B.9}$$

Eq. B.9 is correct in a continuous time model. Its discrete time equivalent is

$$A_n = \begin{pmatrix} 1 & 0 \end{pmatrix} T^n \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{B.10}$$

In the limit of k_{on} and k_{off} much smaller than the polymerase step they are also much smaller than 1 and $e^{n(T-\mathbb{1})} \simeq \mathbb{1} + n(T - \mathbb{1}) \simeq (\mathbb{1} + (T - \mathbb{1}))^n$. In this limit the continuous and discrete time descriptions of Eq. B.9 and Eq. B.10 are equal.

The eigenvalues of $T - \mathbb{1}$ are $[1, \delta]$, where $\delta = 1 - k_{\text{on}} - k_{\text{off}}$ with corresponding eigenfunctions:

$$\begin{pmatrix} P_{\text{on}} \\ P_{\text{off}} \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \tag{B.11}$$

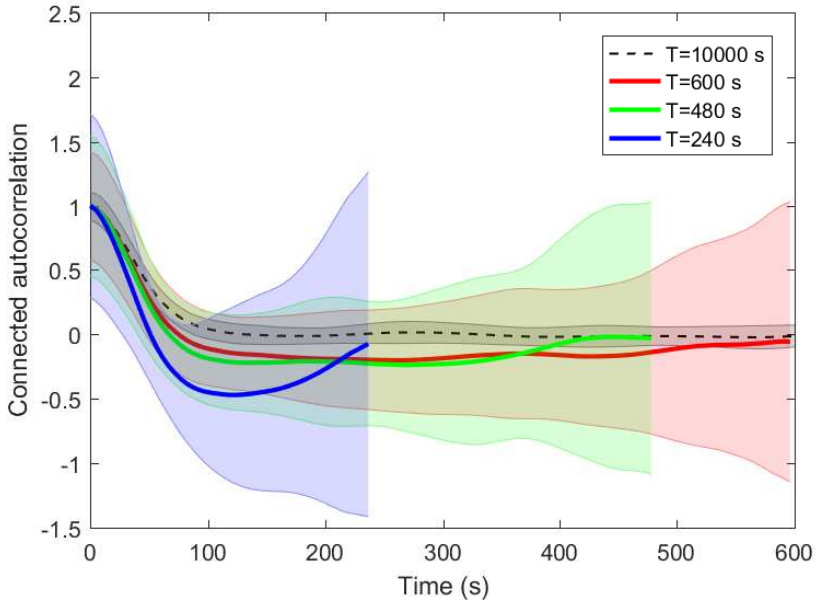

Figure B.4: **Example of the connected autocorrelation function for the two state model calculated for different trace lengths as a function of time T.** The shaded areas denote the standard variation over xx simulated traces. The switching rates $k_{\text{on}} = k_{\text{off}} = 0.01s^{-1}$ and the number of nuclei $M = 500$.

The transition matrix T is

$$T = \frac{1}{P_{\text{on}} + P_{\text{off}}} \begin{pmatrix} P_{\text{on}} & 1 \\ P_{off} & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \delta \end{pmatrix} \begin{pmatrix} 1 & 1 \\ P_{off} & -P_{\text{on}} \end{pmatrix} \tag{B.12}$$

and

$$e^{n(T-\mathbb{1})} = \begin{pmatrix} P_{\text{on}} + e^{n(\delta-1)}P_{\text{off}} & P_{\text{on}} - e^{n(\delta-1)}P_{\text{on}} \\ P_{off} - e^{n(\delta-1)}P_{\text{off}} & P_{\text{off}} + e^{n(\delta-1)}P_{\text{on}} \end{pmatrix} \tag{B.13}$$

resulting in

$$A_n = P_{\text{on}} + e^{n(\delta-1)}P_{\text{off}}. \tag{B.14}$$

In steady state $x_{\text{on}}(s-j) = P_{\text{on}}$ and the connected autocorrelation is:

$$< F(t)F(t+\tau) > - < F(t) >^2 = \sum_{i=1}^{r}\sum_{j=1}^{r} L_i L_j P_{\text{on}} P_{\text{off}} e^{|\tau - j + i|(\delta-1)}. \tag{B.15}$$

Since we already know the ratio of the rates from P_{on}, inferring δ using Eq. B.57 determines k_{on} and k_{off}.

B.3 Computing out of steady state

The autocorrelation approach can be generalized to a case when the system is out of steady state, when the autocorrelation function explicitly depends on the two time points and not only on their difference. During mitosis the gene is OFF and then gets turned ON in early interphase. Motivated by the hunchback expression we will present the calculation assuming the gene is

initially ON, but it is generalizable to any other initial condition. Assuming $t-i<s-j$, we want to calculate the probability that the polymerase is at position i at time t, given that it was at position j at time s. Since the gene is initially OFF, we need to calculate the probability that the gene is ON at time $t-i$. The autocorrelation function of the polymerase position is:

$$\langle a_i(t)a_j(s)\rangle = \begin{pmatrix}1 & 0\end{pmatrix} e^{(s-t+i-j)(T-\mathbb{1})} \begin{pmatrix}1\\0\end{pmatrix} \begin{pmatrix}1 & 0\end{pmatrix} e^{(t-i)(T-\mathbb{1})} \begin{pmatrix}0\\1\end{pmatrix}. \tag{B.16}$$

Using Eq. B.14 and

$$\begin{pmatrix}1 & 0\end{pmatrix} e^{n(T-\mathbb{1})} \begin{pmatrix}0\\1\end{pmatrix} = P_{\text{on}}(1-e^{n(\delta-1)}), \tag{B.17}$$

we obtain:

$$\langle F(t)F(s)\rangle = \sum_{i=1}^{r}\sum_{j=1}^{r} L_iL_jP_{\text{on}}(1-e^{(\delta-1)\min(t-i,s-j)})(P_{\text{on}}+P_{\text{off}}e^{|s-j-t+i|(\delta-1)}) \tag{B.18}$$

B.4 Multiple off states

The calculations presented in Appendix B.2 can be extended to models that include more OFF or ON states as long there are only two production states for the mRNA: one enhanced and one basal production state. The transition matrix T will then be of higher dimension and in practice should be (and has to be for dimensions larger than 3) diagonalized numerically. The exact analytical solution for the autocorrelation function is still valid written in terms of the powers of T.

B.5 Generalized multi step model

A gene with many OFF states can also be described using a reduced model with two effective gene expression states ON and OFF, where the times of transitions between these two state are not exponential but follow a long tailed distribution approximated by a Gamma distribution. The Gamma distribution describes an effective transition over many irreversible transitions between a series of OFF states:

$$\Gamma_{\alpha,\beta}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, \tag{B.19}$$

where β is the scale parameter, α is the shape parameter, and $\Gamma(\alpha)$ is the gamma function. The mean time time spent in the OFF state is $1/k_{\text{on}}^{\text{eff}} = \alpha/\beta$, so the probability for the gene to be in the ON state is:

$$P_{\text{on}} = \frac{k_{\text{on}}^{\text{eff}}}{k_{\text{on}}^{\text{eff}}+k_{\text{off}}} = \frac{1}{1+\alpha k_{\text{off}}/\beta}. \tag{B.20}$$

This model has three parameters, regardless of the number of OFF states, and using Eq. B.20 reduces the number of parameters to two, which greatly simplifies the inference. The remaining two parameters are learned from the autocorrelation function in Eq. B.5, which formally has the same form as Eq. B.7:

$$\langle F(t)F(s)\rangle = \sum_{m=1}^{r}\sum_{n=1}^{r} L_mL_nx_{\text{on}}(t-m|s-n)A_\Gamma(|s-t+mnj|), \tag{B.21}$$

but $A_\Gamma(|s-t+m-n|) = x_{\text{on}}(t-m|s-n)$ is now not memoryless. We limit our presentation to the steady state, but the calculation generalizes to out of steady systems.

We cannot solve the problem in real space, but we compute the Fourier transform of the autocorrelation function of the fluorescence signal:

$$\hat{C}(\xi) = \int_{-\infty}^{+\infty} d\tau(\langle F(t)F(t+\tau)\rangle - \langle F(t)\rangle^2)e^{-2i\pi\tau\xi}, \tag{B.22}$$

which using Eq. B.5

$$\hat{C}(\xi) = x_{\text{on}} \sum_{m,n} L_m L_n 2\Re\left[e^{-2i\pi(m-n)}\hat{A}^*_\Gamma(\xi)\right] \tag{B.23}$$

we reduce to calculating

$$\hat{A}^*_\Gamma(\xi) = \int_0^{+\infty} dt e^{-2i\pi t\xi}(A_\Gamma(t) - P_{\text{on}}). \tag{B.24}$$

We decompose $A_\Gamma(t)$ into a sum over full cycles of the gene turning from ON to OFF, with the constraint that at time t the gene is ON:

$$A_\Gamma(t) = \sum_{k=0}^{\infty} A_{\Gamma k}(t), \tag{B.25}$$

where

$$A_{\Gamma k}(t) = x_{\text{on}}(t|\text{ ON at time } s \text{ \& process has gone though k cycles}) \tag{B.26}$$

Since the first jump is from the ON to OFF, which is exponential it contributes $A_{\Gamma 0}(t) = e^{-k_{\text{off}}t}$.

First we compute an auxiliary probability distribution function of the time it takes the process to go through a full ON-OFF cycle $\eta(t)$ of taking an exponential jump out of the ON state followed by a Gamma distributed jump out of the OFF state:

$$\eta(t) = \int_0^t dx k_{\text{off}} e^{-k_{\text{off}}x} \frac{\beta^\alpha}{\Gamma(\alpha)}(t-x)^{\alpha-1}e^{-\beta(t-x)}. \tag{B.27}$$

The Fourier transform of this distribution is:

$$\hat{\eta}(\xi) = \int_0^{+\infty} dt e^{-2i\pi\xi t}\eta(t) = \frac{k_{\text{off}}}{2i\pi\xi + k_{\text{off}}} \frac{\beta^\alpha}{(2i\pi\xi+\beta)^\alpha}. \tag{B.28}$$

To compute $\hat{A}^*_\Gamma(\xi)$ we need to sum over all the possible times at which the cycles could have occurred, with the constraint that at time t the gene is ON:

$$\hat{A}^*_\Gamma(\xi) = \int_0^{+\infty} dt e^{-2i\pi\xi t}\left[\sum_{k=0}^{\infty}\left(\int_{t_i>0,\sum_{i=1}^k t_i<t} e^{-k_{\text{off}}(t-\sum_i t_i)}\prod_{i=1}^k \eta(t_i)dt_i\right) - P_{\text{on}}\right]. \tag{B.29}$$

We can rewrite the last term in Eq. B.29:

$$\hat{A}^*_\Gamma(\xi) = \int_0^{+\infty} dt e^{-2i\pi\xi t}\left[\sum_{k=0}^{\infty}\left(\int_{t_i>0,\sum_{i=1}^k t_i<t} e^{-k_{\text{off}}(t-\sum_i t_i)}\prod_{i=1}^k \eta(t_i)dt_i\right) - P_{\text{on}}\sum_{k=0}^{\infty}\right.$$
$$\left.\int_{\sum_i t_i<t}(k_{\text{off}})^k e^{-k_{\text{off}}\sum_i t_i}e^{-k_{\text{off}}(t-\sum_i t_i)}\right], \tag{B.30}$$

using the expansion of unity:

$$1 = \sum_{k=0}^{\infty} e^{-k_{\text{off}}t}\frac{(k_{\text{off}}t)^k}{k!} \tag{B.31}$$

$$= \sum_{k=0}^{\infty} \int_{\sum_i t_i < t} (k_{\text{off}})^k e^{-k_{\text{off}} \sum_i t_i} e^{-k_{\text{off}}(t - \sum_i t_i)}, \tag{B.32}$$

with the convention for the $k = 0$ term:

$$\int_{\sum_i t_i < t} (k_{\text{off}})^k e^{-k_{\text{off}} \sum_i t_i} e^{-k_{\text{off}}(t - \sum_i t_i)} = e^{-k_{\text{off}} t}. \tag{B.33}$$

