Variation in protein expression levels in cell populations

Thiago S. Guzella

Dissertation presented to obtain the Ph.D degree in Systems Biology

Instituto de Tecnologia Química e Biológica | Universidade Nova de Lisboa

Apoio financeiro da FCT e do FSE no âmbito do Quadro Comunitário de apoio, BD n. SFRH / BD / 33572 / 2008

Acknowledgements

First of all, I would like to thank my supervisors, Vasco Barreto and Jorge Carneiro for the unique opportunity, their honest comments, guidance and support throughout all these years, and in special for the freedom I was given in order to pursue my ideas. I would not have got to this point if it weren't for the environment they work so hard to provide. To Vasco, thank you for providing me with the opportunity to develop the experimental part of this work, for the immense patience and understanding in dealing with someone coming to the lab from an Electrical Engineering background. For the kindness and wisdom so many times, and for providing much needed calm and lucid insight during difficult times I had with experiments in these years. To Jorge, thank you for being such an amazing mentor in every possible way, for your ability to always look further and to bridge experiments and theory, and your leadership. For always being available, your principled approach in dealing with all matters, scientific or otherwise, the constant optimism, kindness and wisdom.

I would like to thank the past and present members of the two groups (Epigenetics and Soma, Quantitative Organism Biology) I had the immense pleasure to belong to, and who contributed to many aspects of my training and this work. To Clara Pereira, Nadine Caratão, Catarina Cortesão, Inês Trancoso, Raquel Freitas, Filipa Marta, Daniel Espadinha, from the Epigenetics and Soma group, thank you very much for the patience, scientific curiosity and critical thinking all this time. To Nuno Sepúlveda, Tom Weber, Tiago Macêdo, Danesh Tarapore, Pedro Silva, from the Quantitative Organism Biology group, for the nice environment and the many discussions all these years, and in special to Danesh Tarapore for so many discussions and suggestions during the time I was writing the thesis.

I am also very grateful to my thesis committee, Jocelyne Demengeot and Henrique Teotónio, for the guidance and support during the development of this work. For the pragmatic view and for the numerous suggestions, many of which I unfortunately did not manage to follow, but stepping back I realize how valuable your input was, and in particular how lucky I was to have you as my thesis committee. To both of you for so many times making time in your busy schedules to talk to me, whether in the setup of the meeting, or separately. For pushing me to take advantage of the structure of the institute, and especially to Jocelyne, whose group was absolutely fundamental in allowing me to perform the experimental part of this work.

On this matter, I would like to thank the members of the Lymphocyte Physiology group, namely Lisa Bergman, Andreia Lino, Ana Catarina Martins, Ricardo Paiva, Marie Bonnet

and Elodie Mohr, for all the help along these years, with whom I learned so much, and for sharing the enthusiasm in working with such a great experimental system. For the late nights often spent in the flow cytometers, and the help and sympathy in times of technical problems. In special to Catarina, who had to put up with my constant questions and requests for reagents. I am extremely grateful to Jocelyne and Andreia, whose comments at a relatively early stage of the experimental work helped me in trying to be more careful and critical in terms of the experimental setup. I also benefited a lot from the great insight and drive of Andreia, always full of ideas and alternatives for trying to solve a problem. Thank you very much. Many thanks also to Ricardo, who dedicated so much time to clarifying experimental protocols, discussing ideas, and making suggestions.

My special thanks to the director of the PGD-2008 PhD programme Henrique Teotónio, to Prof. António Coutinho, and those who gave me the amazing opportunity to join this PhD programme here in the Instituto Gulbenkian de Ciência (IGC). I also thank Manuela Cordeiro for the assistance, and to the latter directors of the IGC internal PhD programme, Thiago Carvalho and Élio Sucena, for the support. To the colleagues from the PGD-2008, PDBC-2008 and INDP-2008 PhD programmes, for the amazing environment during the lectures and afterwards. And many thanks to Christen Mirth for revising the summary of this thesis.

Many thanks to Daniel Damineli, for the great friendship and all the help in so many things, especially in organizing my way of thinking and improving my way of making a presentation, and also to Maitê Portes, for the kindness and wisdom. To Leila Shirai and Vlad Coroama, for the great insight and late-night discussions. To the closer friends from the PGD-2008 programme, Mariluz Gomez and Cláudia Martinho, for the friendship all these years. And in special to Nadja Pejanovic, for the highly contagious enthusiasm and energy.

I also would like to thank several people from the IGC for sharing insight and the willingness to discuss. To Daniel Damineli and Artemy Kolchinsky for discussions on so many theoretical aspects. To Íris Caramalho, for the kindness, wisdom, help and several suggestions. To many people that took the time to help and discuss experimental plans, ideas and problems, namely Thiago Carvalho, Raffaella Gozzelino, Ivo Marguti, Lígia Gonçalves, Elisabetta Padovan, Joana Rodo, and especially to Rasmus Larsen, for the assistance, but most of all for the numerous words of wisdom. To Research Fellows and Principal Investigators in the institute, namely Miguel Godinho Ferreira, Mónica Betterncourt Dias, Lars Jansen, Carlos Tadokoro, with whom I had the pleasure to discuss about my work, and whose wisdom and comments in diverse times were so illuminating. To those in the

Quantitative Genetics Journal Club for insightful discussions, which greatly contributed to the general discussion of this thesis relating the work developed here with quantitative genetics, and especially to José Álvarez-Castro for the keen insight and the great input on the general discussion.

The experimental part of this work would not have been possible without the outstanding technical competence and great assistance of Rui Gardner and the UIC team, namely Telma Lopes, Ana Margarida Nascimento, Cláudia Bispo and Cláudia De Andrade. Thank you for the great patience and for pushing cell sorting, and the support in flow cytometric analysis in general. I also would like to take the opportunity to acknowledge Manuel Rebelo for animal house management and the assistance in arranging orders for some experiments. And, of course, Rosa Santos and Ana Regalado for antibody production, and for putting up with my tendency to "devour" antibodies.

I would like to acknowledge the financial support of Fundação para a Ciência e Tecnologia (fellowship number SFRH/BD/33572/2008), and the Portuguese people, whose taxes provided the basis for my PhD fellowship and much of the research I did. To IGC, not only for the amazing scientific environment, and also for the financial support in the last year, which allowed me to complete the work.

To one of my first scientific mentors Tomaz Mota-Santos, who encouraged me, while an undergraduate student of Electrical Engineering, to pursue my interests in biology, and eventually in doing a PhD in IGC. And to another mentor, Walmir Caminhas, who provided me at that time with the freedom, support, encouragement and the means to pursue those interests.

And to my family, for the patience, support and understanding, for whom there are no words in English, Portuguese, or perhaps any language that allow me to fully express myself.

Summary

A widespread observation is that of variation in the expression levels of a molecule when analyzing a snapshot of single cells from a population. One possible explanation for this variation is that there are asynchronous fluctuations throughout time in the expression level of each cell, for example due to noise in gene expression. Another explanation is that there are so-called stable variants, groups of cells that have a permanent bias for a limited range of expression levels compared with the levels observed in a snapshot of the population. In an extreme scenario, each cell of the population would be a single stable variant with an expression level that is constant as a function of time, without any fluctuations. Stable variants could occur if the population is genetically diverse, with each clone being associated with a particular expression level. Moreover, stable variants may be present even in an isogenic population, especially in the case of cells from multicellular organisms, akin to differentiation stages or cell lineages that differentially regulate expression of the molecule. This constitutes what will be referred to as epigenetic variation, in a broad context, to conceptualize the underlying process leading to phenotypic differences that persist throughout time. Overall, however, it is not known whether the variation in expression levels observed in a snapshot of single cells from a population is explained by i) fluctuations in the levels of single cells, ii) the expression level being constant as a function of time in each cell, or iii) a combination of the two, namely stable variants with fluctuations in the expression levels of single cells. One of the objectives of this work is to study which one of these cases explains the variation in the expression levels that is observed, which is referred to as the "origin of the variation". Moreover, provided that fluctuations do occur in cells of the population, one would expect the difference between the expression levels of two subsets of cells to change over a certain period of time. Therefore, another question that arises is what is the amount of time necessary for this to take place, which provides a notion of the "timescale of the variation". Despite several experimental works that address, whether explicitly or not, these two questions, the ability to understand how different mechanisms shape the variation in expression levels that is observed in a cell population is limited by the fact that it remains unclear how to formulate the origin and timescale of the variation.

In chapter 2 of this thesis, we develop a quantitative theoretical framework to address these two questions, based on classifying the mechanisms that influence the expression levels in a cell population into two components, one stable and the other unstable. In a simplified way, the stable component represents permanent differences between the expression levels of two subsets of cells; on the other hand, the unstable component represents differ-

ences that are transient, and will eventually vanish over time. This leads to the concept of sub-population, a set of cells in which all variation observed is due to the unstable component. We use these definitions to describe protein expression levels in a heterogeneous cell population, composed of several sub-populations (stable variants), based on a simplified model of constitutive protein expression. In the context of this model, the stable component arises due to variation in the time-averaged rate of protein production of the different sub-populations, and two parameters are defined, termed R^2_α and τ_T, formalizing the notions of the origin and timescale of the variation, respectively. The estimation of these two parameters is done by isolating subsets of cells and quantifying their expression levels over a sufficiently long period of time. Even if analysis is done over a period of time that is not very long, it is possible to interpret the estimates, which constrain parameter values that are compatible with the data. Parameter R^2_α quantifies the relative contribution of the stable component, ranging from 0% (only the unstable component is present) to 100% (stable component only), and the intermediate case $0\% < R^2_\alpha < 100\%$ denotes that a combination of the two components is at play. In turn, parameter τ_T, referred to as the characteristic time of the variation, is related to the time taken for the expression levels in the subsets of cells that have been isolated to relax to their stationary values, with dynamics dictated by the unstable component. This parameter depends on properties that dictate how expression of the protein is regulated in single cells, such as the dynamics of change in the instantaneous rate of protein production and the mean lifetime of the protein.

Afterwards, in chapter 3, we rely on this theoretical framework to address experimentally how different mechanisms shape the variation in expression levels of the T Cell Receptor (TCR) in mouse CD4$^+$ T cells. In a wild-type animal, as a consequence of a process of somatic DNA rearrangement, this cell population is genetically diverse, and hence referred to as a polyclonal population, since it is composed of several clones. In a given lymphocyte population, variation in expression levels of a molecule may come about due to genetic and epigenetic variation, which mold the stable component, and also to fluctuations in the expression level of each cell throughout time, representing the unstable component. We quantify R^2_α and τ_T for the TCR in a polyclonal (genetically diverse) and in two isogenic populations (that are TCR-transgenic and lack the process of somatic rearrangement), and by evaluating the values of R^2_α in these various populations, we assess the impact of genetic and epigenetic variation on the stable component in the polyclonal population. This quantification is done in an *in vitro* setup, in which neither stimulation of the cells nor cell division is expected, limited to a time frame of 3–4 days. After analysis, the description of the data considered most adequate indicates the unstable component as the main contri-

bution in the two isogenic populations studied, but also provides preliminary evidence for the stable component in these two populations. On the other hand, for the polyclonal population, the description of the data in this *in vitro* setup implies that the stable component is the main contribution, and, along with the results of adoptive transfers (*in vivo* setup), provides strong evidence for the stable component in this population. We infer, based on the comparison between the values of R^2_α estimated *in vitro*, that under these conditions genetic variation would be the main explanatory factor for the stable component in the polyclonal population, and that epigenetic variation would have a relatively small impact. Altogether, these results establish the TCR in a polyclonal population of $CD4^+$ T cells as a model system to study how the stable and unstable components contribute to variation in expression levels.

In conclusion, this thesis contributes to understanding how two processes affecting the expression levels in each cell throughout time, stochastic fluctuations and permanent bias, shape the variation that is observed in a snapshot of the population. We develop a theoretical quantitative framework to formalize and quantify certain properties of the expression levels, and put forward an experimental model system to study the interplay of such processes. In this way, this work provides an integrated view of the expression levels as a quantitative trait, bridging concepts from physics, systems biology and quantitative genetics.

Resumo

Uma observação muito comum é a de variação nos níveis de expressão de uma molécula ao se observar instantaneamente diferentes células de uma população. Uma possível explicação para esta variação é a existência de flutuações assíncronas ao longo do tempo no nível de expressão de cada célula, por exemplo, devido a ruído na expressão génica. Outra explicação é que existem as chamadas variantes estáveis, grupos de células que têm um viés permanente numa gama de níveis de expressão que é limitada em comparação com os níveis observados na população. Num caso extremo, cada célula da população seria uma única variante estável com um nível de expressão que é constante em função do tempo, sem qualquer flutuação. Variantes estáveis podem estar presentes se a população é geneticamente diversa, com cada clone associado a um nível de expressão particular. Além disso, as variantes estáveis podem estar presentes mesmo numa população isogénica, em especial no caso de células de organismos multicelulares, de forma semelhante a estágios de diferenciação ou linhagens celulares que regulam diferentemente a expressão da molécula. Isto constitui o que será referido como variação epigenética, num contexto amplo, para conceptualizar os processos subjacentes levando a diferenças fenotípicas que persistem ao longo do tempo. Em geral, no entanto, não se sabe se a variação nos níveis de expressão de células individuais observada num instante de tempo numa população é explicada por i) flutuações no nível de expressão de cada célula, ii) um nível de expressão constante em função do tempo em cada célula, ou iii) uma combinação de ambos, ou seja, variantes estáveis que têm flutuações no nível de expressão de cada célula. Um dos objetivos desse trabalho é estudar qual desses casos explica a variação nos níveis de expressão que é observada, que de agora em diante será referido como a "origem da variação". Além disso, contanto que haja flutuações no nível de células da população, é de se esperar que a diferença entre os níveis de expressão de dois subconjuntos de células mude ao longo de um certo período de tempo. Assim, uma outra pergunta que surge é qual o tempo necessário para que isso ocorra, o que fornece uma noção da "escala de tempo da variação". Apesar de vários trabalhos experimentais que abordam, de uma forma explícita ou não, essas duas questões, a capacidade de compreender como é que diferentes mecanismos moldam a variação nos níveis de expressão é limitada pelo fato de não estar claro como formular e abordar a origem e a escala de tempo da variação.

No capítulo 2 desta tese, desenvolvemos uma base teórica quantitativa para abordar estas duas questões, que se baseia na classificação dos mecanismos que influenciam os níveis de expressão numa população de células em duas componentes, uma estável e a outra instá-

vel. De uma forma simplificada, a componente estável representa diferenças permanentes entre os níveis de dois subconjuntos de células de expressão, enquanto que a componente instável representa diferenças que são transitórias, e eventualmente desaparecem ao longo do tempo. Isto dá origem ao conceito de sub-população, um conjunto de células em que toda a variação observada é devida à componente instável. Estas definições foram usadas para descrever os níveis de expressão de uma proteína numa população heterogénea de células, composta por várias sub-populações (variantes estáveis), com base num modelo simplificado de expressão constitutiva da proteína. No contexto deste modelo, a componente estável surge devido à variação na taxa média, ao longo do tempo, de produção da proteína nas diferentes sub-populações, e dois parâmetros são definidos, denominados R^2_α e τ_T, formalizando as noções de origem e escala de tempo da variação, respectivamente. A estimação desses dois parâmetros é feita isolando-se subconjuntos de células e quantificando os níveis de expressão ao longo de um período de tempo suficientemente longo. Mesmo quando a análise é feita durante um período de tempo não muito longo, é possível interpretar-se as estimativas, que nesse caso restringem os valores dos parâmetros que são compatíveis com os dados. O parâmetro R^2_α quantifica a contribuição relativa da componente estável, variando de 0% (apenas a componente instável está presente) até 100% (componente estável apenas), e o caso intermédio $0\% < R^2_\alpha < 100\%$ indica que há uma combinação das duas componentes. Por sua vez, o parâmetro τ_T, referido como o tempo característico da variação, está relacionado com o tempo necessário para que os níveis de expressão nos subconjuntos de células que foram isolados relaxem aos seus valores estacionários, com uma dinâmica que está relacionada com a componente instável. Este parâmetro depende de propriedades que ditam como a expressão da proteína é regulada nas células, tais como a dinâmica de alterações na taxa instantânea de produção da proteína e o tempo de vida médio da proteína.

No capítulo 3, usámos esta base teórica para abordar experimentalmente como diferentes mecanismos moldam a variação dos níveis de expressão do receptor de células T (RCT) em células T CD4$^+$ de ratinho. Num animal não-manipulado geneticamente, como consequência de um processo de rearranjo somático do ADN, esta população de células é geneticamente diversa, e, portanto, referida como uma população policlonal, uma vez que é composta por vários clones. Numa determinada população de linfócitos, a variação nos níveis de expressão de uma molécula pode acontecer devido à variação genética e epigenética, que moldam a componente estável, e também devido a flutuações no nível de expressão de cada célula ao longo do tempo, o que representa a componente instável. Quantificámos R^2_α e τ_T para o RCT em uma população policlonal (geneticamente diversa)

e em duas populações isogénicas (que são transgénicas para o RCT e têm o rearranjo somático inibido), e, com base nos valores de R^2_α nestas várias populações, avaliámos o impacto de variação genética e epigenética na componente estável na população policlonal. Esta quantificação foi feita numa condição *in vitro*, na qual não se espera haver estimulação das células, nem divisão celular, limitada a um intervalo de tempo de 3–4 dias. Após análise, a descrição dos dados considerada mais adequada indica que a componente instável é a principal contribuição nas duas populações isogénicas estudadas, mas também fornece evidência preliminar para a componente estável nestas duas populações. Por outro lado, a descrição dos dados da população policlonal nesta condição *in vitro* revela que a componente estável é a principal contribuição, e, juntamente com os resultados de transferências adoptivas (in vivo), fornece fortes evidências para a componente estável nesta população. Com base na comparação dos valores de R^2_α estimados *in vitro*, inferimos que na condição *in vitro* a variação genética é o principal factor explicativo para a componente estável nesta população, e que a variação não-genética estável tem um impacto relativamente pequeno. Estes resultados estabelecem o TCR na população policlonal de células T $CD4^+$ como um modelo experimental para estudar como as componentes estável e instável contribuem para a variação nos níveis de expressão.

Em conclusão, este trabalho contribui para a compreensão de como dois processos que afectam os níveis de expressão em cada célula ao longo do tempo, as flutuações estocásticas e o viés permanente, moldam a variação que é observada na população num determinado instante de tempo. Desenvolvemos uma base quantitativa teórica para formalizar e quantificar certas propriedades dos níveis de expressão, e apresentamos um modelo experimental para estudar a ação combinada de tais processos. Desta forma, esta tese oferece uma visão integrada dos níveis de expressão como um traço quantitativo, unindo conceitos de física, biologia de sistemas e genética quantitativa.

Contents

List of Figures

List of Tables

List of Abbreviations

AIC Akaike Information Criterion

AICD Activation-Induced Cell Death

APC Antigen Presenting Cell

CI Confidence interval

DP Double Positive

ER Endoplasmic Reticulum

FACS Fluorescence-Activated Cell Sorting

GFP Green Fluorescent Protein

iPSC Induced pluripotent stem cell

mESC Mouse Embryonic Stem Cell

MHC Major Histocompatibility Complex

mRNA Messenger Ribonucleic Acid

NHEJ Non-Homologous End Joining

OU process Ornstein-Uhlenbeck process

Rag2 Recombination Activating Gene 2

RSS Recombination Signal Sequences

TCR T Cell Receptor

Chapter 1

INTRODUCTION

The regulation of gene expression is a key process in converting genetic information into a more functional form. The amount of a particular protein or mRNA that is expressed by a cell is one property that can be studied in order to better understand this process. In doing so, via a quantitative analysis of gene expression, one aspires to obtain a better understanding of the inner workings of a cell, the most basic unit of life.

Such a microscopic level of study has interesting features. Perhaps most striking is that the property of interest is expected to be time-varying, fluctuating as a function of time, even in the absence of an external, nominal, perturbation. When representing the expression level in a single cell as a time-series, one would hence observe fluctuations. Such fluctuations may arise, at least in part, as a consequence of the fact that gene expression depends on regulatory molecules that are present in small copy number. In order to better understand some of the potential underlying causes for these fluctuations, the next section surveys related works.

1.1 Fluctuations in expression levels in a single cell across time

This section reviews studies that have analyzed fluctuations in expression levels in single cells. In this context, the term *fluctuation* is used to denote variation in the value of a quantity around a mean value as a function of time. As many regulatory molecules may be present in small copy numbers in a cell, the resulting stochastic effects on gene expression are expected to introduce fluctuations in the level of a particular molecule that is expressed by a cell. These specific effects, under the reference of *noise in gene expression*, have been under intense study recently, and are reviewed in section 1.1.1. Despite this extensive emphasis, however, one must keep in mind that noise in gene expression is unlikely to be the only factor underlying the fluctuations that are observed. In this sense, section 1.1.2 highlights that, independently of the specific mechanism, the fundamental property is that the fluctuations would be stochastic, such that differences in the expression levels of two subsets of cells would be transient.

1.1.1 Noise in gene expression

Noise in gene expression, arising due to the small copy number of molecules involved in the processes in gene expression, has been studied from a theoretical point of view using stochastic models of gene expression. A common ground for all stochastic models is the chemical master equation (Gillespie, 2007). This equation is a Markov model that, under

the assumption of a homogeneous (well-stirred) medium, describes the changes in the number of molecules in the system, given a set of possible reactions. Two main approaches have been used for the analysis of stochastic models of gene expression. One is based on simulations using the Gillespie algorithm (Gillesple, 1977; Gillespie, 2007), a Monte Carlo procedure for sampling trajectories described by a chemical master equation. The other approach is based on the Langevin equation, or in the more general form of stochastic differential equations (Arnold, 1974; van Kampen, 2007), which approximate the chemical master equation when the number of molecules is sufficiently large (Gillespie, 2007). In a simplified view of gene expression, the factors shaping the fluctuations expected in single cells can be classified based on three steps of the process (Kaern et al., 2005). First, the rate of transitions of the promoter between being inactive or active for transcription. Second, transcriptional bursting, the average number of mRNA that are transcribed once the promoter is in the active state. Finally, translational bursting, the average number of protein molecules that are synthesized during translation.

The two main techniques that have been used for the quantification of the expression level in each cell are flow cytometry and microscopy (Longo and Hasty, 2006), the latter especially in the form of time-lapse imaging (Locke and Elowitz, 2009). Flow cytometry restricts one to the analysis of snapshots of the expression levels in the population, while time-lapse imaging allows one to analyze the fluctuations in single cells, and, by imaging several cells, also snapshots.

The recent interest in noise in gene expression coincides with some theoretical works (McAdams and Arkin, 1997; Cook et al., 1998) analyzing stochastic models of gene expression, and the use of such models to study the lytic-lysogenic switch in λ phage (Arkin et al., 1998). Besides these early works, two other advancements likely contributed to the theoretical and experimental studies that would follow in the subsequent years. First, the growing availability of techniques for quantifying expression levels in single cells, including the popular green fluorescent protein (GFP) and its variants (Tsien, 1998). On this regard, most of the works have focused on variation in protein expression levels, although some studies have addressed variation in mRNA levels (Golding et al., 2005; Raj et al., 2006; Maamar et al., 2007; Zenklusen et al., 2008; Raj et al., 2010). The second contribution were observations in some of the first works on synthetic biology (Elowitz and Leibler, 2000; Gardner et al., 2000) suggesting that fluctuations in cellular mechanisms had measurable consequences on gene expression.

The typical approach, especially in early works, has been to analyze stochastic models of gene expression, which predict fluctuations in single cells with statistics dependent on

parameters of a model, and then compare the statistics of the levels in a snapshot of the population with values measured experimentally. Works that have relied on this approach are reviewed in section 1.1.1.1, while those that have tracked single cells throughout time are the subject of section 1.1.1.2.

1.1.1.1 Inferring the impact of fluctuations by analyzing variation in the population

In the literature, two approaches for quantifying the variation in a population have been used. The term noise is typically used to refer to the coefficient of variation of expression levels in a population of cells, given by the ratio between the standard deviation and the mean, as perhaps the most straight-forward quantification. Another approach has been to use the so-called noise strength, given by the ratio between the variance and the mean (Ozbudak et al., 2002; Blake et al., 2003; Raser and O'Shea, 2004). This corresponds to the Fano factor, and tends to reveal more clearly the impact of changing a parameter of a model of gene expression (Kaern et al., 2005).

One possible explanation for the variation in expression levels in the population is that differences in mRNA levels between different cells would be amplified via translational bursting. Consistent with this idea, an early study in *Bacillus subtilis* found that variation was affected mostly by changing the rate of translation, remaining essentially constant if changing the rate of transcription. In an early study using *Saccharomyces cerevisiae* (Blake et al., 2003), this impact of the translation rate was confirmed, therefore pointing to a contribution of mRNA fluctuations. However, a non-monotonic dependence of the noise strength on the rate of transcription was also observed (Blake et al., 2003), and was attributed to pulsative mRNA production due to transcriptional re-initiation, which would further amplify mRNA fluctuations.

An important development was the distinction between intrinsic and extrinsic noise (Elowitz et al., 2002; Swain et al., 2002). This distinction is based on an approach known as the dual-reporter method (Elowitz et al., 2002), where two copies of a given promoter, each driving the expression of a different fluorescent protein, are integrated in genetically identical cells, and the levels of the two proteins quantified in each cell. It was shown (Swain et al., 2002) that this allows for the decomposition of the total variation, as the coefficient of variation, into intrinsic and extrinsic noise. The former represents the noise inherent to the activation of each copy of the promoter and the translation of each corresponding transcript, with fluctuations in the expression of the two proteins being statistically independent. Extrinsic noise, on the other hand, is due to variation in the expression levels of factors (regulatory proteins, polymerases, ribosomes, and other cellular components) that

affect both copies. In this method, intrinsic noise is reflected into how uncorrelated are the expression levels of the two reporters in each cell, while extrinsic noise leads to correlation between their levels. In the analysis of engineered *E. coli* strains, intrinsic noise was observed to be inversely proportional to the average level of expression, following a relationship obtained from analysis of a stochastic model of gene expression (Swain et al., 2002), but extrinsic noise had a more complicated dependence. Using the dual-reporter approach in budding yeast (*Saccharomyces cerevisiae*), Raser and O'Shea (2004) showed that chromatin remodeling, which takes place during the induction of some promoters, would result in slow transitions of a promoter in between states of being inactive and active for transcription, shaping the intrinsic noise, and would be an important feature of gene expression in eukaryotes.

The results of some of these studies (Ozbudak et al., 2002; Blake et al., 2003; Raser and O'Shea, 2004) led to the view that in prokaryotes gene expression tends to take place with fast rates of transition between the inactive and activate states, such that translational bursting is the main source of fluctuations (Kaern et al., 2005). On the other hand, slow promoter transitions are expected to be common in eukaryotic cells (Raser and O'Shea, 2004; Kaern et al., 2005).

In terms of intrinsic and extrinsic noise, studies tend to point to the latter as being dominant in both prokaryotes and eukaryotes (Elowitz et al., 2002; Raser and O'Shea, 2004; Volfson et al., 2006). However, methodological biases are also possible, given the tendency to study proteins that are expressed at high levels, which are expected to be dominated by extrinsic noise (Bar-Even et al., 2006). In *S. cerevisiae*, using a promoter that had been previously reported (Raser and O'Shea, 2004) to result in expression levels dominated by extrinsic noise, it was shown (Volfson et al., 2006) that accounting for the histogram of expression levels in a population required considering two sources for such noise. The first is the unequal partitioning of proteins between the mother and daughter cell in budding yeast (Hartwell and Unger, 1977). The second source was linked to the expression levels of an upstream factor, whose fluctuations are transmitted to the downstream product being analyzed (Volfson et al., 2006).

On the issue of the distribution of protein expression levels, Shahrezaei and Swain (2008) derived, based on the chemical master equation, that a promoter that remains mainly in the active state would lead to protein levels following a negative binomial distribution (see also Paulsson and Ehrenberg, 2000), on the limit that the protein has a longer lifetime than the mRNA. For a promoter that transitions between inactive and active states, a bimodal distribution of protein levels may be obtained, if the promoter transitions are

sufficiently slow (Shahrezaei and Swain, 2008). Starting from a formulation of expression levels as a continuous variable, other studies (Friedman et al., 2006; Taniguchi et al., 2010) have derived the gamma distribution, which approximates the negative binomial distribution in the limit of high expression levels (Friedman et al., 2006; Shahrezaei and Swain, 2008; Taniguchi et al., 2010). On the other hand, Paixão (2007) studied models of gene expression based on multi-step regulatory cascades, where the levels of reactants involved in the intermediate steps would fluctuate slowly, analogously to extrinsic noise, and showed that expression levels would converge to a lognormal distribution (see also Krishna et al., 2005; Furusawa et al., 2005). In a recent genome-wide analysis of expression levels in *E. coli*, Taniguchi et al. (2010) compared the lognormal and gamma distributions, and found that the latter provides a better description for proteins expressed at low levels, while both distributions fit equally well those expressed at high levels. Bar-Even et al. (2006) also pointed to the lognormal as the best approximation highly expressed genes in *S. cerevisiae*, in comparison with the normal distribution.