Collecting terms:

$$\sum_{k=0}^{\infty} \left[\int_{t_i > 0} \prod_{i=1}^{k} dt_i \left[\left(\prod_{i=1}^{k} \eta(t_i) - P_{\text{on}} (k_{\text{off}})^k e^{-k_{\text{off}} \sum_i t_i} \right) \int_{t > \sum_i t_i} dt e^{-2i\pi\xi t} e^{-k_{\text{off}}(t - \sum_i t_i)} \right] \right] \tag{B.34}$$

and setting $u = t - \sum_i t_i$ in the last integral:

$$\hat{A}^*_{\Gamma}(\xi) = \sum_{k=0}^{\infty} \left[\int_{t_i > 0} \prod_{i=1}^{k} dt_i \left[\left(\prod_{i=1}^{k} \eta(t_i) - P_{\text{on}} (k_{\text{off}})^k e^{-k_{\text{off}} \sum_i t_i} \right) \int_0^{+\infty} du e^{-2i\pi\xi(u + \sum_i t_i)} e^{-k_{\text{off}} u} \right] \right] \tag{B.35}$$

we obtain:

$$\hat{A}^*_{\Gamma}(\xi) = (k_{\text{off}} + 2i\pi\xi - k_{\text{off}}(1 + \frac{2i\pi\xi}{\beta})^{-\alpha})^{-1} - \frac{P_{\text{on}}}{2i\pi\xi}. \tag{B.36}$$

Using Eq. B.21 we recover Eq. 7.11 in Materials and Methods. For $\alpha = 1$ we recover results of the two state model.

B.6 The autocorrelation of a Poisson polymerase firing model

We compared the auto-correlation function for our models with bursty dynamics to the auto-correlation of a model that assumes in steady state stochastic gene expression with a constant exponentially distributed rate – a Poisson polymerase firing model. We assume that the gene expression rate is memoryless and the transcription interval follows an exponential distribution of mean τ_P:

$$P(t) = \frac{1}{\tau_P} e^{-t/\tau_P}. \tag{B.37}$$

In order to compare the two models we need to reinterpret the statistics introduced for bursty dynamics in the framework of a Poisson model. The quantity P_{on} corresponds to the average occupancy of polymerase sites on the gene. This constant can be computed for a Poisson arrival model. The size of the polymerase is 150 bp and its speed is $\sim$ 25 bp/second, the maximum loading rate of polymerase is one every 6 second. Since the polymerase cannot load faster than once every 6 seconds, we calculate the average occupancy of the gene as the temporal average of probability that the polymerase starts transcribing within 6 seconds:

$$P_{\text{on}} = \int_0^6 dt \frac{1}{\tau_P} e^{-t/\tau_P} = 1 - e^{-6/\tau_P}. \tag{B.38}$$

Here we assume that the next polymerase to bind can be recruited while the previous one is clearing off the binding site. Since the process is memoryless and the Poisson firing process is uncorrelated, its connected autocorrelation is close to a delta function $\delta(\tau = 0)$. However, due to the gene lengthy elongation time, there is a non-flat auto-correlation function of the fluorescence signal. the probability of the polymerase to be at position i at time t, given it the gene to ON as

predicted from the MS2 signal at short times. At steady state, the connected auto-correlation function is:

$$\langle F(t)F(t+\tau)\rangle - \langle F(t)^2\rangle = P_{\text{on}} \sum_{i,j} L_i L_j A_P(j-\tau-i) - \left(P_{\text{on}} \sum_i L_i\right)^2, \tag{B.39}$$

where $A_P(\tau)$ is the probability of the polymerase to be at position i at time τ, given it was at position j at time 0 in the Poisson firing model.

If $\tau < 6s$ then the two positions on the gene, i and j, share the same polymerase with a probability proportional to $|6-\tau|$, taking equally distributed polymerase positions. If $\tau > 6s$, $A_P(\tau)$ is given by the probability that there is a polymerase at the second site, which is independent of what happened at the first site. The two cases give:

$$A_P(\tau) = \frac{\theta(6-|\tau|)}{6}\left[(6-|\tau|) + P_{\text{on}}|\tau|\right] + \theta(|\tau|-6)P_{\text{on}}. \tag{B.40}$$

This function is flat for $\tau > 6s$ and the first part of the right hand side of Eq. B.40 has little effect on the autocorrelation function over a cell cycle (as cell cycle duration is much bigger than $6s$). For this reason we use a flat function as a very good approximation for A_P in our analysis.

From the form of Eqs. B.39 and B.40 and the flat approximation of A_P we see that P_{on} is only a normalizing constant and the shape of the function is completely determined by the loop function L_i, which is known. We can compare the expected autocorrelation function of a Poisson model to data and find that it does not explain experimental results as well as bursty dynamics (although gene switching models have higher numbers of parameters).

From Eq. B.38 we can learn polymerase arrival rates in the anterior and at the boundary of the embryo. We find that the Poisson model would require very high heterogeneity of polymerase arrival times as a function of A-P axis position. At the boundary in particular we expect the mean polymerase arrival time to be above $60s$.

B.7 Numerical simulations

To simulate the time evolution of MCP-GFP loci's intensity, we used the Gillespie algorithm [33, 152] to predict the time it takes for the gene to switch between the states, the active ON state and the inactive OFF states. In all models we assume that the time of the transition from the active to the inactive states, τ_{on} is exponentially distributed with rate k_{off}. The time of the transition from the inactive OFF states to ON state, τ_{off} depends on the model considered:

- for the two-state model τ_{off} is exponentially distributed with rate k_{on}.
- for the three-state model τ_{off} is a sum of two exponential processes with rates k_1 and k_2 that describe the transitions between the two OFF states.
- for the Gamma model τ_{off} is chosen from a the $\Gamma(\alpha,\beta)$ distribution defined in Eq. B.19.

To generate the traces of length T from N nuclei, we first simulate a long trajectory of length $N \times T$, denoted as $X(t)$. To account for the incompressibility of the polymerase, we divide the traces into 6s intervals, which is the time the polymerase needs to cover a region of the gene equal to its own lengths. We assume that at each 6s time point, if the gene is in

the ON state, there is a transcription initiation event by a single RNA polymerase with a full transcription rate, defined as the length of the gene divided by the polymerase velocity, defined in SI section B.1. Following this event, the RNA polymerase will slide along the target gene segment and synthesize a nascent RNA. At time i into this elongation process, the nascent RNA has L_i MS2 binding sites as depicted in Fig. 3 of the main text. To impose $P_{\text{on}} = k_{\text{on}}^{\text{eff}}/(k_{\text{on}}^{\text{eff}} + k_{\text{off}})$ If the gene switches into the OFF state before a full 6s interval, the polymerase transcribes the gene at a reduced rate proportional to the fraction of the 6s interval for which the gene was ON. The number of MS2 binding sites at the transcription locus site is therefore given by the convolution of the gene state and the promoter construct design function L (see Fig. 1 in the main text):

$$F(t) = X(t) * L. \tag{B.41}$$

We assume that the number of MCP-GFP molecules in the nuclei is sufficient to bind to all newly transcribed MS2 binding sites and that the binding process is infinitely fast. The spot intensity is calculated as the number of binding sites produced at the loci (given the intensity of each MPC-GFP dimer equal to 1). Lastly, the long spot intensity traces are divided equally into N smaller traces of length T.

B.8 Correction to the autocorrelation function for finite trace lengths

The short duration of the experimental traces, $v_{\alpha,i}$, where $1 \leq \alpha \leq M$ describes the identity of the trace and $0 < i < K$ denotes the sampling times, coupled with the need to correct for experimental biases by calculating the connected correlation function introduces finite size effects. The true connected correlation function between time points at a distance r, C_r (red line in Fig. B.5), is not equal to the empirical connected correlation function calculated as an average over the M traces, $c(r)$ (blue line in Fig. B.5), of the autocorrelation functions of the finite traces. The theoretical connected autocorrelation function calculated in our model is:

$$C_r = \frac{\langle v_i v_{i+r} \rangle - \bar{v}^2}{\bar{v^2} - \bar{v}^2}, \tag{B.42}$$

where $\langle \cdot \rangle$ denotes an average over random realizations of the process and we assume steady state $\bar{v^k} = \langle v_i^k \rangle = \langle v_{i+j}^k \rangle$. The empirical connected correlation function of each finite trace of length $K << \infty$ has the form:

$$c_\alpha(r) = \left[\frac{\sum\limits_{(i,j),|i-j|=r} \left\{ \left(v_{\alpha i} - \frac{1}{N} \sum\limits_{l=1}^{N} v_{\alpha l} \right) \left(v_{\alpha j} - \frac{1}{N} \sum\limits_{l=1}^{N} v_{\alpha l} \right) \right\}}{\frac{N-r}{N} \sum\limits_{j=1}^{N} \left(v_{\alpha j} - \frac{1}{N} \sum\limits_{l=1}^{N} v_{\alpha l} \right)^2} \right] \tag{B.43}$$

and the empirical connected correlation function calculated averaged over M traces is

$$c(r) = \frac{1}{M} \sum_{\alpha=1}^{M} c_\alpha(r). \tag{B.44}$$

C_r requires knowing the true second moment of the fluorescence signal $\bar{v^2}$. In our data we find that the true variance of the normalized fluorescence signal, $\bar{v^2} - \bar{v}^2$ is well approximated by the

average over traces, so we approximate Eq. B.43 by:

$$c_\alpha(r) = \frac{\sum\limits_{(i,j),|i-j|=r} \left\{ \left(v_{\alpha i} - \frac{1}{K}\sum\limits_{l=1}^{K} v_{\alpha l} \right) \left(v_{\alpha j} - \frac{1}{K}\sum\limits_{l=1}^{K} v_{\alpha l} \right) \right\}}{\bar{v^2} - \bar{v}^2} \tag{B.45}$$

The difference between the theoretical and empirical connected correlation function is independent of our model and arises for the connected correlation function of any random process, as shown in Fig. B.5 for the simplest random process – the Ornstein-Uhlenbeck process. The difference is due to the fact that the short time average induces spurious correlations when calculating averages of the signal taken at different times. When analyzing the data, to avoid describing nucleus-to-nucleus variability that is not connected to the signal, we first subtract the mean steady state fluorescence signal of each trace, normalize this connected autocorrelation function to 1 at time $t = 0$, and then average over traces (Eq. B.45) before averaging over the trace ensemble (Eq. B.44). In steady state, the infinite trace mean equals the ensemble average, $\lim_{K\to\infty} \frac{1}{K}\sum\limits_{i=1}^{K} v_{\alpha i} = \bar{v}$. However, as shown in Fig. 2 of the main text, the short trace mean is not a good approximation to the long term (or ensemble) average, $\frac{1}{K}\sum\limits_{i=1}^{K} v_{\alpha i} \neq \bar{v}$. The points located in the center of the trace are much more correlated with the mean than the points at the beginning and end of the time interval. The correction for each value of r is different and must be separately computed.