Subsequent studies on a larger scale in budding yeast reported an overall inverse relationship between protein abundance and the coefficient of variation (Bar-Even et al., 2006; Newman et al., 2006). Moreover, the data from these studies are consistent with housekeeping genes tending to have small variation, while inducible genes, especially those linked to stress responses, have comparatively large cell-to-cell variation. A more recent study in *E. coli* (Taniguchi et al., 2010), also observed this scaling, for proteins present in less than 10 molecules in average per cell, while there was no scaling for proteins expressed at higher levels. This was interpreted in terms of the former being dominated by intrinsic noise, while the latter would be dominated by extrinsic noise. Moreover, in the analysis of highly abundant proteins, they observed no correlation between the levels of each protein and corresponding mRNA in each cell across the population, which is consistent with proteins being long-lived, and an effect of extrinsic noise on translation (Taniguchi et al., 2010).

1.1.1.2 Analyzing fluctuations in single cells as a function of time

Rather than trying to relate the fluctuations taking place in each cell as as function of time to the variation that is observed at the level of the population, some studies have directly tracked single cells as a function of time. Based on the representation of expression levels in a single cell as a time-series, an important question concerns the timescale, or dynamics, of the fluctuations. This is a property of the fluctuations that cannot be inferred by analyzing a single snapshot of a population. The timescale can be informally described as the average

amount of time taken for the values to change by a given quantify. A formal definition is based on properties of the auto-correlation function (Papoulis and Pillai, 2002), which for a stationary stochastic process is a function of the statistical properties of the observations in two time instants, along with the amount of time in between these two instants. The most common approach has been to quantify the timescale as the so-called auto-correlation time, which is the amount of time elapsed such that the auto-correlation function has decreased by a given percentage, typically 50% (for example, Rosenfeld et al., 2005; Austin et al., 2006; Sigal et al., 2006; Dar et al., 2012).

One of the first works (Rosenfeld et al., 2005) to quantify the timescale relied on a synthetic gene network in *E. coli*, in which extrinsic noise dominated over intrinsic noise. By relying on the dual-reported method (Elowitz et al., 2002), adapted to quantify the rates of protein production, they estimated the auto-correlation time of the intrinsic noise as being relatively short (less than 10 minutes), while the auto-correlation time for the total noise was longer (around 40 minutes), and comparable with the cell-cycle time. It was concluded that extrinsic fluctuations would be slow, resulting in a cellular memory that is relatively long-lasting. Consequently, the expression of a protein at a particular amount could persist for a duration of time close to one cell generation. They also showed that the rates of protein production throughout time were better described by a lognormal distribution than with a normal distribution (Rosenfeld et al., 2005). The question of the dynamics of the fluctuations has also been addressed by some studies using frequency domain analysis, a well-known approach for signal processing based on the Fourier transform, applied to the analysis of stochastic models of gene expression (Cox et al., 2006, 2008). By applying this analysis to a synthetic negative feedback loop in *E. coli*, Austin et al. (2006) concluded that repression of expression levels results in a reduction on the timescale of fluctuations in expression levels (higher frequencies). In that study (Austin et al., 2006), the various auto-correlation times estimated for each cell ranged from around 15 minutes to 1–3 hours in the different conditions.

In a human cancer cell line, Sigal et al. (2006) developed an approach for tagging proteins at their endogenous loci with fluorescent proteins, allowing the generation of a clonal cell library with different fusion proteins for monitoring expression levels. The auto-correlation time for the various proteins analyzed was between 16 and 50 hours (ranging from 0.8 to 2.5 generations). Therefore, in human cells the timescale may be even longer than one cell-cycle length, even in a cancer cell line that is constantly dividing, further suggesting that noise may underlie a long-lasting cellular memory. The application of the dual-reporter method to a particular ribosomal protein led to the conclusion of extrinsic

noise being the main source of variation. An additional interesting observation was that of a positive correlation between the auto-correlation time and the coefficient of variation.

The finding that fluctuations in expression levels could be long-lasting further motivated improvements in the theoretical models to simulate gene expression. One of the simplest stochastic processes to exhibit an analogous property is the Ornstein-Uhlenbeck (OU) process (Arnold, 1974). This is a Gaussian process, often referred to as colored noise, having an exponential auto-correlation function. In other words, observations made in two distinct but sufficiently close time instants are correlated. White noise, on the other hand, whose values in different time instants are uncorrelated, has an Dirac delta auto-correlation function (Papoulis and Pillai, 2002). The OU process has been one of the bases for modeling extrinsic noise (for example, Dunlop et al., 2008; Rausenberger and Kollmann, 2008; Shahrezaei et al., 2008). One of the approaches, put forward by Shahrezaei et al. (2008), is to replace a particular parameter in a model, for example the transcription rate, by a stochastic variable described by the OU process, such that the instantaneous parameter value fluctuates around a given mean with dispersion and dynamics defined by parameters of the underlying process. The Gillespie algorithm was also extended to incorporate extrinsic noise described in such a way (Shahrezaei et al., 2008).

1.1.2 Other sources of fluctuations

As outlined in the previous section, the works on noise in gene expression can be divided into two categories. In one case, grounded on stochastic models of gene expression, studies have tried to relate how these fluctuations would influence variation in expression levels across the population, and then quantify the latter. By introducing different manipulations, such as changing rates of transcription or translation, and quantifying properties based on snapshots of the population, one relates indirectly the fluctuations to the variation that is observed at the level of the population. On the other hand, some works have directly measured the fluctuations taking place in single cells over time.

However, although this suggests a relatively simple relationship between a snapshot of the population and the fluctuations in single cells throughout time, there has been some divergence. As one example, Huh and Paulsson (2011) showed that it is difficult to distinguish between the impact of noise in gene expression from variations in the number of molecules in the two daughter cells upon cell division. Therefore, it remains unclear to which degree the fluctuations that are observed are indeed entirely due to noise in gene expression.

Moreover, and more importantly, in making the connection between fluctuations taking place in single cells and the variation that is observed across the population, an implicit assumption is that every cell has the potential to express any of the levels observed in the population (Huang, 2009). However, in a general cell population, the presence of stable variants (Chang et al., 2008; Huang, 2009) would violate this assumption. Hence, in a general cell population the relationship between the fluctuations in single cells and the variation in the population is not straightforward. Therefore, potential mechanisms that might result in such variants are reviewed in the next section.

1.2 Potential mechanisms resulting in permanent differences in expression levels

As introduced in the previous section, fluctuations in the expression level of each cell across time arise, at least in part, due to noise in gene expression. A corollary of these fluctuations being stochastic is that differences in expression levels of two subsets of cells would be transient. Given the extensive repercussion of the recent works on noise in gene expression, on areas ranging from microbiology (Fraser and Kaern, 2009) to stem cell biology (Tischler and Surani, 2013), it would be tempting to assume that these stochastic fluctuations are the only factor that leads to the variation that is observed at the level of the population. Therefore, it is important to consider other mechanisms that, by introducing permanent differences in the expression levels of different subsets of cells, may contradict this assumption.

One simple scenario is that in which expression of the molecule of interest is regulated by an external signal, such as a soluble factor. If the "environment", in a spatial sense, is composed of a set of *niches* or microenvironments, each niche associated with a unique, constant level of the soluble factor, then the overall effect is that of a cell composed of stable variants for the molecule of interest (see, for example, Stout and Suttles, 2004). Consequently, one would have to account for this impact of the environment on the mapping from fluctuations taking place in each cell with the variation in expression levels observed at the level of the population. However, this is not the case under some experimental scenarios, such as some *in vitro* conditions, where cells are maintained in a homogeneous medium, where any differences in the levels of soluble factors would eventually average out. Hence, unless one is concerned with a cell type capable of secreting soluble factors, or subject to cell-cell interactions, the impact of cell-extrinsic signals can be discarded. Therefore, the remainder of this section focuses on *cell-intrinsic* mechanisms that would result in stable

variants.

Among the mechanisms that may result in permanent differences in the expression levels of two cells, genetic variation is one that comes to mind. However, since essentially all the works on noise in gene expression have relied on populations of genetically identical cells, it is also important to discuss mechanisms that may result in permanent differences even in an isogenic population.

1.2.1 Genetic variation

One possibility is that the cell population has standing genetic variation at a set of loci that influence the expression level of the molecule of interest (for example, Brem et al., 2002). In this case, genetic variation would result in different clones having different expression levels, such that there would be permanent differences in the expression levels of cells from different clones. Genetically heterogeneous cell population may be the case for microorganisms, or cells originated from a single multicellular organism. In the latter case, one may consider cancer cells (Yates and Campbell, 2012), and also T and B lymphocytes in jawed vertebrates.

The genetic diversity of lymphocytes is established by a process of somatic DNA rearrangement, termed V(D)J recombination, which is one of the hallmarks of the adaptive immune system in jawed vertebrates (Tonegawa, 1983; Market and Papavasiliou, 2003). In developing lymphocytes, the loci coding for some of the sub-units of the antigen receptors are rearranged, and hence populations of mature lymphocytes, that have completed their respective developmental processes, are inherently heterogeneous, composed of clones. This is the case for both developing B and T cells, whose antigen receptors are named, respectively, B cell receptor (BCR) and T cell receptor (TCR) (Paul, 2003). The antigen receptors interact with antigens available throughout the body, with the particular receptor expressed by a cell influencing the nature of signal that will be received upon encountering a particular antigen.

1.2.2 Epigenetic mechanisms

Even in a population of genetically identical cells there may be mechanisms that lead to permanent differences in the expression level of cells in a population, resulting in so-called stable variants (Chang et al., 2008; Huang, 2009). This is the very case of cells from multicellular organisms. In these organisms, a hallmark is the process of development, starting from the zygote, a single cell, to give rise to an adult organism. Hence, the genome of a

cell from a multicellular organism has the potential to give rise to distinct cell types, with the ultimate specification of patterns of gene expression specific to each cell type (Reik, 2007). In this sense, development is *epigenetic* (Reik, 2007), a term initially formulated by Waddington (for a historical account, see Haig, 2004; Gilbert, 2012), as the set of mechanisms that essentially link genotype and phenotype (Goldberg et al., 2007), emphasizing the differences within a single organism (Haig, 2004). Nowadays, the term *epigenetics* is often used to denote mitotically and/or meiotically heritable differences in gene expression or any other cellular phenotype that are not due to changes in DNA sequence (Haig, 2004; Goldberg et al., 2007). However, such a term is often used in a more restricted sense in molecular biology in terms of the so-called "epigenetic marks", which include non-covalent histone modifications and DNA methylation (Ledford, 2008). In this thesis, the denomination of epigenetic will be used to refer, in a broad context, to mechanisms that establish and maintain variant cell states (Jablonka and Raz, 2009). where the differences among the various states are not not explained differences by changes in DNA sequence. Therefore, this leads to the concept of epigenetic variation, as underlying phenotypic differences that persist throughout time. The qualifier of heritable may be applied to epigenetic variation, to highlight that the differences not only persist throughout time, but are also mitotically and/or meiotically heritable.

The various patterns of gene expression that are established during development are self-sustainable and heritable, such that the different types are stable, persisting throughout the lifetime of the organism (Hemberger et al., 2009; Barrero et al., 2010; Leeb and Wutz, 2012). In this context, another important concept, also credited to Waddington (see Haig, 2004), is that of an *epigenetic landscape*, a portrait of the sequence of conditions experienced by a cell (Gilbert, 2012). The landscape (figure 1.1) illustrates the hierarchical transitions taking place during development, resulting in the definition of cell fate (Goldberg et al., 2007). In establishing the patterns that are associated with different cell types, the underlying mechanisms ultimately shape the expression levels of molecules in a cell population, not only whether genes are expressed or not.

The idea that cell differentiation would be a consequence of a regulatory system composed of distinct attractors dates back to Max Delbrück, as discussed by Thomas and D'Ari (1990). Further conceptualization of the process of cellular differentiation came with an early study showing that random boolean networks of genes can be quite regular, with relatively few stable states (Kauffman, 1969). Later on, the idea of positive feedback, as a necessary condition for the existence of multiple stable states (Thomas and D'Ari, 1990), provided an important concept for bottom-up studies, relying on the standard genetics ap-

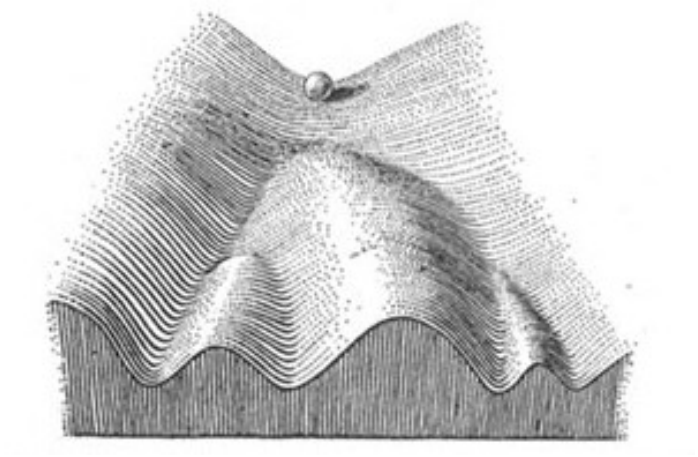

Figure 1.1: Illustration of the concept of *epigenetic landscape*, put forward by Waddington. The landscape represents the set of possible states that can be attained by a cell, symbolized by the ball. The cell starts from the top, representing an uncommitted state (such as that of a stem cell), which provides access to all the possible states. As time goes by, the ball transverses the landscape down, eventually reaching a stable, terminal, state. Figure adapted from Slack (2002).

proach of loss-of-function and gain-of-function. Indeed, the view of a cell as a dynamical system provides an immediate connection with the view of epigenetic landscape. In this way, one can refer to each cell fate or cell type as a stable cellular state, or more precisely an attractor, described by a particular profile of gene expression (Huang, 1999). Furthermore, the epigenetic landscape corresponds to the set of trajectories of the gene expression profile. This constitutes, essentially, a generalization of the property of multistability often studied in the context of small reaction networks, to the entire genome (Huang, 1999). This concept was further emphasized as genomics approaches started to be developed in the end of the 1990's, allowing for relatively straightforward analysis of genome-wide expression profiles (Huang, 1999).

In the analysis of how particular molecules and pathways interact to lead to cellular differentiation, hematopoietic development provides an attractive and very popular model system. In this context, much work has been conducted with progenitor cells that, upon stimulation, differentiate to a particular cell type. The relationship with networks of interacting genes has been further reinforced by recent experimental evidence of a cell fate as an attractor of a gene regulatory network. Using the leukemia cell line HL-60, which can be induced to differentiate into a neutrophil-like state with various stimuli, and using

microarrays for analysis, Huang et al. (2005) showed that two such conditions lead to different trajectories, but the same final pattern of genome-wide gene expression, consistent with the final state being an attractor. Using this same system (the HL-60 cell line), and the up-regulation of CD11b as a marker of differentiation to the neutrophil-like state, later on Chang et al. (2006) addressed whether the process was indicative of a single step. Taking advantage of the fact that, upon stimulation, some cells in the population do not upregulate CD11b (CD11b-low cells), these cells were isolated, using cell sorting, and further re-stimulated. It was observed that a higher percentage of these cells upregulated CD11b, in comparison with the untreated population. On the basis of a transient memory for this higher propensity to differentiate, these data were interpreted as being consistent with the existence of an intermediate state, this state being metastable, with the cells originally reverting to the original state if not re-stimulated. Overall, the view of stable cellular states as attractors is becoming increasingly popular, especially in stem cell biology (Enver et al., 2009). The idea of an attractor provides for a formal view of cell fate or type as an stable state.

One of the first works to provide a more mechanistic understanding of a gene regulatory network governing differentiation, further reinforcing the idea of cell fates as attractors focused on differentiation of hematopoietic progenitors. The differentiation of these progenitors into macrophage and neutrophils requires the transcription factors PU.1 and C/EBPα, respectively, and Laslo et al. (2006) explored the establishment of macrophage fate upon induced expression of PU.1. They showed promiscuous expression of genes associated with the macrophage and neutrophil fates during the early stages of differentiation, even at the level of single cells, but that the expression of neutrophil-associated genes is repressed as a function of time. This repression depended on transcription factors that were induced during the early stage. A mathematical model showing the interaction between the initiating factors, namely PU.1 and C/EBPα, and the induced factors showed as stable states the mixed lineage, characterized by promiscuous expression at low levels, and the terminal states associated with macrophage or neutrophil fate.

Further evidence for stable variants in the hematopoietic system comes from studies of the repopulating capacity. For example, Dykstra et al. (2007) analyzed the propagation of distinct hematopoietic differentiation programs under long-term *in vivo* conditions. In this setup, single long-term reconstituting hematopoietic stem cells are transplanted to recipient mice, and the progeny of these cells are tracked, in terms of B, T and myeloid (granulocytes/monocytes) cell populations. It was observed that the percentages of the three populations in each recipient were clustered into four different categories, with the

patterns of some of these being replicated upon new transfer of the reconstituted population to a second set of recipients. This suggests that the pattern of reconstitution characteristic of each cell is a stable, cell-intrinsic property. Expanding on this issue, recently Naik et al. (2013) used a cellular barcoding approach to analyze the ability of lymphoid-primed multipotent progenitor (LMPP, identified based on the expression of a set of markers) cells to produce different hematopoietic cell types. In this approach, cells are isolated, lentivirus-marked with a unique DNA barcode sequence from a library, and transferred to irradiated recipients. Afterwards, the sequences recovered in samples from populations of dendritic cells, B cells and myeloid cells were analyzed. It was observed that the LMPPs had very heterogeneous patterns in terms of output population produced, ranging from only one type to all types analyzed. This was not the case for hematopoietic stem cells, which were mostly capable of producing the multiple types. Interestingly, by pre-expanding a pool of barcode-marked LMPPs *in vitro*, and splitting them to transfer to a pair of recipients, they observed that the pattern of the cell types produced by each barcode was consistent in the two repicients of each pair, showing that the bias to produce a certain combination of cell types is, at least in part, an intrinsic, imprinted property of each LMPP. However, in the case of the works by Dykstra et al. (2007) and Naik et al. (2013), the underlying mechanism that explains the stable property of cells remains unclear.

These works provide evidence that mechanisms that are epigenetic, in a broad sense, may result in stable variants. However, it is important to highlight that many of these studies have mostly relied on approaches for analysis based on population averages. Therefore, while in many cases the data can and has been interpreted as being consistent with cells being in a single attractor, it does not necessarily rule out a multiplicity of such states. Moreover, another important concern is to which degree these variants are indeed stable, since some reports raise the possibility of metastability of such states.

Besides transcription factors, the association between the so-called “epigenetic marks” (especially histone modifications and DNA methylation) and gene expression has been described in the context of the specification of cell types (Reik, 2007). In eukaryotic cells, DNA is packaged, along with a set of proteins, into chromatin. Nucleosomes are composed of an octamer including two copies of each of the four canonical histones, H3, H4, H2A and H2B, and constitute an important component of this packaging, each nucleosome being wrapped around 146–147 bp of DNA (Luger et al., 1997; Richmond and Davey, 2003). Three types of epigenetic marks have been better studied: i) methylation of DNA, particularly in the case of CG-rich regions of the genome, termed CpG islands (Bird, 2002); ii) histone modifications, in which key residues of the N-terminal tails of histones are modi-

fied (Kouzarides, 2007); iii) histone variants, which replace some of the canonical histones in certain regions of the genome (Talbert and Henikoff, 2010). In the following, each one of these marks is briefly introduced.

One of the epigenetic marks that has been studied in most detail is DNA methylation (Bird, 2002), and to which a definition of epigenetics in the sense of molecular biology dates back (as discussed in Haig, 2004; Holliday, 2006). DNA methylation is often described as a mark that directly silences genes, although recent works point to a more complex picture, with the impact of its presence being dependent on the specific location in which it occurs; for example, rather than blocking, methylation in the gene body may stimulate transcription elongation (Jones, 2012). There is considerable evidence for DNA methylation being a heritable mark, an aspect often discussed in the context of so-called maintenance methylation, to refer to the overall process of replicating the pattern of DNA methylation in between cell divisions (Bird, 2002).

Histone modifications can either affect physical properties of the chromatin fiber, due to changes in charge, but also recruit, stabilize or block the binding of partners (Goldberg et al., 2007; Barrero et al., 2010). In a typical case, the modification trimethylation of histone H3 at lysine 4 (H3K4me3), associated with gene activation, with some of the partners recruited being protein complexes such as histone acetyltransferase (HAT) and ATP-dependent remodellers. This leads to histone acetylation, increasing the mobility of histones or mediating their eviction, thereby effectively opening chromatin structure (Barrero et al., 2010). On the other hand, modifications associated with gene repression, such as trimethylation of histone H3 at lysine 9 (H3K9me3) or lysine 27 (H3K27me3) ultimately lead to the recruitment of complexes that mediate histone deatecylation (HDAC) and ATP-dependent remodeling (Barrero et al., 2010). Important mediators of methylation of H3K27 and H3K4 are Polycomb group (PcG) and Trithorax (TrxG) proteins, respectively (Goldberg et al., 2007).

Finally, histone variants tend to replace the so-called canonical histones in certain regions of the genome, some well-studied variants being H3.3 (which may replace H3), H2A.Z and H2A.X (which may replace H2A), and CenH3 (also known as CENP-A in humans, and may replace H3) (Talbert and Henikoff, 2010). Variant histones can modify the structure and stability of the nucleosome, an effect that may impinge on the regulation of transcription (Banaszynski et al., 2010; Talbert and Henikoff, 2010).

Overall, a general question concerns the stability of both the underlying mechanisms (Barrero, 2012) and of terminal cell fate (Holmberg and Perlmann, 2012). This question is highlighted by the ability to reprogram a cell, as initially shown in the 1960's, by somatic

nuclear transfer in *Xenopus* (Gurdon et al., 1958; Gurdon and Melton, 2008), and also even more strikingly with the ability to reprogram terminally differentiated cell types an embryonic stem cell-like state, marked by pluripotency, in what has been known as induced pluripotent stem cells (iPSCs; Takahashi and Yamanaka, 2006).

Hence, there are mechanisms, at least in cells from multicellular organisms, with the potential to specify persistent differences in the expression levels of two subsets of cells, even in an isogenic population.In an extreme, hypothetical case, each cell of an adult organism would be essentially unique, with its particular level of gene expression. This would be analogous to the well-known case of the extreme reproducibility in the development of cell lineages in *Caenorhabditis elegans* (Emmons, 1996; Kipreos, 2005) . In the presence of these additional mechanisms, two subsets of cells that initially have distinct levels of expression may not become identical, even after a very long time.

1.3 Dynamics of expression levels in mammalian cell populations

In this section, we introduce in more details some selected works that have addressed in some detail the dynamics of expression levels, focusing on mammalian cell population. In this way, a common picture is that of epigenetic states of expression, with a corresponding potential landscape representing the transition between these metastable states.

One of the first studies focused on the expression of Sca1 in a hematopoetic progenitor cell line, named EML cells (Chang et al., 2008). To analyze the expression levels in the population as a function of time, they relied on cell sorting to isolate subsets of cells expressing different levels of Sca1, and then tracked the expression levels in these subsets as a function of time. They found that changes in the expression levels of the different subsets of cells were still taking place after more than 10 days. In the end, based on the histograms of expression levels of cells that had low, intermediate and high expression levels at the time of isolation, it was argued that all such subsets had histograms of Sca1 expression levels that were very similar to that of the original population from which they were isolated, suggesting that these subsets could reconstitute all expression levels in the original population. Furthermore, they found that this variation in the level of Sca1 was correlated with the ability of the cells to differentiate to various cell lineages upon stimulation, and argued that cells would be transiently prone to differentiate, given the slow dynamics of the fluctuations in Sca1 levels. Upon analysis of genome-wide expression levels using micro-arrays, they found that the expression levels of Sca1 were correlated with markedly different tran-

scriptional signatures, but that over time would converge to a single transcriptome. Overall, it was conclude that this was evidence of the existence of a multitude of metastable expression states, with cells randomly switching between the different states throughout time. Finally, the reconstitution of the histogram of Sca1 expression levels would be consistent with a lack of stable variants in the cell population, at least in terms of the expression of this molecule.

Another study using the EML cells (Pina et al., 2012) addressed whether the ability to self-renewal, typical of an "unprimed", multipotent state, would co-exist with the lineage-biased transcriptome in cells that exhibited such biases. They analyzed the expression of Sca1 in these cells, and found that differences in expression between cells initially isolated with high and low levels were still clear even after 2 weeks. They also pointed to a component of terminal differentiation, with some of the cells expressing low levels of Sca1 expressing a master regulator of the erythroid lineage and lacking the ability of self-renewal. Hence, they showed that, even though there is an overall tendency for transient bias, as concluded by (Chang et al., 2008), some cells in the population show markers of terminal differentiation. These cells seemed to be in an intermediate state, since they are closer, in terms of transcriptome-wide signature, to uncommitted cells, than to fully differentiated cells. To further analyze this, they quantified mRNA levels of a panel of genes in various single cells isolated, and observed marked variation of genes important for lineage commitment, suggesting that commitment can occur even without the expression of all lineage-associated genes, suggesting multiple possible trajectories for commitment. These data are consistent with the view that, once cells enter this terminal differentiation state, they are not capable anymore of expressing high Sca1 levels anymore and lose the potential to self-renew.

In parallel, observations in mouse embryonic stem cells (mESCs) suggested that the expression level of Nanog, a transcription factor involved in pluripotency, is very dynamic (Chambers et al., 2007). To study this process in more detail, a mESC cell line, referred to as TNG-A, was constructed, where a GFP reporter replaces one of the Nanog alleles. Using this cell line, Kalmar et al. (2009) showed that the overall bimodal pattern of GFP expression is very robust, with culture conditions changing merely the percentage of cells in each mode, and that cells expressing low levels of Nanog have a greater tendency to differentiate, in comparison with cells expressing high levels. In terms of the expression levels, they found that clones grown out of single cells had a strong tendency to reconstitute the original pattern observed in the initial population. They related this dynamic expression pattern of Nanog to a gene regulatory network with noise-driven excitability, modeling the

interactions between Nanog and Oct4. Another study (Luo et al., 2012) with the TNG-A cell line used cell sorting to isolate cells expressing intermediate Nanog levels to infer the parameters of a gaussian mixture model for expression levels of the population measured in different time-points. An interesting feature of this mixture model approach is that it allows the visualization of the system in terms of a potential landscape, providing an intuitive representation of trajectories of single cells subject to noise. Luo et al. (2012) relied on various experimental treatments to modulate signaling pathways, which result in different patterns of expression on the population. Besides the two states of high and low expression levels reported by Kalmar et al. (2009), they included an intermediate state, resulting in a mixture model with 3 states, where the location of each state remained the same under all experimental conditions tested, but that the frequency of cells and variance of each state were modulated.

Sisan et al. (2012) also developed an approach based on the potential landscape, applied to study the expression levels of GFP driven by the tenascin-C (an extracellular matrix protein) promoter in genetically identical fibroblasts. GFP showed a bimodal expression pattern in the population, well-approximated by a mixture of two lognormal distributions, representing two states of expression, one being GFP-negative and the other with high GFP levels. They used cell sorting to isolate cells initially expressing certain ranges of GFP levels, and flow cytometry to track GFP levels in the cells up to several days after isolation. They observed that it took more than one month for the different populations to start having similar GFP expression patterns, that would also approximate that of the original population. Since a model based on random switching between the two states, with exponential residence times in each of the two states, could describe the data reasonably well, but not the dynamics in early instants of some of the populations that were isolated. They then used a model based on the Langevin equation to describe the dynamics of expression levels in single cells, calibrated using short-term imaging. This calibration showed that the best representation of the Langevin equation by using a logarithmic transformation of the data. This model provided a much better description of the dynamics of the isolated populations, along with the corresponding potential landscape. Furthermore, by analyzing the percentage of cells falling in the range of values corresponding to the GFP-negative state, up to 100 days after cell sorting, they argued that the isolated populations would relax back to the steady-state pattern of the original population.