In analyzing our data we use the finite size correction for the mean derived below that expresses the empirical connected correlation function $c(r)$ in terms of the theoretical connected correlation function C_r. For $K \to \infty$ the empirical connected correlation function becomes the infinite time connected correlation function, however our traces are very short. These corrections are valid for all time dependent data sets so for completeness the finite size correction for the variance is derived in SI Section B.9 but is not used in the analysis.

The number of pairs of time points of distance r in a trace of length N is simply $N - r$ and the combination of Eqs. B.44 and Eqs. B.45 becomes:

$$\begin{aligned}
c(r) &= \frac{1}{M(N-r)(\bar{v^2} - \bar{v}^2)} \sum_{\alpha=1}^{M} \left[\sum_{i=1}^{N-r} \left\{ \left(v_{\alpha i} - \frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right) \left(v_{\alpha(i+r)} - \frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right) \right\} \right] \\
&= \frac{1}{M(N-r)(\bar{v^2} - \bar{v}^2)} \sum_{\alpha=1}^{M} \Big[\sum_{i=1}^{N-r} \Big\{ v_{\alpha i} v_{\alpha(i+r)} - v_{\alpha i} \left(\frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right) - \left(\frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right) v_{\alpha(i+r)} + \\
&\quad (N-r) \left(\frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right)^2 \Big\} \Big] \\
&= \left\langle \frac{1}{(\bar{v^2} - \bar{v}^2)} \left\{ \sum_{i=1}^{N-r} \frac{v_{\alpha i} v_{\alpha(i+r)}}{N-r} - \sum_{i=1}^{N-r} \frac{v_{\alpha i}}{N-r} \left(\frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right) - \sum_{i=r+1}^{N} \frac{v_{\alpha i}}{N-r} \left(\frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right) + \frac{1}{N^2} \left(\sum_{l=1}^{N} v_{\alpha l} \right)^2 \right\} \right\rangle_\alpha
\end{aligned} \tag{B.46}$$

where we have explicitly written out the terms and in the last line we introduced the average over traces $\langle \cdot \rangle_\alpha = 1/M \sum_{\alpha=1}^{M} \cdot$. In steady state due to time invariance:

$$\left\langle \sum_{i=N-r+1}^{N} \frac{v_{\alpha i}}{N-r} \left(\frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right) \right\rangle_\alpha = \left\langle \sum_{i=1}^{r} \frac{v_{\alpha i}}{N-r} \left(\frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right) \right\rangle_\alpha \tag{B.47}$$

and the theoretical (not connected) correlation between two points is a function only of the distance between these two points:

$$\tilde{C}_r = \langle v_i v_{i+r} \rangle = 1/M \sum_{\alpha=1}^{M} v_{\alpha i} v_{\alpha i+r}. \tag{B.48}$$

We have assumed that M is large and a population average over the M traces for points separated by r on each trace approximates the $M \to \infty$ limit of the theoretical average over different realizations of the process. Using Eq. B.48 we obtain:

$$c(r) = \frac{\tilde{C}_r}{\bar{v^2} - \bar{v}^2} + \frac{1}{(\bar{v^2} - \bar{v}^2)} \left\langle \sum_{i=1}^{r} \frac{2 v_{\alpha i}}{N-r} \left(\frac{1}{N} \sum_{l=1}^{N} v_{\alpha l} \right) + \frac{1}{N} \left(\frac{1}{N} - \frac{2}{(N-r)} \right) \left(\sum_{l=1}^{N} v_{\alpha l} \right)^2 \right\rangle_{\alpha}. \tag{B.49}$$

To rewrite $\left\langle \left(\sum_{l=1}^{N} v_{\alpha l} \right)^2 \right\rangle_{\alpha}$ as a sum over C_r we calculate the number of pairs of time points separated by a distance k in the whole trace of length N. For $k = 0$ it is equal to N and for $1 \leq k \leq N-1$ it is equal to $2(N-k)$:

$$\left\langle \left(\sum_{l=1}^{N} v_{\alpha l} \right)^2 \right\rangle_{\alpha} = N\tilde{C}_0 + \sum_{k=1}^{N-1} 2(N-k)\tilde{C}_k. \tag{B.50}$$

Similarly

$$\left\langle \sum_{i=1}^{r} \sum_{l=1}^{N} v_{\alpha i} v_{\alpha l} \right\rangle_{\alpha} = \left\langle \left(\sum_{i=1}^{r} v_{\alpha i} \right)^2 \right\rangle_{\alpha} + \left\langle \sum_{i=1}^{r} \sum_{l=r+1}^{N} v_{\alpha i} v_{\alpha l} \right\rangle_{\alpha} \tag{B.51}$$

$$= r\tilde{C}_0 + \sum_{k=1}^{r-1} 2(r-k)\tilde{C}_k + \sum_{l=r+1}^{N} \sum_{i=1}^{r} \tilde{C}_{|l-i|} \tag{B.52}$$

$$= r\tilde{C}_0 + \sum_{k=1}^{r-1} 2(r-k)\tilde{C}_k + \sum_{m=1}^{N-1} \tilde{C}_m [\min(m+r, N) - \max(r, m)] \tag{B.53}$$

Collecting the empirical connected autocorrelation function in Eq. B.44 is expressed in terms of the theoretical non-connected correlation function in Eq B.48 as:

$$c(r) = \frac{1}{\bar{v^2} - \bar{v}^2} \left[\tilde{C}_r + \frac{1}{N} \left(\frac{1}{N} - \frac{2}{(N-r)} \right) \left(N\tilde{C}_0 + \sum_{k=1}^{N-1} 2(N-k)\tilde{C}_k \right) \right.$$
$$\left. + \frac{2}{N(N-r)} \left(r\tilde{C}_0 + \sum_{k=1}^{r-1} 2(r-k)\tilde{C}_k + \sum_{m=1}^{N-1} \tilde{C}_m [\min(m+r, N) - \max(r, m)] \right) \right]. \tag{B.54}$$

B.9 Correction to the autocorrelation function from correlations in the variance

In SI Section B.8 we calculated the finite size correction due to short traces for the empirical connected correlation function assuming that differences between he empirical variance and the theoretical variance for infinite traces do not affect the connected autocorrelation function. This

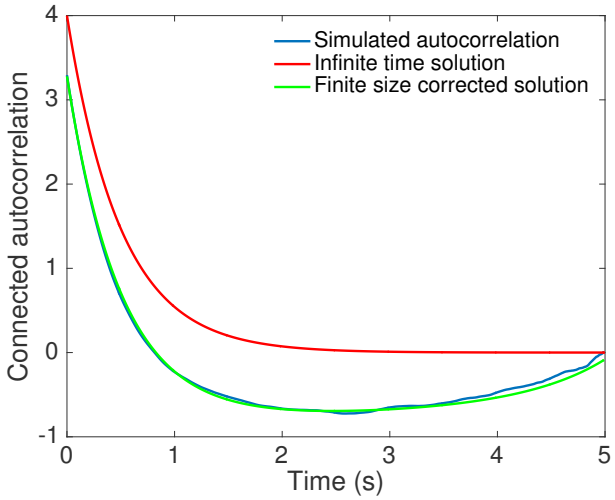

Figure B.5: **The finite trace effect for the Ornstein-Uhlenbeck process.** The connected autocorrelation function $C_r = \exp(-t/\tau)$ (red line) compared to the connected autocorrelation function calculated from short time traces as described in SI Section B.8 (blue line) and the corrected connected autocorrelation function (Eq. B.54 green line). $\lambda = 2\text{s}^{-1}$, $\gamma = 4\text{s}^{-1/2}$ and the short trace length is 5s where the Ornstein-Ulhenbeck process is $\partial_t x = -\lambda x + \gamma\xi$ and ξ is Gaussian white noise.

approximation is valid for our data. For completeness we now calculate the finite size correction coming from spurious correlations in the variance obtained when computing the variance trace by trace, before averaging over the traces (Eq. B.44). Analyzing the data, we normalize the autocorrelation function of each trace before taking the average over all traces because of potential nucleus-to-nucleus variability in the signal calibration. This is equivalent to dividing each autocorrelation function by its variance, before averaging over the traces and can introduce errors.

The empirical connected correlation function in Eqs. B.44 and B.43 can be rewritten by adding and subtracting 1 in the denominator as:

$$c(r) = \frac{N}{(N-r)(\bar{v^2}-\bar{v}^2)}\Big\langle \left[\frac{\sum_{(i,j),|i-j|=r} \left\{ \left(v_{\alpha i} - \frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right) \left(v_{\alpha j} - \frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right) \right\}}{1 + \frac{1}{\bar{v^2}-\bar{v}^2} \left(\sum_{j=1}^{N} \left(v_{\alpha j} - \frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right)^2 - (\bar{v^2}-\bar{v}^2) \right)} \right] \Big\rangle_\alpha,$$

where the average $\langle\cdot\rangle_\alpha$ is over M traces as defined in SI Section B.8. Assuming the true variance of the process is close to the empirical variance we linearize the denominator :

$$c(r) = \frac{N}{(N-r)(\bar{v^2}-\bar{v}^2)}\Big\langle \left[\sum_{(i,j),|i-j|=r} \left\{ \left(v_{\alpha i} - \frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right) \left(v_{\alpha j} - \frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right) \right\} \right] \times \left[2 - \frac{1}{\bar{v^2}-\bar{v}^2} \sum_{j=1}^{N} \left(v_{\alpha j} - \frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right)^2 \right] \Big\rangle_\alpha. \quad \text{(B.55)}$$

We first term is proportional to the connected correlation function in Eq. B.54 we calculated in

SI Section B.8 assuming constant variance. We focus on the second term:

$$d(r) = \Big\langle \left[\sum_{(i,j),|i-j|=r} \left\{ \left(v_{\alpha i} - \frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right) \left(v_{\alpha j} - \frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right) \right\} \right] \cdot \left[\sum_{j=1}^{N} \left(v_{\alpha j} - \frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right)^2 \right] \Big\rangle_\alpha$$
$$= \Big\langle \left\{ \sum_{i=1}^{N-r} v_{\alpha i} v_{\alpha(i+r)} - \sum_{i=1}^{N-r} v_{\alpha i} \left(\frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right) - \sum_{i=r+1}^{N} v_{\alpha i} \left(\frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right) + \frac{N-r}{N^2} \left(\sum_{l=1}^{N} v_{\alpha l} \right)^2 \right\} \times$$
$$\left[\sum_{j=1}^{N} v_{\alpha j}^2 - \frac{2}{N}\sum_{j=1}^{N}\sum_{l=1}^{N} v_{\alpha j} v_{\alpha l} + \frac{N}{N^2}\sum_{j=1}^{N}\sum_{l=1}^{N} v_{\alpha j} v_{\alpha l} \right] \Big\rangle_\alpha .$$

Using time invariance at steady state (Eq. B.47) in the first factor and simplifying the algebra in the second factor:

$$d(r) = \Big\langle \left[\sum_{i=1}^{N-r} v_{\alpha i} v_{\alpha(i+r)} - 2\sum_{i=1}^{N-r} v_{\alpha i} \left(\frac{1}{N}\sum_{l=1}^{N} v_{\alpha l} \right) + \frac{N-r}{N^2} \left(\sum_{l=1}^{N} v_{\alpha l} \right)^2 \right] \cdot \left[\sum_{j=1}^{N} v_{\alpha j}^2 - \frac{1}{N}\sum_{j=1}^{N}\sum_{l=1}^{N} v_{\alpha j} v_{\alpha l} \right] \Big\rangle_\alpha$$
$$= \Big\langle \sum_{j=1}^{N}\sum_{i=1}^{N-r} v_{\alpha j}^2 v_{\alpha i} v_{\alpha(i+r)} - \frac{1}{N}\sum_{i=1}^{N-r}\sum_{j,l=1}^{N} v_{\alpha j} v_{\alpha l} v_{\alpha i} v_{\alpha(i+r)} - \frac{2}{N}\sum_{i=1}^{N-r}\sum_{j,l=1}^{N} v_{\alpha i} v_{\alpha l} v_{\alpha j}^2 + \frac{2}{N^2}\sum_{i=1}^{N-r}\sum_{j,k,l=1}^{N} v_{\alpha i} v_{\alpha j} v_{\alpha k} v_{\alpha l}$$
$$+ \frac{N-r}{N^2}\sum_{j,k,l=1}^{N} v_{\alpha j}^2 v_{\alpha k} v_{\alpha l} - \frac{N-r}{N^3}\sum_{j,k,l,m=1}^{N} v_{\alpha j} v_{\alpha k} v_{\alpha l} v_{\alpha m} \Big\rangle_\alpha .$$

The final correction for correlation due to correlations in the variance coming from short time traces is easily evaluated terms of four-points correlation function $F(s,t,u) = \overline{v_i v_{i+s} v_{i+s+t} v_{i+s+t+u}}$.

B.10 Cross-correlation

The presented correlation analysis can also be extended to constructs with two colored promoters inserted at two difference positions on the same gene. In this case, each construct can have a different loop design function L_i^ν, where $\nu = 1, 2$, and the cross-correlation of the normalized fluorescence intensity is:

$$\langle F_1(t) F_2(s) \rangle = \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} L_i^1 L_j^2 < a_i(t) a_j(s) > . \tag{B.56}$$

The L_i^ν functions start at the same point (the one describing the downstream construct is 0 for the first steps).

After the loop design functions L_i^ν have been defined, the calculation of the theoretical cross-correlation function and auto-correlation rely only on calculating the correlations of the gene expression state, which is the same for both. So the results presented for the particular models are valid, after correcting for the two different loops functions. For examples, the steady state connected cross-correlation function of the two state model is:

$$\langle F_1(t) F_2(t+\tau) \rangle - \langle F_1(t) \rangle^2 = \sum_{i=1}^{r}\sum_{j=1}^{r} L_i L_j P_{\text{on}} P_{\text{off}} e^{|\tau - j + i|(\delta - 1)}, \tag{B.57}$$

where P_{on} and $\langle F_1(t) \rangle^2 = \langle F_2(t) \rangle^2$ can be independently calculated from either probe, which provides an independent estimate of the experimental noise.

The differences in the use of the cross-correlation function and auto-correlation function arise when calculating the finite size corrections from short traces, because assumptions about the statistical time invariance of the signal in steady state are no longer valid. The non-connected theoretical correlation function (equivalent of Eq. B.48) is now defined on two signals, v_i and w_i:

$$\tilde{C}_r = \langle v_{\alpha,i} w_{\alpha,i+m} \rangle, \tag{B.58}$$

where $\langle \cdot \rangle$ define the average over random realizations of the process and in steady state is independent of i. Unlike for the auto-correlation function, $\tilde{C}_r$ is no longer symmetric with exchange of v_i and w_i. The empirical cross-correlation function is (assuming the variance is well approximated by the empirical variance):

$$c(r) = \left\langle \frac{1}{(\bar{v^2} - \bar{v}^2)} \left\{ \sum_{i=1}^{N-r} \frac{v_{\alpha i} w_{\alpha(i+r)}}{N-r} - \sum_{i=1}^{N-r} \frac{v_{\alpha i}}{N-r} \left(\frac{1}{N} \sum_{l=1}^{N} w_{\alpha l} \right) \right. \right. \\ \left. \left. - \sum_{i=r+1}^{N} \frac{w_{\alpha i}}{N-r} \left(\frac{1}{N} \sum_{l=1}^{N} v_{\alpha l} \right) + \frac{1}{N^2} \left(\sum_{l=1}^{N} v_{\alpha l} \right) \left(\sum_{l=1}^{N} w_{\alpha l} \right) \right\} \right\rangle_{\alpha}, \tag{B.59}$$

which in terms of the $\tilde{C}_m$ is:

$$c(r) = \frac{1}{(\bar{v^2} - \bar{v}^2)} \left\{ \tilde{C}_r - \frac{1}{N(N-r)} \sum_{i=1}^{N-r} \sum_{l=1}^{N} \tilde{C}_{l-i} - \frac{1}{N(N-r)} \sum_{i=r+1}^{N} \sum_{l=1}^{N} \tilde{C}_{i-l} + \frac{1}{N^2} \sum_{i,l=1}^{N} \tilde{C}_{i-l} \right\}. \tag{B.60}$$

Repeating the steps in SI Section B.8 we obtain the finite size correction for the cross-correlation function.

$$c(r) = \frac{1}{(\bar{v^2} - \bar{v}^2)} \left\{ \tilde{C}_r - \frac{1}{N(N-r)} \sum_{k=-N+r+1}^{N-r-1} (N - r - |k|) \tilde{C}_k \right. \\ - \frac{1}{N(N-r)} \sum_{i=1}^{N-r} \sum_{l=N-r+1}^{N} \tilde{C}_{l-i} - \frac{1}{N(N-r)} \sum_{k=-N+r+1}^{N-r-1} (N - r - |k|) \tilde{C}_k \\ \left. - \frac{1}{N(N-r)} \sum_{i=r+1}^{N} \sum_{l=1}^{r} \tilde{C}_{i-l} + \frac{1}{N^2} \sum_{k=-N+1}^{N-1} (N - |k|) \tilde{C}_k \right\}$$

$$= \frac{1}{(\bar{v^2} - \bar{v}^2)} \left\{ \tilde{C}_r - \frac{2}{N(N-r)} \sum_{k=-N+r+1}^{N-r-1} (N - r - |k|) \tilde{C}_k - \frac{1}{N(N-r)} \sum_{m=1}^{N-1} \tilde{C}_m [\min(m + N - r, N) - \max(N - r, \right. \\ \left. - \frac{1}{N(N-r)} \sum_{m=1}^{N-1} \tilde{C}_m [\min(m + r, N) - \max(r, m)] + \frac{1}{N^2} \sum_{k=-N+1}^{N-1} (N - |k|) \tilde{C}_k \right\}.$$

B.11 Precision of the translational process

The precision of the total mRNA produced during a cell cycle presented in the main text is proportional to the activity of the gene and requires a careful calculation of the variability of the probability of the gene to be ON in different nuclei at the same position. The total activity of a nucleus, defined as the integral of the normalized fluoresce $\sum_i^K F_i$, where $i < K$

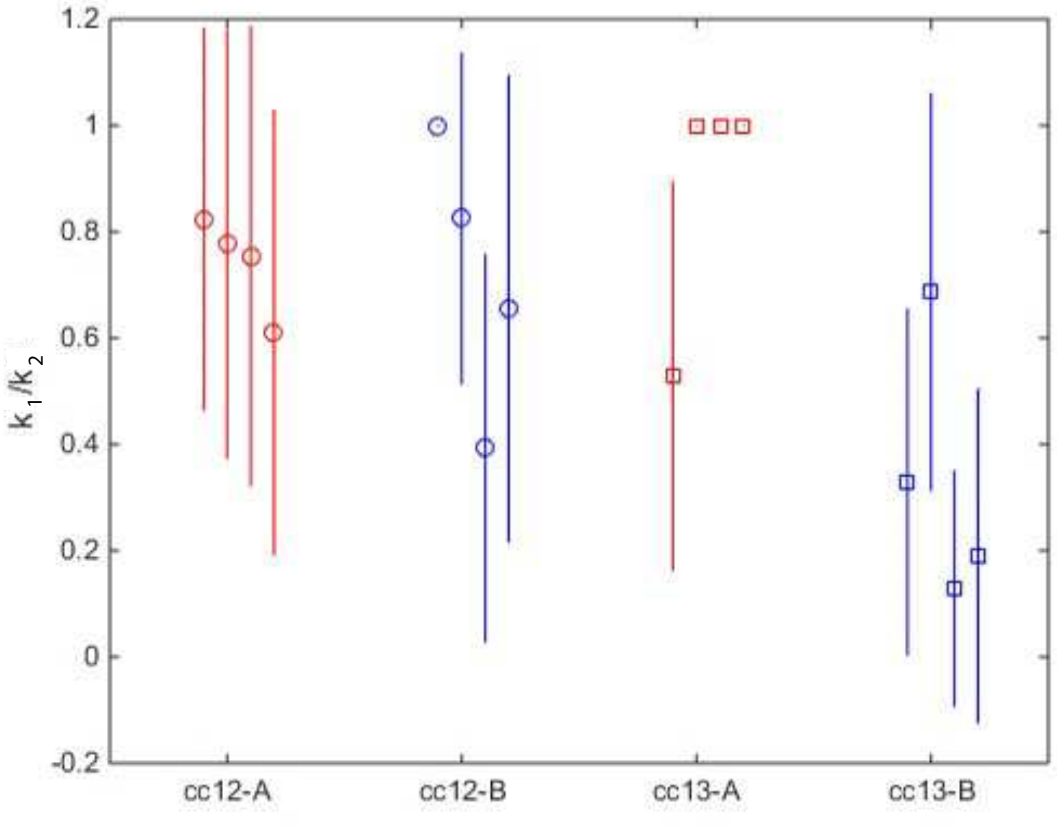

Figure B.6: **The fit of the three state cycle model to the data** The fit of the ratio of the two rates for leaving the two OFF states k_1/k_2 to the steady state traces from four embryos in the anterior and boundary region of cell cycle 12 and 13. Each point is data from one embryo. The error bar represent the standard deviation of the inferred value. The fit is for a randomized 60% of the data. The sum of the switching rates $k_{\text{on}} + k_1 + k_2$ is shown in Fig. 5B of the main text.

are the sampling times in steady state window of the cycle, in steady state is proportional to the probability of the gene to be ON in a given trace, P^{α}_{on}. To keep our analysis independent of normalization, we will calculate the relative error defined as the variance over the mean of P^{α}_{on}, $\text{var}(P^{\alpha}_{\text{on}})/\langle P^{\alpha}_{\text{on}}\rangle_{\alpha}$, where the averages are taken over traces.

First, we can calculate the relative error of the probability of the gene to be ON P^{α}_{on} directly from the traces. We compute the mean and standard deviation of the distribution of P^{α}_{on} in a given window along the AP axis. P^{α}_{on} for each trace is calculated from Eq. B.4.