Therefore, many of the works considered here focus on molecules that show a bimodal expression pattern in the population, with different states of expression of the gene or reporter of interest. This itself provides evidence of the ability to form distinct states of

expression.

1.4 Aims of this thesis

Hence, considering that there are mechanisms with the potential to introduce permanent differences between the expression levels of two subsets of cells, and given that stochastic fluctuations on the expression level of a single cell are expected, at least in principle, one can consider two extreme scenarios. If fluctuations are the only factor explaining the variation observed, the differences in expression levels between two subsets of cells are transient. One of the mechanisms that is expected to introduce such fluctuations is noise in gene expression, due to small copy number effects on gene expression. On the other hand, if the fluctuations are negligible, such that the level of every cell remains completely constant across time, and each cell of the population has a unique level of expression, then the initial differences between two subsets of cells would not change. Since these two scenarios are non-exclusive, their combination constitutes a third one, in which both fluctuations and unique levels of expression are present. In this case, it is expected that two subsets of cells will become more similar across time, but never identical. Understanding which of these scenarios underlies the variation observed, termed the "origin of the variation", is one of the objectives of this work. Moreover, as long as there are some fluctuations in the level expressed by a single cell, one would expect the difference between two subsets of cells to change as a function of time, with some dynamics. Thus, another objective of this work is to address what is the amount of time needed for this change to take place, which will be referred to as the "timescale of the variation". Despite several experimental works that address, whether explicitly or not, such aspects, the fundamental questions of how to formulate and determine the origin and timescale of the variation, as stated above, remain unanswered. Consequently, overall it is not clear to which degree the variation observed is due to fluctuations in the levels of single cells or due to the presence of distinct, stable variants. Moreover, although some works have provided some estimates of the time for changes in expression levels to take place, a consistent approach for quantification is lacking.

Since the two questions are initially framed in an informal way, an important first step is to ground them more precisely. This is done in chapter 2, by developing a quantitative theoretical framework to describe the expression dynamics in a cell population. This is one important contribution of this thesis. A key starting point is the classification of the mechanisms regulating expression levels in the cell population into two components, one

stable and the other unstable. This allows for framing the the origin of the variation in terms of the contributions of the stable and unstable components. It is also the basis for deriving a model to describe protein expression in a cell population. The properties of this model are studied, resulting in the definition of the origin and timescale of the variation as two parameters describing a cell population. Building on this, an approach for quantifying these two parameters is formulated.

Given the considerable focus of experimental studies on noise in gene expression in cell populations, a further understanding of the contributions to variation in expression levels will require studying a system where there are both stable and unstable components. Hence, chapter 3 puts forward T cells (more specifically, $CD4^+$ T cells) as a model system to study these two components. This is done by quantifying the relative contribution of the stable component of variation in TCR levels in a polyclonal (genetically heterogeneous) and in two isogenic T cell populations in an *in vitro* setting, in which neither stimulation of the cells nor cell division is expected, and limited to a time frame of 3–4 days. In the isogenic populations, stable component is due, by definition, to what has been referred to as epigenetic variation in section 1.2.2 (page 12), to refer to processes leading to phenotypic differences that persist throughout time, but are not explained by differences in DNA sequence. After analysis, the description of the data considered most adequate indicates the unstable component as the main contribution in the two isogenic populations studied, but also provides preliminary evidence for the stable component in these populations. On the other hand, this description points to the stable component being the main contribution in the polyclonal population, and, indeed, differences in expression levels of subsets of cells from this population can indeed persist for a very long period of time, under *in vivo* conditions. These results establish the TCR in a polyclonal population of $CD4^+$ T cells as a model system to study how the stable and unstable components contribute to variation in expression levels.

In summary, this thesis develops a conceptual and theoretical framework to address the origin and timescale of the variation in expression levels. In this context a particularly important concept is that of a stable component of variation, which results in permanent differences between the expression levels of two subsets of cells. This framework is then used to analyze the mechanisms that shape expression of the T cell receptor in $CD4^+$ T cells. Finally, chapter 4 presents the general discussion of the work developed in this thesis, in the context of attractors in gene regulatory networks, variation in expression levels in T cells and expression levels as quantitative traits.

In terms of the overall organization of the thesis, a brief note is in order. Given the

interdisciplinary nature of the work developed here, detailed theoretical derivations and details of experimental methods and data are presented in the respective methods sections and appendices, so as to improve the readability of the text across a mixed audience.

1.5 Mathematical notation

Throughout this thesis, $\log(\cdot)$ denotes the natural logarithm. Random variables are represented as bold symbols, as in $\boldsymbol{x}$. Vectors are represented as $\vec{v}$, and defined as column vectors (i.e., $\vec{v} \in \Re^{n \times 1}$). Moreover, a vector-valued random variable is denoted as $\vec{\boldsymbol{x}}$.

Let $\mathbb{E}[\boldsymbol{x}]$ denote the expected value of a random variable $\boldsymbol{x}$, $\mathbb{V}[\boldsymbol{x}]$ the variance and $\mathbb{C}[\boldsymbol{x}, \boldsymbol{y}]$ the covariance between the random variables $\boldsymbol{x}$ and $\boldsymbol{y}$. Moreover, let $\mathbb{K}[\boldsymbol{x}] = \sqrt{\mathbb{V}[\boldsymbol{x}]}/\mathbb{E}[\boldsymbol{x}]$ denote the coefficient of variation of $\boldsymbol{x}$. Being a normalized quantity, the coefficient of variation may be presented as a percentage, for convenience.

Let $\boldsymbol{y} \sim \mathcal{N}(\mu_y, \sigma_y)$ denote a normally-distributed random variable, having probability density function:

$$f(y) = \frac{1}{\sqrt{2\,\pi}\,\sigma_y}\,\exp\left(-\frac{1}{2\,\sigma_y^2}\,(y-\mu_y)^2\right) \tag{1.1}$$

and expected value μ_y and variance σ_y^2. Then, $\boldsymbol{z} = \exp(\boldsymbol{y})$ follows a lognormal distribution with parameters μ_z and σ_z, denoted as $\boldsymbol{z} \sim \mathcal{LN}(\mu_z, \sigma_z)$, with density (Papoulis and Pillai, 2002):

$$f(z) = \frac{1}{\sqrt{2\,\pi}\,\sigma_z\,z}\,\exp\left(-\frac{1}{2\,\sigma_z^2}\,(\log(z)-\mu_z)^2\right) \tag{1.2}$$

Bibliography

Arkin, A., Ross, J., and McAdams, H. H. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. *Genetics*, 149(4):1633–48.

Arnold, L. (1974). *Stochastic differential equations: theory and applications*. Wiley, 1 edition.

Austin, D. W., Allen, M. S., McCollum, J. M., Dar, R. D., Wilgus, J. R., Sayler, G. S., Samatova, N. F., Cox, C. D., and Simpson, M. L. (2006). Gene network shaping of inherent noise spectra. *Nature*, 439(7076):608–11.

Banaszynski, L. A., Allis, C. D., and Lewis, P. W. (2010). Histone variants in metazoan development. *Developmental cell*, 19(5):662–74.

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y., and Barkai, N. (2006). Noise in protein expression scales with natural protein abundance. *Nature Genetics*, 38(6):636–43.

Barrero, M. J. (2012). The stability of the induced epigenetic programs. *Comparative and Functional Genomics*, 2012:434529.

Barrero, M. J., Boué, S., and Izpisúa Belmonte, J. C. (2010). Epigenetic mechanisms that regulate cell identity. *Cell Stem Cell*, 7(5):565–70.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & Development*, 16(1):6–21.

Blake, W. J., Kaern, M., Cantor, C. R., and Collins, J. J. (2003). Noise in eukaryotic gene expression. *Nature*, 422(6932):633–7.

Brem, R. B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–5.

Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L., and Smith, A. (2007). Nanog safeguards pluripotency and mediates germline development. *Nature*, 450(7173):1230–4.

Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E., and Huang, S. (2008). Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, 453(7194):544–7.

Chang, H. H., Oh, P. Y., Ingber, D. E., and Huang, S. (2006). Multistable and multistep dynamics in neutrophil differentiation. *BMC Cell Biology*, 7:11.

Cook, D. L., Gerber, A. N., and Tapscott, S. J. (1998). Modeling stochastic gene expression: implications for haploinsufficiency. *PNAS*, 95(26):15641–6.

Cox, C. D., McCollum, J. M., Allen, M. S., Dar, R. D., and Simpson, M. L. (2008). Using noise to probe and characterize gene circuits. *PNAS*, 105(31):10809–14.

Cox, C. D., McCollum, J. M., Austin, D. W., Allen, M. S., Dar, R. D., and Simpson, M. L. (2006). Frequency domain analysis of noise in simple gene circuits. *Chaos*, 16(2):026102.

Dar, R. D., Razooky, B. S., Singh, A., Trimeloni, T. V., McCollum, J. M., Cox, C. D., Simpson, M. L., and Weinberger, L. S. (2012). Transcriptional burst frequency and burst size are equally modulated across the human genome. *PNAS*, 109(43):17454–9.

Dunlop, M. J., Cox, R. S., Levine, J. H., Murray, R. M., and Elowitz, M. B. (2008). Regulatory activity revealed by dynamic correlations in gene expression noise. *Nature Genetics*, 40(12):1493–8.

Dykstra, B., Kent, D., Bowie, M., McCaffrey, L., Hamilton, M., Lyons, K., Lee, S.-J., Brinkman, R., and Eaves, C. (2007). Long-term propagation of distinct hematopoietic differentiation programs in vivo. *Cell Stem Cell*, 1(2):218–29.

Elowitz, M. B. and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–8.

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584):1183–6.

Emmons, S. W. (1996). Simple worms, complex genes. *Nature*, 382(6589):301–2.

Enver, T., Pera, M., Peterson, C., and Andrews, P. W. (2009). Stem cell states, fates, and the rules of attraction. *Cell Stem Cell*, 4(5):387–97.

Fraser, D. and Kaern, M. (2009). A chance at survival: gene expression noise and phenotypic diversification strategies. *Molecular Microbiology*, 71(6):1333–40.

Friedman, N., Cai, L., and Xie, X. (2006). Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression. *Physical Review Letters*, 97(16):1–4.

Furusawa, C., Suzuki, T., Kashiwagi, A., Yomo, T., and Kaneko, K. (2005). Ubiquity of log-normal distributions in intra-cellular reaction dynamics. *Biophysics*, 1:25–31.

Gardner, T. S., Cantor, C. R., and Collins, J. J. (2000). Construction of a genetic toggle switch in Escherichia coli. *Nature*, 403(6767):339–42.

Gilbert, S. F. (2012). Commentary: 'The epigenotype' by C.H. Waddington. *International Journal of Epidemiology*, 41(1):20–3.

Gillespie, D. T. (2007). Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58:35–55.

Gillesple, D. T. (1977). Exact Stochastic Simulation of Coupled Chemical Reactions. *The Journal of Physical Chemistry*, 93555(1):2340–2361.

Goldberg, A. D., Allis, C. D., and Bernstein, E. (2007). Epigenetics: a landscape takes shape. *Cell*, 128(4):635–8.

Golding, I., Paulsson, J., Zawilski, S. M., and Cox, E. C. (2005). Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–36.

Gurdon, J. B., Elsdale, T. R., and Fischberg, M. (1958). Sexually mature individuals of Xenopus laevis from the transplantation of single somatic nuclei. *Nature*, 182:64–65.

Gurdon, J. B. and Melton, D. A. (2008). Nuclear reprogramming in cells. *Science*, 322(5909):1811–5.

Haig, D. (2004). The (dual) origin of epigenetics. *Cold Spring Harbor Symposia on Quantitative Biology*, 69:67–70.

Hartwell, L. H. and Unger, M. W. (1977). Unequal division in Saccharomyces cerevisiae and its implications for the control of cell division. *The Journal of Cell Biology*, 75(2 Pt 1):422–35.

Hemberger, M., Dean, W., and Reik, W. (2009). Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. *Nature Reviews Molecular Cell Biology*, 10(8):526–37.

Holliday, R. (2006). Epigenetics: a historical overview. *Epigenetics*, 1(2):76–80.

Holmberg, J. and Perlmann, T. (2012). Maintaining differentiated cellular identity. *Nature Reviews Genetics*, 13(6):429–39.

Huang, S. (1999). Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. *Journal of Molecular Medicine*, 77(6):469–80.

Huang, S. (2009). Non-genetic heterogeneity of cells in development: more than just noise. *Development*, 136(23):3853–62.

Huang, S., Eichler, G., Bar-Yam, Y., and Ingber, D. E. (2005). Cell Fates as High-Dimensional Attractor States of a Complex Gene Regulatory Network. *Physical Review Letters*, 94(12):128701.

Huh, D. and Paulsson, J. (2011). Non-genetic heterogeneity from stochastic partitioning at cell division. *Nature Genetics*, 43(2):95–100.

Jablonka, E. and Raz, G. (2009). Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *The Quarterly Review of Biology*, 84(2):131–76.

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–92.

Kaern, M., Elston, T. C., Blake, W. J., and Collins, J. J. (2005). Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6(6):451–64.

Kalmar, T., Lim, C., Hayward, P., Muñoz Descalzo, S., Nichols, J., Garcia-Ojalvo, J., and Martinez Arias, A. (2009). Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biology*, 7(7):e1000149.

Kauffman, S. (1969). Homeostasis and differentiation in random genetic control networks. *Nature*, 224(5215):177–8.

Kipreos, E. T. (2005). C. elegans cell cycles: invariance and stem cell divisions. *Nature Reviews Molecular Cell Biology*, 6(10):766–76.

Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, 128(4):693–705.

Krishna, S., Banerjee, B., Ramakrishnan, T. V., and Shivashankar, G. V. (2005). Stochastic simulations of the origins and implications of long-tailed distributions in gene expression. *PNAS*, 102(13):4771–6.

Laslo, P., Spooner, C. J., Warmflash, A., Lancki, D. W., Lee, H.-J., Sciammas, R., Gantner, B. N., Dinner, A. R., and Singh, H. (2006). Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell*, 126(4):755–66.

Ledford, H. (2008). Language: Disputed definitions. *Nature*, 455(7216):1023–8.

Leeb, M. and Wutz, A. (2012). Establishment of epigenetic patterns in development. *Chromosoma*, 121(3):251–62.

Locke, J. C. W. and Elowitz, M. B. (2009). Using movies to analyse gene circuit dynamics in single cells. *Nature Reviews Microbiology*, 7(5):383–92.