We can compare the results of the empirically estimated relative error to predictions of the steady state models. We know that the expected average over traces $\sum_{alpha=1}^{M} P^{\alpha}_{\text{on}}$ is P_{on}. Within the assumption of our model presented in SI Section B.2, the expectation value of the square of the P^{α}_{on} is expressed in terms of the expression states of the gene, $X(t)$:

$$\langle P^{\alpha,2}_{\text{on}}\rangle_{\alpha} = \left\langle \frac{1}{T^2}\int_0^T dt \int_0^T ds X(t)X(s) \right\rangle_{\alpha}, \tag{B.61}$$

where the average is over M traces and T is the total duration of the trace in real time. In terms

of the probability that the gene is ON at time τ given that it was ON at time 0, $A(\tau)$ defined in Eq. B.6, we obtain

$$\langle P_{\rm on}^{\alpha,2}\rangle_\alpha = \left\langle \frac{1}{T^2}\int_0^T dt \int_0^T ds P_{\rm on} A(t-s) \right\rangle, \tag{B.62}$$

where $A(\tau)$ has units of seconds. The relative error is obtained by replacing $A(\tau)$ by the appropriate function for each model. For the two state model:

$$\langle P_{\rm on}^{\alpha,2}\rangle_\alpha = \frac{P_{\rm on}}{T^2}\int_0^T dt \int_0^T ds (P_{\rm on} + P_{\rm off} e^{-|t-s|(k_{\rm on}+k_{\rm off})}). \tag{B.63}$$

Integrating and substracting the mean squared we obtain the relative error:

$$\frac{\delta P_{\rm on}}{P_{\rm on}} = \frac{1}{T}\sqrt{2\frac{k_{\rm off}}{k_{\rm on}(k_{\rm on}+k_{\rm off})}(T - \frac{1-e^{-T(k_{\rm on}+k_{\rm off})}}{k_{\rm on}+k_{\rm off}})}. \tag{B.64}$$

The probability of the gene to be on is proportional to the total mRNA produced and for large T we reproduce the result in Eq. 4 in the main text:

$$\frac{\delta{\rm mRNA}}{\rm mRNA} = \sqrt{\frac{2}{T}\frac{k_{\rm off}}{k_{\rm on}(k_{\rm on}+k_{\rm off})}} = \sqrt{2\frac{\tau_i(1-P_{\rm on})}{TP_{\rm on}}}. \tag{B.65}$$

For the three state cycle model the same calculation is valid until Eq. B.62 and is then carried out numerically.

Precision from static (Fluorescent In Situ Hybridization – FISH) images is calculated as the variance over the mean of the distribution of a binary variable, which for each nucleus is 1 if the gene is on in the static image and 0 if it off [139, 120, 125]. The signal in FISH datasets in an average over an unknown timeframe. To compare our analysis of the time dependent signal to these previous measurements, we use a binary variable, which is 1 for each nucleus that was ON during the steady state interphase and 0 for each nucleus that was always OFF. The results of the relative error as a function of position obtained using this empirical analysis in SIFig. B.7 show agreement with previous reports [125]: for most traces the relative error in the anterior is zero – all nuclei in a given AP axis window express, and it increases to $\sim 50\%$ at the boundary.

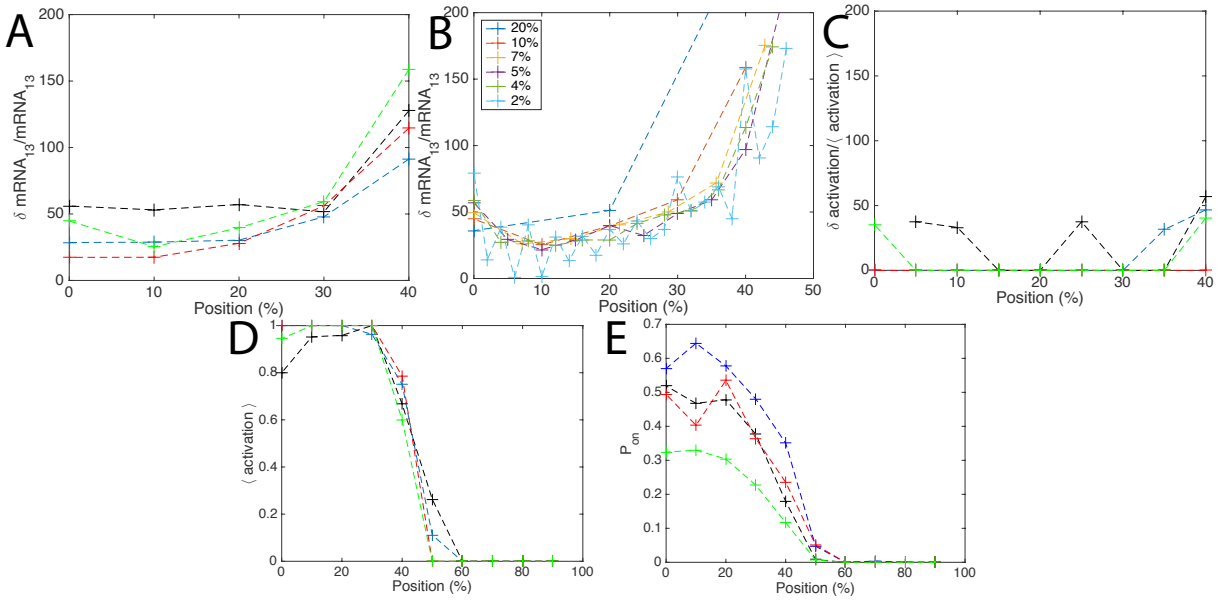

Figure B.7: **The relative error of gene expression.** A. The conclusions about precision do not depend on the embryo. The relative error of the total mRNA produced in cell cycle 13 as a function of position for windows equal to 10% of the embyo length. Each colored line represents one embryo. The same data plotted as an average over embryos with the variance as error bars is shown in Fig. 7 of the main text. B. The conclusions about precision do not depend on the window size. The total mRNA produced in cell cycle 13 as a function of position for different window sizes. Except for very large scales (20%) and very small scales comparable to one nuclear width (2%, the relative error as a function of position is reproducible. C.The relative error of the discrete variable that describes the probability of the gene to be ON at any time during the cell cycle as function of position. The relative error is much lower in the anterior compared to the error in the total produced mRNA, but remains high at the boundary. D. The mean probability of the gene to be ON at any time during the cell cycle as a function of the embryo length (binary approximation). E. The mean probability for the gene to be ON averaged over the cell cycle. In C-E each colored lines describe different embryos.

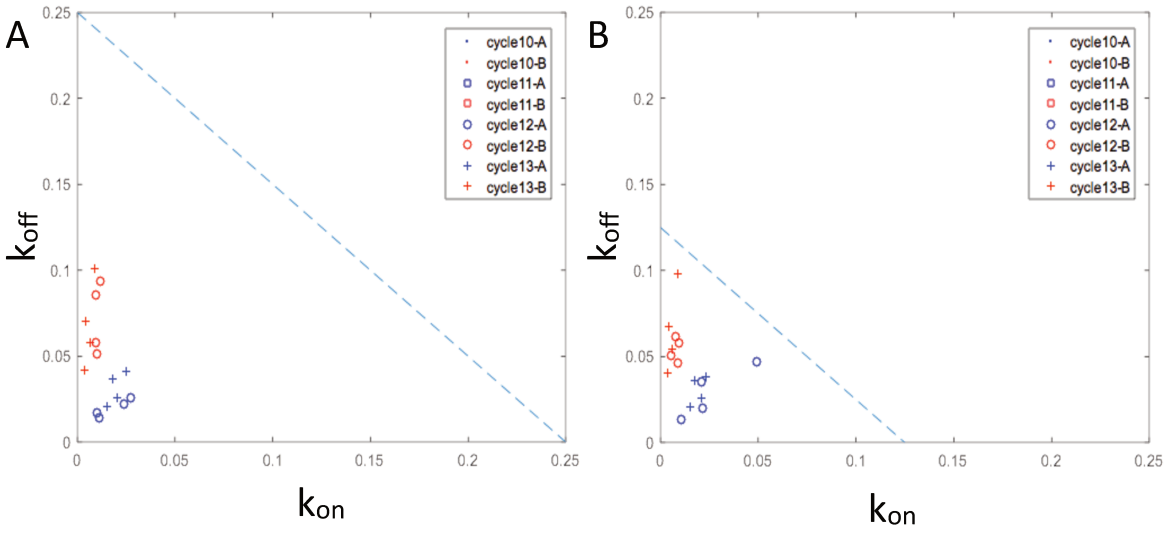

Figure B.8: **The dependence of the data fit on polymerase buffering time.** Assuming different buffering times for the polymerase does not strongly affect the fit of the switching rates: a fit with $\tau_{\text{buffering}} = 4s$ (A) and $\tau_{\text{buffering}} = 8s$. $\tau_{\text{buffering}} = 6s$ is used in the main text in Fig. 5D.

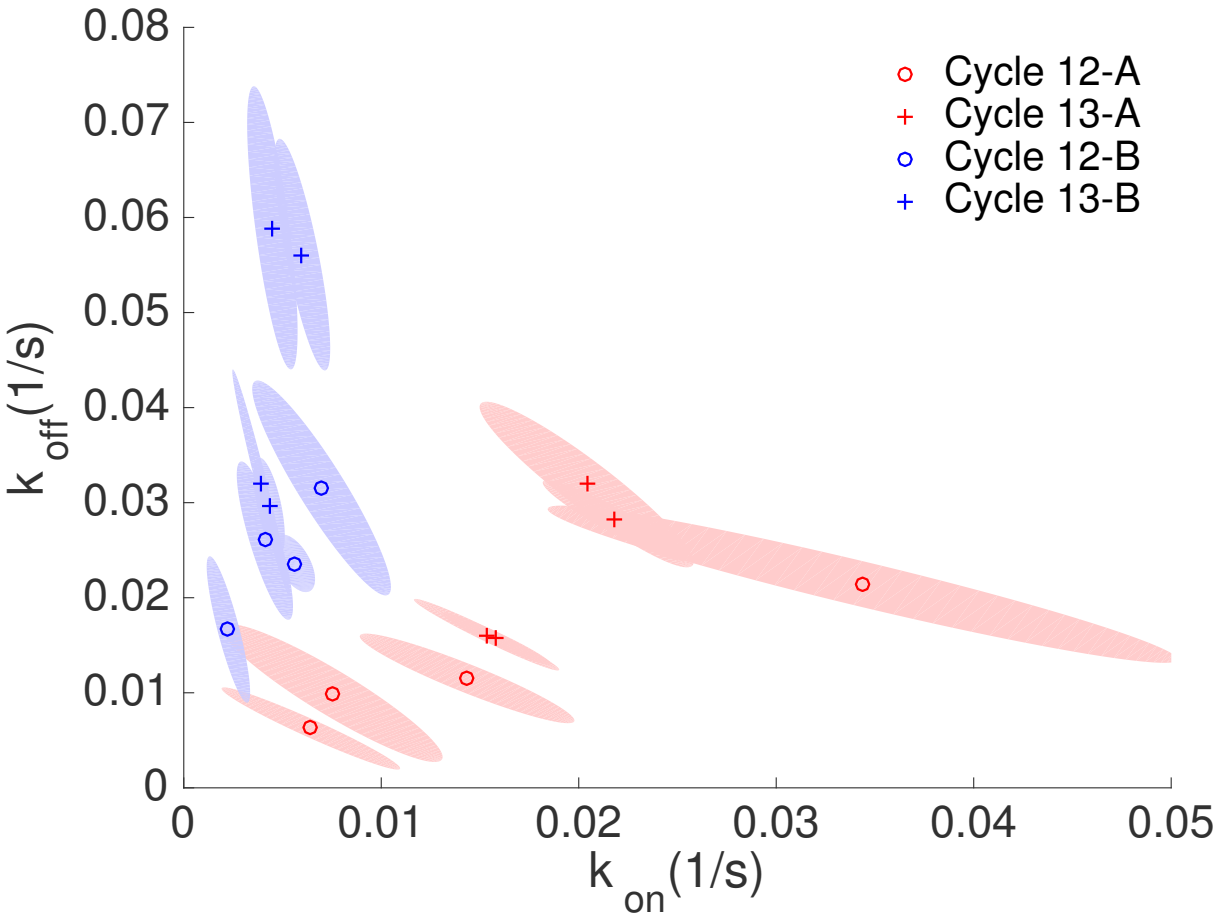

Figure B.9: **Fit of switching parameters with variability for the two-state model.** This graph presents the result of the two-state model data analysis in terms of gene switching rates with error zones.

Bibliography

[1] *The miscellaneous botanical works of Robert Brown: Volume 1.* R. Hardwicke, London., 1866. 11

[2] C. W. Gardiner. *Handbook of stochastic methods for physics, chemistry and the natural sciences*, volume 13 of *Springer Series in Synergetics*. Springer-Verlag, Berlin, third edition, 2004. 11, 18

[3] A. Einstein. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, 322(8):549–560, 1905. 11

[4] M. von Smoluchowski. Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen. *Annalen der Physik*, 1906. 11

[5] A. J. Lotka. *Elements of Physical Biology*. Williams Wilkins Co., Baltimore, 1925. 11

[6] V. Volterra. Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Mem. R. Accad. Naz. dei Lincei*, 2, 1926. 11

[7] Sewall Wright. Statistical genetics and evolution. *Bulletin of the American Mathematical Society*, 8(4):223–246, 1942. 11

[8] R A Fisher. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Royal Society of Edinburgh*, 1918. 11

[9] Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1983. Cambridge Books Online. 11, 53

[10] M Kimura. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, 1977. 11

[11] H C Berg and E M Purcell. Physics of chemoreception. *Biophysical journal*, 20(2):193–219, November 1977. 11, 12, 14, 72, 93

[12] Albert-lászló Barabási and Zoltán N Oltvai. Network Biology: understanding the cell's functional organization. *Nature reviews, Genetics*, 5(February), 2004. 12

[13] Thomas G. Kurtz David F. Anderson. *Continuous time Markov chain models for chemical reaction networks, Design and Analysis of Biomolecular Circuits*. Springer New York, 2011. 12

[14] Alan S Perelson and George F Oster. Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *Journal of theoretical biology*, 81(4):645–670, 1979. 12, 60

[15] Alan S. Perelson and Patrick W. Nelson. Mathematical Analysis of HIV-I : Dynamics in Vivo. *SIAM review*, 41(1):3–44, 2009. 12

[16] C Janeway. *Immunobiology*. Garland Science, 2005. 12, 33, 42

[17] Jonathan Desponds, Thierry Mora, and Aleksandra M. Walczak. Fluctuating fitness shapes the clone size distribution of immune repertoires. *Proceedings of the National Academy of Sciences*, 113(2):274–279, 2016. 13, 14, 41

[18] Tanguy Lucas, Teresa Ferraro, Baptiste Roelens, Jose De Las Heras Chanes, Aleksandra M Walczak, Mathieu Coppey, and Nathalie Dostatni. Live imaging of bicoid-dependent transcription in Drosophila embryos. *Current Biology*, 23(21):2135–9, 2013. 14, 73, 80, 82, 95, 99

[19] Hernan G Garcia, Mikhail Tikhonov, Albert Lin, and Thomas Gregor. Quantitative imaging of transcription in living Drosophila embryos links polymerase activity to patterning. *Current Biology*, 23(21):2140–5, 2013. 14, 73, 77, 80, 82, 85, 88, 95, 96

[20] Cheryll Tickle Lewis Wolpert and Alfonso Martinez Arias. *Principles of development Fifth Edition*. Oxford University Press, 2015. 14

[21] C. H. Waddington. *The strategy of the genes. A discussion of some aspects of theoretical biology*. London: George Allen Unwin, Ltd., 1957. 14

[22] Aude Porcher and Nathalie Dostatni. The Bicoid morphogen system. *Current Biology*, 20(5):249–254, 2010. 14, 95

[23] Jonathan Desponds, Huy Tran, Teresa Ferraro, Tanguy Lucas, Carmina Perez Romero, Aurelien Guillou, Cecile Fradin, Mathieu Coppey, Nathalie Dostatni, and Aleksandra M Walczak. Precision of readout at the hunchback gene. *bioRxiv*, 2016. 14, 15, 79

[24] NG Van Kampen. *Stochastic processes in physics and chemistry*. North Holland, 2007. 17, 19

[25] N. G. van Kampen. Itô versus Stratonovich. *Journal of Statistical Physics*, 24(1):175–187, 1981. 23

[26] H. C. Tuckwell. Study of some diffusion models of population growth. *Theoretical Population Biology*, 1974. 23

[27] Riccardo Mannella and Peter V. E. McClintock. Itô versus Stratonovich: 30 years later. *Fluct. Noise Lett.*, 11(1):1–10, 2012. 23

[28] Christopher Jarzynski. Rare events and the convergence of exponentially averaged work values. *Phys. Rev. E*, 73:046105, Apr 2006. 24

[29] D Andrieux, P Gaspard, S Ciliberto, N Garnier, S Joubaud, and A Petrosyan. Entropy production and time asymmetry in nonequilibrium fluctuations. *Physical Review Letters*, 98(15), 2007. 24

[30] Armita Nourmohammad, Jakub Otwinowski, and Joshua B Plotkin. Host-Pathogen Coevolution and the Emergence of Broadly Neutralizing Antibodies in Chronic Infections. *PloS Genet*, 12(7), 2016. 24

[31] F Mancini, M Marsili, and A M Walczak. Trade-offs in delayed information transmission in biochemical networks . *Journal of Statistical Physics*, 2015. 24

[32] 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. 26

[33] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977. 26, 138

[34] Daniel T Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, 115(4):1716–1733, 2001. 27

[35] N A Sinitsyn, Nicolas Hengartner, and Ilya Nemenman. Adiabatic coarse-graining and simulations of stochastic biochemical networks. *Proceedings of the National Academy of Sciences*, 2009. 27

[36] Ord K. Stuart A. *Kendall's Advanced Theory of Statistics: Volume I—Distribution Theory*. Edward Arnold, 1994. 30

[37] Alan S Perelson and Gérard Weisbuch. Immunology for physicists. *Reviews of modern physics*, 69(4):1219–1267, 1997. 33

[38] Iren Bains, Rustom Antia, Robin Callard, and Andrew J Yates. Quantifying the development of the peripheral naive CD4 + T-cell pool in humans. *Blood*, 113(22):5480–5487, 2009. 34, 45, 64, 124

[39] Yuval Elhanati, Anand Murugan, Curtis G. Callan Jr, Thierry Mora, and Aleksandra M. Walczak. Quantifying selection in immune receptor repertoires. *Proceedings of the National Academy of Sciences*, 111(27):9875–9880, 2014. 35, 38, 55

[40] Anand Murugan, Thierry Mora, Aleksandra M Walczak, and Curtis G Callan. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences*, 109(40):16161–16166, 2012. 35, 45

[41] Westera et al. Closing the gap between T-cell life span estimates from stable isotope-labeling studies in mice and humans. *Blood*, 122(13):2205–2212, 2013. 36

[42] Rob J. De Boer and Alan S. Perelson. Quantifying T lymphocyte turnover. *J. Theor. Biol.*, 327:45–87, June 2013. 36, 50

[43] AR McLean, MM Rosado, Fabien Agenes, R Vasconcellos, and Antonio A Freitas. Resource competition as a mechanism for B cell homeostasis. *Proceedings of the National Academy of Sciences*, 94(11):5792–5797, 1997. 36, 56

[44] Andreas Mayer, Thierry Mora, Olivier Rivoire, and Aleksandra M Walczak. Diversity of immune strategies explained by adaptation to pathogen statistics. *Proceedings of the National Academy of Sciences*, 2016. 36

[45] Frederik W Wiegel and Alan S Perelson. Some Scaling Principles for the Immune System. *Immunology and Cell Biology*, 82:127–131, 2004. 36

[46] Joshua A Weinstein, Ning Jiang, Richard A White, Daniel S Fisher, and Stephen R Quake. High-throughput sequencing of the zebrafish antibody repertoire. *Science*, 324(5928):807–810, 2009. 36, 37, 42, 44, 47, 52, 113, 114

[47] Mikhail V et al Pogorelyy. Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. *arXiv preprint*. 37

[48] Wilfred Ndifon et al. Chromatin conformation governs T-cell receptor Jβ gene segment usage. *Proceedings of the National Academy of Sciences*, 109(39):15865–70, Sep 2012. 37, 42

[49] Niclas Thomas et al. Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics*, 30(22):3181–8, aug 2014. 37, 42

[50] Veronika I. Zarnitsyna, Brian D. Evavold, Louis N. Schoettle, Joseph N. Blattman, and Rustom Antia. Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. *Front. Immunol.*, 4(DEC):485, January 2013. 37, 38, 42, 44

[51] René L. Warren, J. Douglas Freeman, Thomas Zeng, Gina Choe, Sarah Munro, Richard Moore, John R. Webb, and Robert A. Holt. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Research*, 21(5):790–797, 2011. 37, 42

[52] Evgeny S Egorov et al. Quantitative Profiling of Immune Repertoires for Minor Lymphocyte Counts Using Unique Molecular Identifiers. *J Immunol*, 194(12):6155–63, 2016. 37

[53] Thierry Mora and William Bialek. Are Biological Systems Poised at Criticality ? *J Stat Phys*, 144(May):268–302, 2011. 38

[54] David J Schwab, Ilya Nemenman, and Mehta Pankaj. Zipf ' s Law and Criticality in Multivariate Data without Fine-Tuning. *Physical Review Letters*, 113, 2014. 38

[55] Yuval Elhanati, Quentin Marcou, Thierry Mora, and Aleksandra M Walczak. repgenHMM : a dynamic programming tool to infer the rules of immune receptor generation from sequence data. *Bioinformatics*, 32(13):1943–1951, 2016. 38

[56] Rhys M Adams, Justin B Kinney, Thierry Mora, and Aleksandra M Walczak. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *bioRxiv*, 2016. 38

[57] R J De Boer, a a Freitas, and a S Perelson. Resource competition determines selection of B cell repertoires. *Journal of theoretical biology*, 212(3):333–343, 2001. 39, 59

[58] Rob J De Boer and Alan S Perelson. Competitive control of the self-renewing T cell repertoire. *International immunology*, 9(5):779–790, 1997. 39, 59

[59] Grant Lythe, Robin E Callard, Rollo L Hoare, and Carmen Molina-París. How many TCR clonotypes does a body maintain ? *Journal of Theoretical Biology*, 389:214–224, 2016. 40

[60] Kevin Larimore, Michael W McCormick, Harlan S Robins, and Philip D Greenberg. Shaping of Human Germline IgH Repertoires Revealed by Deep Sequencing. *J Immunol*, Aug 2012. 42

[61] A. M Sherwood, C Desmarais, R. J Livingston, J Andriesen, M Haussler, C. S Carlson, and H Robins. Deep Sequencing of the Human TCR and TCR Repertoires Suggests that TCR Rearranges After and T Cell Commitment. *Sci Transl Med*, 3(90):90ra61–90ra61, Jul 2011. 42

[62] Harlan S Robins, Paulo V Campregher, Santosh K Srivastava, Abigail Wacher, Cameron J Turtle, Orsalem Kahsai, Stanley R Riddell, Edus H Warren, and Christopher S Carlson. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood*, 114(19):4099–107, Nov 2009. 42

[63] Ivan V. Zvyagin, Mikhail V. Pogorelyy, Marina E. Ivanova, Ekaterina A. Komech, Mikhail Shugay, Dmitry A. Bolotin, Andrey A. Shelenkov, Alexey A. Kurnosov, Dmitriy B. Staroverov, Dmitriy M. Chudakov, Yuri B. Lebedev, and Ilgar Z. Mamedov. Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing. *Proceedings of the National Academy of Sciences*, 2014. 42, 52

[64] Thierry Mora, Aleksandra M Walczak, William Bialek, and Curtis G Callan. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 107(12):5405–5410, 2010. 42, 47