Longo, D. and Hasty, J. (2006). Dynamics of single-cell gene expression. *Molecular Systems Biology*, 2:64.

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, 389(6648):251–60.

Luo, Y., Lim, C. L., Nichols, J., Martinez-Arias, A., and Wernisch, L. (2012). Cell signalling regulates dynamics of Nanog distribution in embryonic stem cell populations. *Journal of the Royal Society, Interface*, 10(78):20120525.

Maamar, H., Raj, A., and Dubnau, D. (2007). Noise in gene expression determines cell fate in Bacillus subtilis. *Science*, 317(5837):526–9.

Market, E. and Papavasiliou, F. N. (2003). V(D)J recombination and the evolution of the adaptive immune system. *PLoS Biology*, 1(1):E16.

McAdams, H. H. and Arkin, A. (1997). Stochastic mechanisms in gene expression. *PNAS*, 94(3):814–9.

Naik, S. H., Perié, L., Swart, E., Gerlach, C., van Rooij, N., de Boer, R. J., and Schumacher, T. N. (2013). Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature*, pages 2–6.

Newman, J. R. S., Ghaemmaghami, S., Ihmels, J., Breslow, D. K., Noble, M., DeRisi, J. L., and Weissman, J. S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–6.

Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D., and van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. *Nature Genetics*, 31(1):69–73.

Paixão, T. (2007). *The Stochastic Basis of Somatic Variation*. PhD thesis, University of Porto.

Papoulis, A. and Pillai, S. U. (2002). *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 4 edition.

Paul, W. E. (2003). *Fundamental Immunology*. Lippincott Williams & Wilkins, 5 edition.

Paulsson, J. and Ehrenberg, M. (2000). Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Physical Review Letters*, 84(23):5447–50.

Pina, C., Fugazza, C., Tipping, A. J., Brown, J., Soneji, S., Teles, J., Peterson, C., and Enver, T. (2012). Inferring rules of lineage commitment in haematopoiesis. *Nature Cell Biology*, 14(3):287–94.

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, 4(10):e309.

Raj, A., Rifkin, S. A., Andersen, E., and van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. *Nature*, 463(7283):913–8.

Raser, J. M. and O'Shea, E. K. (2004). Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678):1811–4.

Rausenberger, J. and Kollmann, M. (2008). Quantifying origins of cell-to-cell variations in gene expression. *Biophysical Journal*, 95(10):4523–8.

Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143):425–32.

Richmond, T. J. and Davey, C. A. (2003). The structure of DNA in the nucleosome core. *Nature*, 423(6936):145–50.

Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S., and Elowitz, M. B. (2005). Gene regulation at the single-cell level. *Science*, 307(5717):1962–5.

Shahrezaei, V., Ollivier, J. F., and Swain, P. S. (2008). Colored extrinsic fluctuations and stochastic gene expression. *Molecular Systems Biology*, 4(196):196.

Shahrezaei, V. and Swain, P. S. (2008). Analytical distributions for stochastic gene expression. *PNAS*, 105(45):17256–61.

Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Liron, Y., Rosenfeld, N., Danon, T., Perzov, N., and Alon, U. (2006). Variability and memory of protein levels in human cells. *Nature*, 444(7119):643–6.

Sisan, D. R., Halter, M., Hubbard, J. B., and Plant, A. L. (2012). Predicting rates of cell state change caused by stochastic fluctuations using a data-driven landscape model. *PNAS*, 109(47):19262–7.

Slack, J. M. W. (2002). Timeline: Conrad Hal Waddington: the last Renaissance biologist? *Nature Reviews Genetics*, 3(11):889–895.

Stout, R. D. and Suttles, J. (2004). Functional plasticity of macrophages: reversible adaptation to changing microenvironments. *Journal of Leukocyte Biology*, 76(3):509–13.

Swain, P. S., Elowitz, M. B., and Siggia, E. D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. *PNAS*, 99(20):12795–800.

Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–76.

Talbert, P. B. and Henikoff, S. (2010). Histone variants–ancient wrap artists of the epigenome. *Nature Reviews Molecular Cell Biology*, 11(4):264–75.

Taniguchi, Y., Choi, P. J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X. S. (2010). Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–8.

Thomas, R. and D'Ari, R. (1990). *Biological Feedback*. CRC Press.

Tischler, J. and Surani, M. A. (2013). Investigating transcriptional states at single-cell-resolution. *Current Opinion in Biotechnology*, 24(1):69–78.

Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature*, 302(5909):575–81.

Tsien, R. Y. (1998). The green fluorescent protein. *Annual Review of Biochemistry*, 67:509–44.

van Kampen, N. G. (2007). *Stochastic Processes in Physics and Chemistry*. Elsevier, 3 edition.

Volfson, D., Marciniak, J., Blake, W. J., Ostroff, N., Tsimring, L. S., and Hasty, J. (2006). Origins of extrinsic variability in eukaryotic gene expression. *Nature*, 439(7078):861–4.

Yates, L. R. and Campbell, P. J. (2012). Evolution of the cancer genome. *Nature Reviews Genetics*, 13(11):795–806.

Zenklusen, D., Larson, D. R., and Singer, R. H. (2008). Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural & Molecular Biology*, 15(12):1263–71.

Chapter 2

A THEORETICAL FRAMEWORK FOR QUANTIFYING THE CONTRIBUTIONS TO VARIATION IN EXPRESSION LEVELS

Author contributions

The theoretical formulation and simulations done in this chapter were planned by the author of the thesis and the supervisor Dr. Jorge Carneiro, with input from the supervisor Dr. Vasco Barreto. The author performed all the theoretical analysis and simulations. The results were analyzed and interpreted by the author and the supervisors Dr. Vasco Barreto and Dr. Jorge Carneiro.

This chapter is an extended version of the theoretical framework presented in the following manuscript:

Guzella, T. S., Barreto, V. B., and Carneiro J. (2013). Quantifying the Contributions and Dynamics Underlying Variation in Expression Levels in a Cell Population. *In preparation, under final review by the co-authors*

Abstract

The analysis and quantification of variation in expression levels in cell populations is currently the focus of intense investigation. However, it remains unclear how to quantify the contributions to this variation in the presence of stable variant. In this work, we develop a theoretical framework, centered on stable and unstable components, to quantify properties describing heterogeneous cell populations. It is shown that the isolation of a subset of cells from a starting population, such as via cell sorting, provides for estimating two parameters. The first is the relative contribution of the stable component, termed R^2_α. The second parameter, denoted as τ_T, and referred to as the characteristic time of the variation, accounts for the dynamics of the expression levels in a population that has been isolated. The present work provides a solid framework for quantifying the contributions to variation in expression levels, and paves the way for analyzing how different mechanisms modulate each component.

Chapter 2

2.1 Introduction

Recent works on noise in gene expression (Elowitz et al., 2002; Raser and O'Shea, 2004; Pedraza and van Oudenaarden, 2005; Sigal et al., 2006; Raj et al., 2010), reviewed in (Raj and van Oudenaarden, 2008), have brought to light that, even in isogenic organisms and cell populations, one can observe variation in expression levels, at least in part due to the small copy number of molecules involved in the process. Hence, the expression level of a molecule would fluctuate asynchronously as a function of time in each cell of a population, reflecting the inherent stochasticity of gene expression. Consequently, a snapshot taken in a single instant of time of the level of a particular molecule in cells of the population, would therefore show variation in the amount that is expressed.

However, noise in gene expression constitutes merely one of the several mechanisms that might underlie variation in expression levels of a molecule, even though most of the quantitative approaches developed up to this date have focused on noise in gene expression as the only mechanism explaining the variation observed (for example, Swain et al., 2002; Pedraza and van Oudenaarden, 2005; Sigal et al., 2006; Dunlop et al., 2008; Munsky et al., 2009; Komorowski et al., 2013). Consequently, it remains unclear to which degree stable variants in the cell population contribute to the variation. By stable variants (Chang et al., 2008), we refer to subsets of cells that are biased, by whichever mechanism, to have a limited range of expression levels, compared with those observed in the population. One example would be the case of a genetically diverse cell population, such as tumors (Yates and Campbell, 2012) and cells of the immune system (Market and Papavasiliou, 2003), where each clone could be associated with a particular expression level. Besides this, especially in the case of cells from multicellular organisms, "epigenetic" mechanisms may have a similar effect. The notion of epigenetic mechanism, and hence that of epigenetic variation in the population, is used in a broad sense to refer to processes such as differentiation stages or cell lineages (Orkin, 2000; Hemberger et al., 2009), that lead to phenotypic differences that are not explained by differences in DNA sequence, but that persist throughout time.

A case of particular interest is a study focusing on the expression of Sca1 in an isogenic hematopoetic cell line (Chang et al., 2008). It has been argued (Chang et al., 2008) that subsets of cells tend to reconstitute the histogram of expression levels of Sca1 of the starting population, albeit with very slow dynamics. In principle, complete reconstitution would be consistent with a lack of stable variants in the population, at least in terms of expression levels of Sca1. However, recent evidence (Pina et al., 2012) shows that, even after 2 weeks, this reconstitution is not fully complete. More importantly, these authors

(Pina et al., 2012) showed that some cells in this hematopoetic cell line express markers indicative of terminal differentiation, and have limited capacity for cell division. This points to an additional component of heterogeneity in the population. An important aspect is that these works have relied mainly on qualitative comparisons of histograms of expression levels in order to compare cell populations, without a rigorous approach for quantification. Consequently, it is not clear how to analyze such data, such that the degree to which the reconstitution takes place remains unclear. Moreover, it becomes difficult to test the impact of experimental manipulations, which may influence the expression levels in a cell population in multiple ways. Finally, a quantitative approach is also important in order to provide more formal concepts on which to ground subsequent studies analyzing expression levels in cell populations.

To quantify the properties of a cell population, this work lumps the molecular mechanisms regulating expression levels in a cell population into two components, one stable and another unstable. In informal terms, the stable component leads to permanent differences between the expression levels of two subsets of cells, while the unstable component, on the other hand, represents transient differences in the expression levels, which will vanish over time. This allows us to study the inference in terms of how these two components influence the expression levels in the population.

This chapter is organized in the following way: in order to formalize the concept of the stable and unstable components, section 2.2 first decomposes the mean and variance of expression levels of a population in terms of the properties of groups of cells in that population. Section 2.3 then uses a simplified model of constitutive protein expression to describe expression levels in a population with both the stable and unstable components. In this model, the stable component arises due to the fact that different groups of cells have different average rates of protein production throughout time. Afterwards, section 2.4 defines a single parameter, termed R^2_α, which summarizes the contribution of the stable and unstable components to variation in expression levels in a cell population. It then considers approaches for estimating R^2_α based on the isolation (physical separation) of cells and analysis of their expression levels. Section 2.5 then studies more precisely how to quantify R^2_α when multiple cells are isolated. In this context, the characteristic time of the variation is defined as a parameter τ_T, related to the dynamics of changes of the expression levels in the cells that have been isolated. Finally, section 2.6 presents the conclusions of this chapter.

2.2 Partitioning the contributions to variation in expression levels

We assume that a biological cell population, hereafter referred to as full population, is a mixture of sub-populations. In this way, each cell belongs to a single sub-population throughout all time, without switching between sub-populations. Using a mixture model formulation, each sub-population is indexed by $i = 1, 2, \ldots, N$, and described by three parameters (μ_i, v_i, w_i): the mean μ_i and variance v_i of expression levels, and the relative frequency w_i of cells in the full population that belong to this sub-population. The latter is given by:

$$w_i = \frac{n_i}{\sum_{j=1}^{N} n_j} \tag{2.1}$$

where n_i is the number of cells in the i-th sub-population and $\sum_{j=1}^{N} n_j$ is the total number of cells in the full population.

The parameters (μ_i, v_i, w_i) describing a sub-population are taken as random variables $(\boldsymbol{\mu}, \boldsymbol{v}, \boldsymbol{w})$ (see methods for details of the notation used) following a particular multivariate distribution. Then, one can relate the mean μ_F and variance v_F of expression levels of the full population to the properties of the sub-populations, as detailed in appendix 2.A. In the limit of large N, provided that there is no covariance between the frequencies $(\boldsymbol{w})$ and either the means $(\boldsymbol{\mu})$, the squared means $(\boldsymbol{\mu^2})$ and the variances $(\boldsymbol{v})$ of the sub-populations, it follows that (appendix 2.A):

$$\mu_F = \mathbb{E}[\boldsymbol{x}] = \mathbb{E}[\boldsymbol{\mu}] \tag{2.2}$$

$$v_F = \mathbb{V}[\boldsymbol{x}] = \underbrace{\mathbb{E}[\boldsymbol{v}]}_{\substack{\text{Expected variance} \\ \text{within each} \\ \text{sub-population}}} + \underbrace{\mathbb{V}[\boldsymbol{\mu}]}_{\substack{\text{Variance among} \\ \text{the means of the} \\ \text{sub-populations}}} \tag{2.3}$$

where the subscript F is used to highlight that these are properties of the full population. Therefore, under these conditions, the mean of the full population is simply the expected value of the means of the sub-populations, while the variance of the full population is partitioned into the expected variance within the sub-populations $(\mathbb{E}[\boldsymbol{v}])$ and the variance among the means of the sub-populations $(\mathbb{V}[\boldsymbol{\mu}])$.

It is important to highlight that equations 2.2 and 2.3 are independent of the precise definition of a sub-population. However, the two terms in equation 2.3 suggest a specific definition, in which only the unstable component is present in each sub-population. In

this way, the term of the expected variance within a sub-population becomes the expected contribution of the unstable component to the variance of the full population, while the variation among the means of the sub-populations is the contribution of the stable component. These constitute general definitions, in the context of the decomposition of the variance of the full population. The derivation of a model of protein expression in a full population then becomes straightforward. As will be adopted in section 2.3, expression levels within each sub-population will be described by a stochastic model, while the different sub-populations will have different means based on one of the parameters of the stochastic model.

2.3 A model for constitutive protein expression in a heterogeneous cell population

In this section, a particular instance of the representation of a full population outlined in section 2.2 is derived. The expression levels of cells in cells belonging to a sub-population are described by a stochastic model (section 2.3.1), representing the unstable component (variation within a sub-population). The final model is then obtained in section 2.3.2, by having each sub-population with a different mean, controlled by one of the parameters of the stochastic model. This leads to variation among sub-populations, representing the stable component.