[65] Emily R. Stirk, Grant Lythe, Hugo A. van den Berg, and Carmen Molina-París. Stochastic competitive exclusion in the maintenance of the naïve T cell repertoire. *Journal of Theoretical Biology*, 265(3):396–410, 2010. 42, 56

[66] Emily R. Stirk, Carmen Molina-París, and Hugo A. van den Berg. Stochastic niche structure and diversity maintenance in the T cell repertoire. *J. Theor. Biol.*, 255(2):237–249, November 2008. 42

[67] Rob J De Boer, Antonio A Freitas, and Alan S Perelson. Resource competition determines selection of B cell repertoires. *Journal of theoretical biology*, 212(3):333–343, 2001. 42, 43, 63

[68] Afonso R. M. Almeida, Inês F. Amado, Joseph Reynolds, Julien Berges, Grant Lythe, Carmen Molina-París, and Antonio a. Freitas. Quorum-Sensing in CD4+ T Cell Homeostasis: A Hypothesis and a Model. *Frontiers in Immunology*, 3(May):1–15, 2012. 42

[69] Tharindi Hapuarachchi, Joanna Lewis, and Robin E. Callard. A mechanistic model for naive CD4 T cell homeostasis in healthy adults and children. *Frontiers in Immunology*, 4(NOV):2–7, 2013. 42

[70] Joseph Reynolds, Mark Coles, Grant Lythe, and Carmen Molina-París. Deterministic and stochastic naïve T cell population dynamics : symmetric and asymmetric cell division. *Dynamical Systems*, 27(1):75–103, 2012. 42

[71] Amy E Troy and Hao Shen. Cutting edge: homeostatic proliferation of peripheral T lymphocytes is regulated by clonal competition. *Journal of immunology (Baltimore, Md. : 1950)*, 170(2):672–676, 2003. 43

[72] T.W. Mak and M.E. Saunders. *The Immune Response: Basic and Clinical Principles*. Number vol. 1 in The Immune Response: Basic and Clinical Principles. Elsevier/Academic, 2006. 43

[73] R J De Boer and A S Perelson. T cell repertoires and competitive exclusion. *Journal of theoretical biology*, 169(4):375–390, 1994. 43

[74] António A. Freitas, Maria Manuela Rosado, Anne Claire Viale, and Alf Grandien. The role of cellular competition in B cell survival and selection of B cell repertoires. *European Journal of Immunology*, 25(6):1729–1738, 1995. 43, 52

[75] Andrej Kosmrlj, Abhishek K Jha, Eric S Huseby, Mehran Kardar, and Arup K Chakraborty. How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. *Proc Natl Acad Sci USA*, 105(43):16671–6, Oct 2008. 45

[76] Andrej Kosmrlj, Elizabeth L Read, Ying Qi, Todd M Allen, Marcus Altfeld, Steven G Deeks, Florencia Pereyra, Mary Carrington, Bruce D Walker, and Arup K Chakraborty. Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection. *Nature*, 465(7296):350–354, 2010. 45

[77] Kaja Murali-Krishna, John D. Altman, M. Suresh, David J D Sourdive, Allan J. Zajac, Joseph D. Miller, Jill Slansky, and Rafi Ahmed. Counting antigen-specific CD8 T cells: A reevaluation of bystander activation during viral infection. *Immunity*, 8(2):177–187, 1998. 45

[78] Susan M Kaech, E John Wherry, and Raft Ahmed. Effector and memory T-cell differentiation: implications for vaccine development. *Nature reviews. Immunology*, 2(4):251–262, 2002. 45

[79] Steve Pressé, Kingshuk Ghosh, Julian Lee, and Ken A. Dill. Principles of maximum entropy and maximum caliber in statistical physics. *Reviews of Modern Physics*, 85(3):1115–1141, 2013. 46

[80] Andrea Cavagna, Irene Giardina, Francesco Ginelli, Thierry Mora, Duccio Piovani, Raffaele Tavarone, and Aleksandra M. Walczak. Dynamical maximum entropy approach to flocking. *Phys. Rev. E*, 89:042707, Apr 2014. 46

[81] Aaron Clauset, Cosma Rohilla Shalizi, and M.E. J. Newman. Power-law distributions in empirical data. *SIAM Rev*, 51(4):661–703, 2009. 47, 113

[82] Olesya V. Bolkhovskaya, Daniil Yu. Zorin, and Mikhail V. Ivanchenko. Assessing T Cell Clonal Size Distribution: A Non-Parametric Approach. *PLoS ONE*, 9(10):e108658, 2014. 47, 113

[83] K S Schluns, W C Kieper, S C Jameson, and L Lefrançois. Interleukin-7 mediates the homeostasis of naïve and memory CD8 T cells in vivo. *Nature immunology*, 1(5):426–432, 2000. 48, 52

[84] J T Tan, E Dudl, E LeRoy, R Murray, J Sprent, K I Weinberg, and C D Surh. IL-7 is critical for homeostatic proliferation and survival of naive T cells. *Proceedings of the National Academy of Sciences*, 98(15):8732–8737, 2001. 48, 52

[85] Benedict Seddon and Rose Zamoyska. TCR signals mediated by Src family kinases are essential for the survival of naive T cells. *Journal of immunology*, 169(6):2997–3005, 2002. 52

[86] C Tanchot, F A Lemonnier, B Pérarnau, A A Freitas, and B Rocha. Differential requirements for survival and proliferation of CD8 naïve or memory T cells. *Science*, 276(5321):2057–2062, 1997. 52

[87] D Nesić and S Vukmanović. MHC class I is required for peripheral accumulation of CD8+ thymic emigrants. *Journal of immunology*, 160(8):3705–3712, 1998. 52

[88] Katherine Best, Theres Oakes, James M. Heather, John-Shawe Taylor, and Benny Chain. Sequence and primer independent stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *bioRxiv*, 2014(20):1–7, 2014. 52

[89] Christopher Vollmers, Rene V Sit, Joshua A Weinstein, Cornelia L Dekker, and Stephen R Quake. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proceedings of the National Academy of Sciences*, 110(33):13463–8, 2013. 52

[90] Brandon J Dekosky, Takaaki Kojima, Alexa Rodin, Wissam Charab, Gregory C Ippolito, Andrew D Ellington, and George Georgiou. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nature Medicine*, 21(1):1–8, 2014. 52

[91] Didier Sornette and Rama Cont. Convergent multiplicative processes repelled from zero: power laws and truncated power laws. *J Phys I France*, 1997. 53

[92] Matteo Marsili, Sergei Maslov, and Yi-Cheng Zhang. Dynamical Optimization Theory of a Diversified Portfolio. *Physica A*, 253:9, 1998. 53

[93] Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226 – 251, 2004. 53

[94] Ivana Cvijović, Benjamin H Good, Elizabeth R Jerison, and Michael M Desai. The fate of a mutation in a fluctuating environment. *Preprint*, 2015. 53

[95] Anna Melbinger and Massimo Vergassola. Evolutionary fitness in variable enviroments. *Arxiv*, 92093(1):16, 2015. 53

[96] Stanislas Leibler and Edo Kussell. Individual histories and selection in heterogeneous populations. *Proc. Natl. Acad. Sci. U. S. A.*, 107(29):13183–13188, July 2010. 53

[97] Ville Mustonen and Michael Lässig. Fitness flux and ubiquity of adaptive evolution. *Proc. Natl. Acad. Sci. U. S. A.*, 107(9):4248–4253, March 2010. 53

[98] Olivier Rivoire and Stanislas Leibler. A model for the generation and transmission of variations in evolution. *Proc. Natl. Acad. Sci. U. S. A.*, 111(19):E1940–9, May 2014. 53

[99] Bahram Houchmandzadeh. Clustering of diffusing organisms. *Phys. Rev. E*, 2002. 55

[100] Petrov DA Messer PW. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol Evol.*, 2013. 55

[101] Paolo Sansoni et al. New advances in CMV and immunosenescence. *Exp. Gerontol.*, 55:54–62, 2014. 61

[102] Teresa Ferraro, Tanguy Lucas, Marie Clémot, Jose De Las Heras Chanes, Jonathan Desponds, Mathieu Coppey, Aleksandra M Walczak, and Nathalie Dostatni. New methods to image transcription in living fly embryos: the insights so far, and the prospects. *Wiley interdisciplinary reviews. Developmental biology*, February 2016. 69, 73, 80, 82, 84, 88

[103] Wolfgang Driever and Christiane Nu. The bicoid Protein Determines Position in the Drosophila Embryo in a Concentration-Dependent Manner. *Cell*, 54:95–104, 1988. 69, 80

[104] Julien O Dubuis, Gasper Tkacik, Eric F Wieschaus, Thomas Gregor, and William Bialek. Positional information , in bits. *Proceedings of the National Academy of Sciences*, 110(41), 2013. 69

[105] Gasper Tkacik, Julien O. Dubuis, Mariela D. Petkova, and Thomas Gregor. Positional information, positional error, and readout precision in morphogenesis: A mathematical framework. *Genetics*, 199(1):39–59, 2015. 69

[106] Patrick Hillenbrand, Gasper Tkacik, and Ulrich Gerland. Beyond the French Flag Model: Exploiting Spatial and Gene Regulatory Interactions for Positional Information. *arXiv preprint*, 2016. 70

[107] Etay Ziv, Ilya Nemenman, and Chris H Wiggins. Optimal Signal Processing in Small Stochastic Biochemical Networks. *PLoS ONE*, 2(10), 2007. 70

[108] Aleksandra M. Walczak, Ga šper Tkačik, and William Bialek. Optimizing information flow in small genetic networks. II. Feed-forward interactions. *Phys. Rev. E*, 81:041905, Apr 2010. 70

[109] Hongtao Chen, Zhe Xu, Constance Mei, Danyang Yu, and Stephen Small. A System of Repressor Gradients Spatially Organizes the Boundaries of Bicoid-Dependent Target Genes. *Cell*, 149(3):618–629, 2012. 70

[110] Angelike Stathopoulos and Michael Levine. Genomic Regulatory Networks and Animal Development. *Cell*, 9:449–462, 2005. 70

[111] Julia Zeitlinger, Robert P Zinzen, Alexander Stark, Manolis Kellis, Hailan Zhang, Richard A Young, and Michael Levine. Whole-genome ChIP – chip analysis of Dorsal , Twist , and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes and development*, (510):385–390, 2007. 70

[112] Eran Segal, Tali Raveh-Sadka, Mark Schroeder, Ulrich Unnerstall, and Ulrike Gaul. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, 451(7178):535–40, 2008. 72

[113] Johannes Jaeger. The gap gene network. *Cellular and Molecular Life Sciences*, 68(2):243–274, 2011. 72, 80

[114] William Bialek and Sima Setayeshgar. Physical limits to biochemical signaling. *Proceedings of the National Academy of Sciences*, 2005(102):10040–10045, 2005. 72

[115] Kazunari Kaizu, Wiet De Ronde, Joris Paijmans, Koichi Takahashi, and Filipe Tostevin. The Berg-Purcell Limit Revisited. *Biophysical journal*, 106(4):976–985, 2014. 72, 73

[116] Thomas Gregor, David W Tank, Eric F Wieschaus, and William Bialek. Probing the limits to positional information. *Cell*, 130(1):153–64, 2007. 73, 93, 98

[117] Feng He, Ying Wen, Jingyuan Deng, Xiaodong Lin, Long Jason Lu, Renjie Jiao, and Jun Ma. Probing Intrinsic Properties of a Robust Morphogen Gradient in Drosophila. *Developmental Cell*, 15(4):558–567, 2008. 73

[118] Michael W Perry, Alistair N Boettiger, and Michael Levine. Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. *Proceedings of the National Academy of Sciences*, 108(33):1–12, 2011. 73