2.3.1 Introducing variation within a sub-population

The stochastic model for protein expression considered here is based on previous work (Shahrezaei et al., 2008), and is defined by the two equations:

$$dx_t = \left\{ \alpha \exp\left(y_t - \frac{1}{2}\sigma^2 \right) - \frac{1}{\beta} x_t \right\} dt \tag{2.4}$$

$$dy_t = -\frac{1}{\tau} y_t \, dt + \frac{\sigma}{\sqrt{\tau/2}} dW_t \tag{2.5}$$

where x_t is the amount of protein expressed at time t, and y_t, which influences protein levels, is a stochastic variable following the Ornstein-Uhlenbeck process. In equation 2.5, W_t is the Wiener process (Arnold, 1974). The parameters for the model are presented in table 2.1, along with their respective units. In the following, the interpretation of these parameters is presented in more detail.

Parameter	Description	Unit
α	Average rate of production	Amount of molecules / time
β	Mean lifetime of the protein	Time
σ	Normalized dispersion of the rate of production	Non-dimensional
τ	(Approximate) Auto-correlation time of the rate of production, related to the dynamics of changes in the instantaneous rate of protein production in single cells	Time

Table 2.1: Description of the parameters of the stochastic model of protein expression defined by equations 2.4 and 2.5.

The equation governing dx_t has two terms, the rate of production:

$$\alpha \exp\left(y_t - \frac{1}{2}\sigma^2\right) \tag{2.6}$$

which depends on the stochastic process y_t, and protein degradation:

$$x_t/\beta \tag{2.7}$$

following first-order kinetics with rate $1/\beta$, β being the mean protein lifetime. A model with a similar overall structure has recently been reported (Singh et al., 2012), in which mRNA transcription and degradation have also been explicitly incorporated. Equation 2.4 can be re-written as:

$$\frac{dx_t}{dt} = \alpha\, z_t - \frac{1}{\beta}x_t \tag{2.8}$$

where z_t, defined as:

$$z_t = \exp\left(y_t - \frac{1}{2}\sigma^2\right) \tag{2.9}$$

denotes the instantaneous normalized rate of protein production, which has unity expected value. All processes governing protein production (promoter transitions, transcription and translation, among others) are lumped together into the average rate α and the instantaneous normalized rate given by z_t. The representation in equation 2.8, highlighting the contribution of lumped upstream factors, has been applied early on in the analysis of mod-

els of stochastic gene expression (for example, Pedraza and van Oudenaarden, 2005; Sigal et al., 2006). Equation 2.8 denotes that, in a single cell, the instantaneous rate of protein production is proportional to the instantaneous levels of these lumped upstream factors, and fluctuates as a function of time, with auto-correlation time approximately equal to τ (Shahrezaei et al., 2008). As illustrated in figure 2.1A, parameter τ is related to the dynamics of fluctuations, or changes, in the instantaneous normalized rate of protein production z_t: small values of τ lead to z_t having fluctuations with "fast" dynamics, while greater values of τ lead to "slower" changes. These dynamics are defined independently of the instantaneous rates that are observed in a snapshot of cells from a given sub-population (figure 2.1B), which are determined by parameter σ. The fluctuations in the instantaneous rate of protein production are then propagated downstream, resulting in fluctuations in protein levels, with dynamics dictated by τ (through z_t) and β. For simplicity, protein degradation is assumed to be deterministic, with the same rate $1/\beta$ for all cells. Therefore, the protein level in a single cell is a function of the stochastic process representing production, whose dynamics are dictated by parameter τ, and the expected lifetime of the protein once expressed, defined by parameter β. In this model, β determines how fast the instantaneous protein level will change in response to changing the instantaneous rate of protein production. If the protein is relatively short-lived, such that $\beta \ll \tau$, the instantaneous level quickly changes when the instantaneous rate of production is altered. On the other hand, in the limit of the protein being extremely long-lived, such that $\beta \gg \tau$, the instantaneous level remains constant, determined simply by the time-averaged rate of production, even though the instantaneous rate of production is constantly changing. Therefore, the model envisages the range of biological scenarios ranging from a very long-lived ($\beta \gg \tau$) to a very short-lived protein ($\beta \ll \tau$), relative to the dynamics of changes in the instantaneous rate of production.

Finally, it follows from equation 2.9 that:

$$\boldsymbol{z_t} \sim \mathcal{LN}\left(-\frac{1}{2}\sigma^2, \sigma\right),\ t \to \infty \tag{2.10}$$

and therefore the stationary rate of protein production follows a lognormal distribution in cells of a sub-population, consistent with a report of lognormal rates of protein expression (Rosenfeld et al., 2005). Equations 2.4 and 2.5 constitute a simple model that generates, for a wide range of parameter values, a lognormal-like distribution of protein levels, compatible with the widespread observation of the lognormal distribution in cell populations (Paixão, 2007). In this scenario, in terms of the log-transformed protein levels (appendix

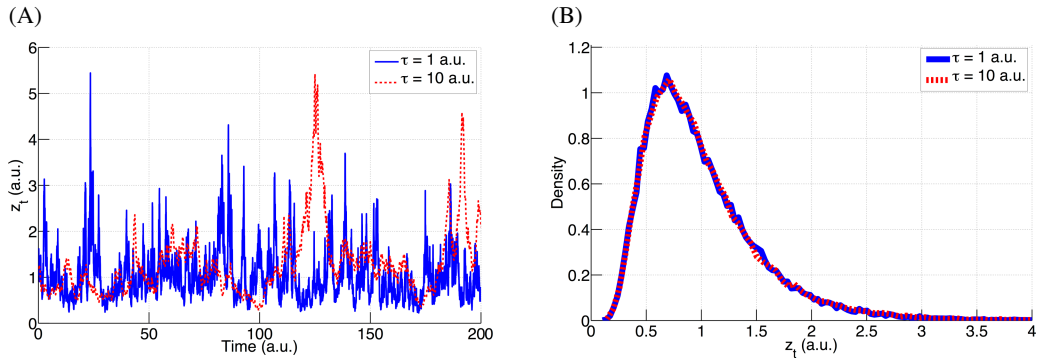

Figure 2.1: Illustration of the instantaneous normalized rate of protein production (z_t), for two different values of τ (for $\tau = 1$ and $\tau = 10$). **(A)** Comparison between two realizations of the stochastic process z_t (considering $\sigma = 0.5$), representing the instantaneous normalized rate of protein production in two single cells, for $\tau = 1$ (full blue line) and $\tau = 10$ (dashed red line). **(B)** Histograms of the stationary ($t \to \infty$) values of z_t (40000 realizations), considering $\sigma = 0.5$, for $\tau = 1$ (full blue line) and $\tau = 10$ (dashed red line). The histograms were normalized so as to have unity area, thereby providing estimates of the stationary probability density functions for z_t. Values of time and the parameters τ are shown in arbitrary units.

2.B), the mean and variance of a stationary sub-population are given by equations 2.11 and 2.12, respectively:

$$\mu_{\log} = \mathbb{E}\left[\log\left(\boldsymbol{x_t}\right)\right] = \log\left(\alpha\,\beta\right) - \frac{1}{2}\,\sigma_W^2 \tag{2.11}$$

$$v_{\log} = \mathbb{V}\left[\log\left(\boldsymbol{x_t}\right)\right] = g(\sigma^2, \tau/\beta) = \sigma_W^2 \tag{2.12}$$

where the subscript W will be used hereafter to denote that the variation is due to the stochastic process influencing the instantaneous rate of protein production. The dependence of equation 2.12 on the ratio τ/β follows from the analysis done in appendix 2.C. In equation 2.12, $g(\cdot,\cdot)$ is an arbitrary function, which can be estimated via simulation. In particular, this function reflects the previously mentioned scenario that, for $\tau \ll \beta$, such a very long-lived protein essentially buffers the fluctuations in the instantaneous rate of protein production. Therefore, for constant σ, σ_W will decrease if τ/β also decreases. Because of this effect, we restrict the analysis to the cases of $0.1 \leq \tau/\beta \leq 10$. In this sense, the limit $\tau/\beta = 0.1$ lumps all biological cases of a very long-lived protein, and similarly $\tau/\beta = 10$ envisages very short-lived proteins.

2.3.2 Combining with variation among sub-populations

As discussed in section 2.2, the stable component arises due to variation in the means of the sub-populations. Therefore, we assume that parameter α in equation 2.4 is distributed in the full population, becoming a random variable, denoted by $\boldsymbol{\alpha}$. Consequently, each sub-population is described by one value of α, resulting in different average rates of production, and hence different mean expression levels.

For simplicity, we consider the case that:

$$\boldsymbol{\alpha} \sim \mathcal{LN}\left(\mu_\alpha, \sigma_\alpha\right) \tag{2.13}$$

For the i-th sub-population, having parameter α_i, the mean and variance then follow from equations 2.11 and 2.12:

$$\mu_{i,\log} = \log\left(\alpha_i\,\beta\right) - \frac{1}{2}\,\sigma_W^2 \tag{2.14}$$

$$v_{i,\log} = \sigma_W^2 \tag{2.15}$$

where σ_W^2 is assumed, also for simplicity, to be the same for all sub-populations. In terms of log-transformed values, plugging equations 2.11 and 2.12 into equation 2.3, one obtains the variance of the full population:

$$v_{F,\log} = \sigma_T^2 = \underbrace{\sigma_W^2}_{\substack{\text{Variance due to}\\ \text{the unstable}\\ \text{component}}} + \underbrace{\sigma_\alpha^2}_{\substack{\text{Variance due to}\\ \text{the stable}\\ \text{component}}} \tag{2.16}$$

An important property of equation 2.16, which is based on log-transformed values, is that the parameters that represent the variances due to the stable and unstable components (σ_α^2 and σ_W^2, respectively) remain separate. This is a key feature, greatly simplifying the process of analysis and inference throughout this work. Therefore, in this way the stable component arises due to variation in the time-averaged rate of protein production of the different sub-populations, more precisely due to the variance of the log-transformed time-averaged rates of protein production. As detailed in appendix 2.D, the equivalent of equation 2.16 considering protein levels without any transformation has an additional term, dependent on σ_α^2 and σ_W^2. This additional term arises since the variance of each sub-population in this case depends on the value of α. For this reason, we consider, from this point on, the analysis based on log-transformed values only.

2.4 Isolating cells to quantify the contributions to the variation in a cell population

The previous section showed that, for the lognormal model of protein expression, the variance of the full population is simply the sum of the variances due to the stable and unstable components. This section considers some approaches for estimating the contributions of these two components to the variation. The first step is the definition of R^2_α, which simplifies the problem of quantifying the contributions, which is then reduced to the estimation of a single parameter. Afterwards, strategies for estimating R^2_α are considered, based on the isolation (physical separation) of cells and analysis of their expression levels. Hereafter, the full population whose value of R^2_α will be estimated is referred to as starting population, while the term "isolated population" denotes a set of one or more cells that have been isolated from the starting population.

2.4.1 Defining the relative contribution of the stable component

Section 2.3.2 showed that the variance of log-transformed expression levels of the full population is simply the sum of variances due to the stable and unstable components (equation 2.16). In this context, in analogy with the R^2 quantification of the variance explained by a linear regression model, we define R^2_α as:

$$R^2_\alpha = \frac{\sigma^2_\alpha}{\sigma^2_T},\ 0 \leq R^2_\alpha \leq 1 \tag{2.17}$$

to denote the proportion of the observed variance that is explained by the stable component.

Hence, R^2_α formalizes and quantifies the relative contribution of the stable component to the total variance of the full population. Quantifying the contribution of the stable and unstable components is therefore reduced to estimating a single parameter. In the case of $R^2_\alpha = 0\%$, variation in expression levels arises due to the unstable component alone; conversely, the stable component explains all the observed variation if $R^2_\alpha = 100\%$. Finally, in the intermediate case $0\% < R^2_\alpha < 100\%$, a combination of the two components is at play.

2.4.2 Strategies for estimating the relative contribution of the stable component based on isolating cells

After defining R^2_α, strategies for its estimation are considered. Since the starting population is assumed to be heterogeneous, being composed of several sub-populations, a natural

approach for estimation is to isolate a subset of cells, quantify a property of expression levels in this subset, and compare the value to that of the starting population. To simplify the presentation, it is assumed that the property is quantified based on a sufficiently large number of cells, such that sampling effects are neglected, to a first approximation.

This section discusses two basic strategies for isolation, which have been used by previous experimental works (Chang et al., 2008; Huang, 2009; Kalmar et al., 2009; Pina et al., 2012; Sisan et al., 2012): single (section 2.4.2.1) or multiple cells (section 2.4.2.2). In the latter approach, which will be the focus of the remainder of this work, the question of which property needs to be quantified in order to estimate R_α^2 is addressed in section 2.5.

2.4.2.1 Isolating single cells

One strategy is to simply separate the full population into the sub-populations that compose it. This can be achieved by isolating a single cell, and allowing it to expand sufficiently, so as to use a sufficiently large number of cells for the quantifications. As defined in the framework, the resulting expanded population, denoted by the subscript d, is the sub-population to which the original cell belongs to, and therefore all variation in the expanded population is due to the unstable component ($\sigma_{T,d}^2 = \sigma_W^2$). Hence, in this case, one has the condition $\sigma_{\alpha,d}^2 = 0$, and R_α^2 may be estimated by considering the variance of the starting population (σ_T^2):

$$R_\alpha^2 = 1 - \frac{\sigma_{T,d}^2}{\sigma_T^2} \tag{2.18}$$

As a further refinement, one may replicate the process, by isolating several single cells and allowing them to expand in parallel. Then, the value of σ_α^2 can be estimated by taking the variance of the means of the expanded populations (as in the derivation of equation 2.16). Since the variance of each expanded population is σ_W^2, R_α^2 can be estimated based on the definition in equation 2.17, without relying on an explicit estimate of σ_T^2. On the other hand, if equation 2.18 is used, it is expected that the values of R_α^2 estimated out of the expanded populations considered will be all identical, otherwise arguing against the assumption of the sub-populations in the full population all having identical variances equal to σ_W^2. An advantage of this approach is that it can properly deal with this case of different variances, provided that a sufficiently large number of expanded populations is analyzed.

In either case, a fundamental limitation of this approach is that it can only be used with a cell population that can self-reconstitute efficiently from single cells, such as cell lines, for example. Moreover, it does not provide any information on the dynamics of the

expression levels. Hence, we consider an alternative, based on the isolation of multiple cells.