[119] Michael W. Perry, Jacques P. Bothma, Ryan D. Luu, and Michael Levine. Precision of hunchback expression in the Drosophila embryo. *Current Biology*, 22(23):2247–2252, 2012. 73, 95

[120] Aude Porcher and Nathalie Dostatni. The bicoid morphogen system. *Current Biology*, 20(5):R249–54, 2010. 73, 80, 93, 146

[121] Edouard Bertrand, Pascal Chartrand, Matthias Schaefer, Shailesh M. Shenoy, Robert H. Singer, and Roy M. Long. Localization of ASH1 mRNA particles in living yeast. *Molecular Cell*, 2(4):437–445, 1998. 73, 99

[122] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–6, 2002. 74, 80

[123] Thomas Gregor, Hernan G Garcia, and Shawn C Little. The embryo as a laboratory: quantifying transcription in Drosophila. *Trends in Genetics*, 30(8):364–75, 2014. 80

[124] Mikhail Tikhonov, Shawn C Little, and Thomas Gregor. Only accessible information is useful: insights from patterning Subject Category. *Royal Society Open Science*, 2(150486), 2015. 80

[125] Shawn C Little, Mikhail Tikhonov, and Thomas Gregor. Precise developmental gene expression arises from globally stochastic transcriptional activity. *Cell*, 154(4):789–800, 2013. 80, 82, 93, 95, 98, 146

[126] Ertugrul M Ozbudak, Mukund Thattai, Iren Kurtser, Alan D Grossman, and Alexander van Oudenaarden. Regulation of noise in the expression of a single gene. *Nature Genetics*, 31(1):69–73, 2002. 80

[127] JM Raser and EK O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304:1811, 2004. 80

[128] Olivier Crauk and Nathalie Dostatni. Bicoid determines sharp and precise target gene expression in the Drosophila embryo. *Current Biology*, 15(21):1888–98, 2005. 80

[129] 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):472–4, 2011. 80, 82, 85, 95

[130] Benjamin Zoller, Damien Nicolas, Nacho Molina, and Felix Naef. Structure of silent transcription intervals and noise characteristics of mammalian genes. *Molecular Systems Biology*, 11(7):823, 2015. 80, 82, 85, 95

[131] Yuichi Taniguchi, Paul J Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X Sunney Xie. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–8, 2010. 80

[132] Meenakshisundaram Kandhavelu, Jason Lloyd-Price, Abhishekh Gupta, Anantha-Barathi Muthukrishnan, Olli Yli-Harja, and Andre S Ribeiro. Regulation of mean and noise of the in vivo kinetics of transcription under the control of the lac/ara-1 promoter. *FEBS Letters*, 586(21):3870–5, 2012. 80, 82, 85

[133] Anantha-Barathi Muthukrishnan, Meenakshisundaram Kandhavelu, Jason Lloyd-Price, Fedor Kudasov, Sharif Chowdhury, Olli Yli-Harja, and Andre S Ribeiro. Dynamics of transcription driven by the tetA promoter, one event at a time, in live Escherichia coli cells. *Nucleic Acids Research*, 40(17):8472–83, 2012. 80, 82, 85

[134] Shasha Chong, Chongyi Chen, Hao Ge, and X Sunney Xie. Mechanism of transcriptional bursting in bacteria. *Cell*, 158(2):314–26, 2014. 80

[135] Tanguy Lucas and et al. in preparation. 2016. 82, 95, 98

[136] Javier Estrada, Felix Wong, Angela DePace, and Jeremy Gunawardena. Information Integration and Energy Expenditure in Gene Regulation. *Cell*, 166(1):234–44, 2016. 85

[137] Antoine Coulon, Matthew L Ferguson, Valeria de Turris, Murali Palangat, Carson C Chow, and Daniel R Larson. Kinetic competition during the transcription cycle results in stochastic RNA processing. *eLife*, 3:1–22, 2014. 85

[138] Antoine Coulon and David R Larson. Fluctuating Analysis: Dissecting Transcriptional Kinetics with Signal Theory. *Methods in Enzymology*, 03:1–33, 2016. 88

[139] Thomas Gregor, Eric F Wieschaus, Alistair P McGregor, William Bialek, and David W Tank. Stability and nuclear dynamics of the bicoid morphogen gradient. *Cell*, 130(1):141–52, 2007. 92, 93, 146

[140] Thorsten Erdmann, Martin Howard, and Pieter Rein ten Wolde. Role of spatial averaging in the precision of gene expression patterns. *Physical Review Letters*, 103(25):258101, 2009. 93, 98

[141] Takashi Fukaya, Bomyi Lim, and Michael Levine. Enhancer Control of Transcriptional Bursting. *Cell*, pages 1–11, 2016. 96

[142] Yufang Wang, Kimberly C Tu, N P Ong, Bonnie L Bassler, and Ned S Wingreen. Protein-level fluctuation correlation at the microcolony level and its application to the Vibrio harveyi quorum-sensing circuit. *Biophysical Journal*, 100(12):3045–53, 2011. 96

[143] Yihan Lin, Chang Ho Sohn, Chiraj K. Dalal, Long Cai, and Michael B. Elowitz. Combinatorial gene regulation by modulation of relative pulse timing. *Nature*, 527(7576):54–8, 2015. 96

[144] Mary J Dunlop, Robert Sidney Cox, Joseph H Levine, Richard M Murray, and Michael B Elowitz. Regulatory activity revealed by dynamic correlations in gene expression noise. *Nature Genetics*, 40(12):1493–8, 2008. 96

[145] B Munsky, G Neuert, and A van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–7, 2012. 96

[146] Teresa Ferraro, Emilia Esposito, Laure Mancini, Sam Ng, Tanguy Lucas, Mathieu Coppey, Nathalie Dostatni, Aleksandra M Walczak, Michael Levine, and Mounia Lagha. Transcriptional Memory in the Drosophila Embryo. *Current Biology*, 26(2):212–8, 2016. 96, 99

[147] Thomas Gregor and Gasper et al Tkacik. oral communication. 2016. 98

[148] Yurie Okabe-Oho, Hiroki Murakami, Suguru Oho, and Masaki Sasai. Stable, precise, and reproducible patterning of bicoid and hunchback molecules in the early Drosophila embryo. *PLoS Computational Biology*, 5(8):e1000486, 2009. 98

[149] Susan M. Janicki, Toshiro Tsukamoto, Simone E. Salghetti, William P. Tansey, Ravi Sachidanandam, Kannanganattu V. Prasanth, Thomas Ried, Yaron Shav-Tal, Edouard Bertrand, Robert H. Singer, and David L. Spector. From silencing to gene expression: Real-time analysis in single cells. *Cell*, 116(5):683–698, 2004. 99

[150] Katerina R Katsani, Roger E Karess, Nathalie Dostatni, and Valerie Doye. In Vivo Dynamics of Drosophila Nuclear Envelope Components. *Molecular Biology of the Cell*, 19:3652–3666, 2008. 99

[151] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, NY, USA, 1991. 110

[152] Dmitri Bratsun, Dmitri Volfson, Lev S Tsimring, and Jeff Hasty. Delay-induced stochastic oscillations in gene regulation. *Proceedings of the National Academy of Sciences*, 102(41):14593–14598, 2005. 138

Résumé

Nous présentons deux problèmes de biologie faisant appel à un traitement de données et des modèles issus de la physique statistique : la dynamique des populations en immunologie et la régulation génétique dans le développement embryonnaire. En immunologie, nous étudions le problème de la sélection somatique dans le système immunitaire adaptatif: la sélection cellulaire et la compétition qui s'y opèrent, constituant un système quasi Darwinien au sein de l'organisme. Dans un premier temps, nous considérons différentes hypothèses sur la dynamique selective : signaux déclenchant la division ou la mort cellulaire par liaison antigénique ou par cytokines, paramètres dynamiques de division, mort et fluctuations environnementales. Nous explorons leur influence sur la taille des clones dont la distribution à queue lourde a été observée à travers les espèces et les types de cellules. Deux familles de modèles émergent : un premier dans lequel le bruit est cohérent à l'échelle du clone et un second dans lequel le bruit varie de cellule à cellule. Nous montrons dans quelle mesure la distribution de taille de clones permet de déterminer le meilleur modèle et relions la forme de la distribution ainsi que l'exposant apparent de la loi de puissance aux paramètres biologiques. Dans un second temps, nous explorons les caractéristiques du réseau complexe et aléatoire formé par les clones et les antigènes : dimension, adjacence, dynamique. Nous nous intéressons à l'effet de la sélection dans le temps et à la vitesse d'évolution des clones.
La deuxième partie de cette thèse est consacrée au développement embryonnaire. Dans l'embryon, il est essentiel pour le noyau de déterminer sa position avec une grande précision pour orienter la différentiation et construire un organisme structuré viable. Cette information positionnelle est acquise, transmise et conservée par la diffusion de protéines et l'activa- tion de circuits génétiques. Plus précisément, la formation de l'axe antéro-postérieur chez la Drosophile est déterminée entre autres par l'activation du gène *hunchback* par la protéine Bicoid. Nous analysons des données issues d'expériences d'imagerie fluorescente dynamique dans les premiers cycles cellulaires de l'embryon. Nous contruisons un modèle spécifique permettant d'analyser la fonction d'autocorrélation des traces temporelles de fluorescence qui prend en compte toutes les difficultés biologiques et expérimentales (bruit, calibration traces courtes, structure du gène artificiel) pour extraire les paramètre dynamiques d'activation de *hunchback*. Nous examinons différentes dynamiques potentielles (poisonnienne, markovienne ou non markovienne) et leur implication pour l'information dont la cellule dispose sur sa position ainsi que la précision de la lecture du gradient de Bicoid.

Mots Clés

Immunologie, Développement, Dynamique des populations, Circuits génétiques

Abstract

This work presents two problems of biology requiring data analysis and models from statistical mechanics: population dynamics in immunology and gene regulation in embryo development. In immunology I study the problem of somatic evolution in the adaptive immune system: selection of and competition among cells that form a close-to-Darwinian system within one individual. First, I consider different potential hypotheses for selective dynamics: division and death signals through antigen binding or cytokines, dynamical parameters for division, death and fluctuations of the environment. I explore their impact on clone sizes. Experimentally, these clone sizes show heavy tail distributions for different species and different pools of cells. Two families of models emerge: models where noise is consistent at the level of the clone and models where it varies from cell to cell. I show how clone size distributions help discriminate between these models and relate the shape of the distribution and the exponent of the power law to biological parameters. Second, I explore the specifics of the complex stochastic network of clones and antigens: its dimensionality, connectivity and dynamics. I study the effect of selection at different time scales and the speed of evolution of the clones.
The second part of this dissertation concerns embryo development. In the fly embryo, it is crucial that nuclei can evaluate their position within the organism accurately to determine cell fate and build a healthy organism. This positional information is obtained, transferred, and maintained through diffusion of proteins and activation of genetic networks. More specifically, the patterning of the antero-posterior axis in drosophila requires the *hunchback* gene, activated by the Bicoid protein. I analyze data from fluorescent live imaging in the early cell cycles of the embryo. I build a tailor-made model to analyze autocorrelation functions of fluorescence time traces overcoming all biological and experimental challenges (noise, calibration, short traces, transgene construct) to extract the parameters of *hunchback* activation. I examine several potential types of dynamics for gene switiching (Poisson, Markovian or non-Markovian) and predict their impact on positional information and the accuracy of bicoid gradient readout.

Keywords

Immunology, Development, Population dynamics, Gene regulatory network

Numéro national de thèse:
xxxxx