2.4.2.2 Isolating multiple cells

A second strategy relies on the isolation of multiple cells, based on their expression levels. More specifically, the isolated population corresponds to cells between percentiles p_1 and p_2 of expression levels of the starting population. Without loss of generality, it is assumed hereafter that $p_1 < p_2$. By isolating multiple cells, one sacrifices the simplicity of obtaining a single sub-population, since all sub-populations that have cells in the range defined by p_1 and p_2 are isolated, but may perform inference in non-dividing cell populations. A commonly used technique for isolation based on expression levels is fluorescence-activated cell sorting (FACS).

In a hypothetical experiment, let a time reference t be defined beginning from the instant of isolation. Since the estimation is done based on a property of expression levels, in principle the isolation can be done based on any two percentiles p_1 and p_2, as long as it does not constitute simply random sampling of cells from the starting population. Therefore, the two percentiles should satisfy $0\% \leq p_1 < p_2 < 100\%$ or $0\% < p_1 < p_2 \leq 100\%$. This ensures that at least one of the isolated populations, at time $t = 0$, is not identical to the starting population. For such an isolated population, three outcomes are possible. If only the unstable component is present ($R^2_\alpha = 0\%$), after waiting a sufficiently long amount of time, the isolated population will become identical to the starting population. On the other hand, if the observed variation is fully explained by the stable component alone ($R^2_\alpha = 100\%$), the isolated population will not change as a function of time, remaining identical to just after being isolated, and always different from the starting population. Finally, if both the stable and unstable components are present in the starting population ($0 < R^2_\alpha < 1$), the isolated population will change as a function of time after isolation, but without ever becoming identical to the starting population.

Besides being applicable to even non-dividing cell types, an additional interesting feature of this strategy is that it allows for also considering the dynamics of changes in expression levels in the isolated population. By definition, these changes are related to the dynamics of the unstable component, as the expression levels in the sub-populations that have been isolated relax to their stationary values. To capture this notion, the timescale of the variation will be referred to as a quantity that is related to this relaxation time, needed to reach stationarity. However, when isolating multiple cells, it is unclear which property of the expression levels of an isolated population, such as the mean or variance, needs to

be quantified in order to estimate R^2_α. Moreover, the choice of the specific approach for isolation, in terms of the percentiles p_1 and p_2, needs to be considered. These key aspects are studied in the next section.

2.5 Estimating the relative contribution of the stable component

To study which property of the expression levels needs to be quantified in order to estimate R^2_α, this section relies on simulations of the process of isolating cells, and then following the temporal dynamics of expression levels. Since all derivations are based on equation 2.16, the analysis herein relies on log-transformed values of protein levels. In the simulations, protein expression levels are described by the model derived in section 2.3. Therefore, at least in principle, the approach for inference is only applicable to this particular model. Finally, for simplicity, cell division is neglected.

The isolated population considered at first for inference here is composed of the 10% of cells with the highest expression levels in the starting population (percentiles $p_1 = 90\%$ and $p_2 = 100\%$, following the notation of section 2.4.2), hereafter referred to as "high expressors". The choice of 10% is arbitrary, and is deemed to represent, at least in principle, a good compromise between resolution and number of cells obtained. Moreover, the use of the "all expressors", which are a random sample of the starting population, provides a reference. The inference of R^2_α is initially addressed via an asymptotic (stationary) property of high expressors, and subsequently extended to a time-dependent formulation, to allow considering information on the dynamics of the expression levels. In these analyses, the central results for the quantification of the contribution of the stable component are derived.

2.5.1 Analysis of the means of the isolated populations

We first address how the asymptotic (stationary) means of the isolated populations are related to the value of R^2_α. The analysis first focuses on the dynamics of the mean of log-transformed protein levels, shown in figure 2.2 for the high expressors, along with the reference values of the all expressors. Since σ^2_T is constant in figure 2.2, all isolated populations have the same initial mean, and this value remains constant throughout time in the case of $R^2_\alpha = 100\%$. For $R^2_\alpha = 0\%$, the mean eventually becomes equal to that of the starting population, while in the intermediate case $0\% < R^2_\alpha < 100\%$, the asymptotic mean reaches a value between these two extremes.

To analyze this relationship more closely, the asymptotic (stationary) mean of log-

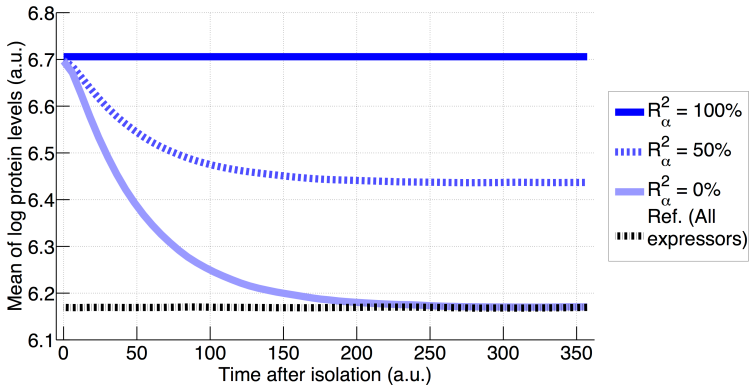

Figure 2.2: Mean (log values) of "high expressors" after isolation as 10% of starting populations with different values of R_α^2, but constant σ_T^2. Results are shown for $\tau = 50$, $\beta = 5$ and $\sigma_T = 0.3$.

transformed protein levels is considered as a function of R_α^2 (figure 2.3) for various isolated populations, including also "low expressors" (10% of cells with the lowest expression levels; $p_1 = 0\%$ and $p_2 = 10\%$). Since the mean of either high or low expressors at time instant 0 is independent of R_α^2, and for $R_\alpha^2 = 100\%$ remains constant as a function of time, the linear relationship in figure 2.3 allows one to define a simple approach for estimating R_α^2. Let $\mu_C(t)$ denote the mean of log-transformed expression levels of an isolated population C, at time t. Defining $\Delta_{A,B}(t)$ as the difference between the means of log-transformed values of two isolated populations A and B, respectively, at time instant t:

$$\Delta_{A,B}(t) = \mu_A(t) - \mu_B(t) \tag{2.19}$$

then R_α^2 can be estimated via:

$$R_\alpha^2 = \frac{\lim_{t\to\infty} \Delta_{A,B}(t)}{\Delta_{A,B}(0)},\ \Delta_{A,B}(0) \neq 0 \tag{2.20}$$

The condition $\Delta_{A,B}(0) \neq 0$ for using equation 2.20 implies that the two isolated populations being compared must have different means just after isolation ($t = 0$).

Equation 2.20 is analogous to the breeder's equation, an important result in quantitative genetics relating the response of a population of individuals to selection acting on a single trait, given the so-called narrow-sense heritability of that trait (Lynch and Walsh, 1998). The breeder's equation is commonly used in the context of animal breeding and so-called artificial selection, in which certain individuals of a given population are selected

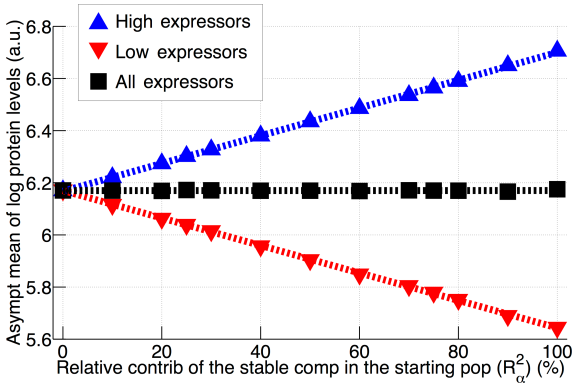

Figure 2.3: Asymptotic (stationary) mean expression levels (log values) of high and low expressors, isolated in the simulations as 10% of the starting population, and also of all expressors. The symbols represent the values obtained from the simulations, while the dotted lines represent the fitting of a straight line, to confirm the linear relationship between the asymptotic mean and R^2_α. Results are shown for $\tau = 50$, $\beta = 5$ and $\sigma_T = 0.3$.

to reproduce, based on their values of a particular trait of interest, such as meat yield (Hill and Kirkpatrick, 2010). This provides an analogy for the approach of isolating cells, in which cells are chosen ("selected") based on their expression levels, constituting the isolated populations referred to as high and low expressors, and the remaining cells being discarded.

From the inequality in equation 2.17, an additional relationship for $\Delta_{A,B}(t)$ holds:

$$\lim_{t\to\infty} \left(\Delta_{A,B}(t)\right) \leq \Delta_{A,B}(0) \tag{2.21}$$

Therefore, the stationary difference between the means of log-transformed expression levels of the isolated populations A and B is expected to be, under the present formulation, lower than or equal to the difference just after isolation.

This formulation has an important consequence in terms of experimental design. In order to maximize resolution in the estimation of R^2_α, one should maximize the value of $\Delta_{A,B}(0)$. For a given percentage of cells that are isolated (the difference $p_2 - p_1$; see section 2.4.2), this is obtained by comparing high and low expressors. Therefore, a key result is that, to estimate R^2_α, one may simply calculate the ratio between the initial and asymptotic difference between the means of log-transformed protein levels in these two isolated populations. Consequently, the remainder of this work focuses on this case, by always relying on the function $\Delta_{H,L}(t)$ for estimation. As shown in figure 2.7 (appendix 2.F), the

linear relationship between $\lim_{t\to\infty}\Delta_{H,L}(t)$ and R^2_α holds for essentially all parameter values of τ/β and σ_T considered, with the effect of increasing the latter being simply that of increasing the initial difference $\Delta_{H,L}(0)$.

2.5.2 A time-dependent formulation for estimation based on the means

To consider the dynamics of expression levels in an isolated population, we introduce the time-dependent function $\Omega_{H,L}(t)$ given by:

$$\Omega_{H,L}(t) = \frac{\Delta_{H,L}(t)}{\Delta_{H,L}(0)},\ \Delta_{H,L}(0) \neq 0 \tag{2.22}$$

Being based on the means of log-transformed values of two populations that have been isolated, $\Delta_{H,L}(t)$ follows an approximately exponential decay (figure 2.4; see appendix 2.E for a rationale). Using the approximation of exponential decay, and defining the characteristic time of the variation as τ_T, we shall write that:

$$\Omega_{H,L}(t) = \underbrace{R^2_\alpha}_{\substack{\text{Relative contribution of the stable component in the starting population}}} + \underbrace{(1-R^2_\alpha)}_{\substack{\text{Relative contribution of the unstable component in the starting population}}} \underbrace{\exp\left(-t/\tau_T\right)}_{\substack{\text{Relaxation of the unstable component (timescale term)}}} \tag{2.23}$$

such that $\Omega_{H,L}(t)$ converges asymptotically to R^2_α. The characteristic time is therefore undefined in the case of $R^2_\alpha = 100\%$, since $\Delta_{H,L}(t)$ does not change as a function of time after isolation. Since the characteristic time τ_T is related to the time needed for the initial difference $\Delta_{H,L}(0)$ to approach the asymptotic value $\lim_{t\to\infty}\left\{\Delta_{H,L}(t)\right\}$, it provides a formal notion of the timescale of the variation.

To analyze the relationship between the characteristic time of the variation τ_T and the parameters of the model, a detailed simulation was conducted. Figure 2.5 summarizes the results of these simulations, by comparing the ratio between τ_T and the sum $\tau+\beta$, as a function of τ/β, for various values of R^2_α and τ_T. Altogether, the data in figure 2.5 show that the sum $\tau+\beta$ provides a lower bound on τ_T:

$$\tau_T \geq \tau + \beta \tag{2.24}$$

Therefore, if the mean lifetime of the protein (β) is known, and τ_T is estimated experimentally by fitting $\Delta_{H,L}(t)$, one can use equation 2.24 to obtain an upper bound on τ,

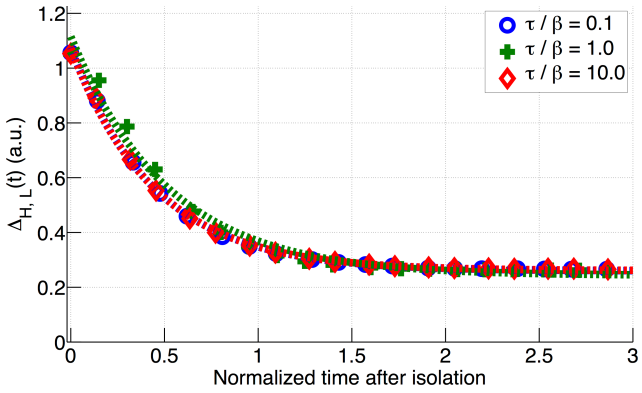

Figure 2.4: The function $\Delta_{H,L}(t)$ decays with approximately exponential dynamics. Simulations of the isolation of cells were done, for various values of τ and β, with $R_\alpha^2 = 25\%$ and $\sigma_T = 0.3$. Shown are simulation results (symbols), along with the results of fitting the model of exponential decay $\Delta(t) = a + b\,\exp(-t/\tau_T)$ to the simulation data (dotted lines), where a and b are constants. Time is normalized in each case by the instant t^* such that $\Delta_{H,L}(t^*)$ has decayed by 90%.

the auto-correlation time of the rate of protein production. The increasing values of the ratio $\tau_T/(\tau+\beta)$, with $\tau/\beta < 1$, for increasing σ_T and decreasing R_α^2, is associated with the increase in parameter σ (see equation 2.12). In the case of $\sigma_T \leq 0.3$, which corresponds to a coefficient of variation of protein levels around 30% (equation 2.54, appendix 2.B), consistent with reported values for some molecules in mammalian cells (Sigal et al., 2006; Feinerman et al., 2008), the relationship between τ_T and parameters τ and β can be approximated, with a bias of at most 5%, as:

$$\tau_T \approx \tau + \beta,\ \sigma_T \leq 0.3 \tag{2.25}$$

The relative contribution of the stable component (R_α^2) and the characteristic time of the variation (τ_T) can be visualized in a single plot, derived from equation 2.23. As shown in figure 2.6, R_α^2 corresponds to the asymptotic value of $\Omega_{H,L}(t)$, while τ_T corresponds to the instant of time that satisfies:

$$1 - \Omega_{H,L}(\tau_T) = (1 - \exp(-1))\ \left(1 - R_\alpha^2\right) \approx 0.63\ \left(1 - R_\alpha^2\right) \tag{2.26}$$

Since equation 2.23 features an exponential decay, it follows that the plateau is reached after an instant of time t satisfying $t \geq 3\,\tau_T$, as evident in figure 2.6. Furthermore, the

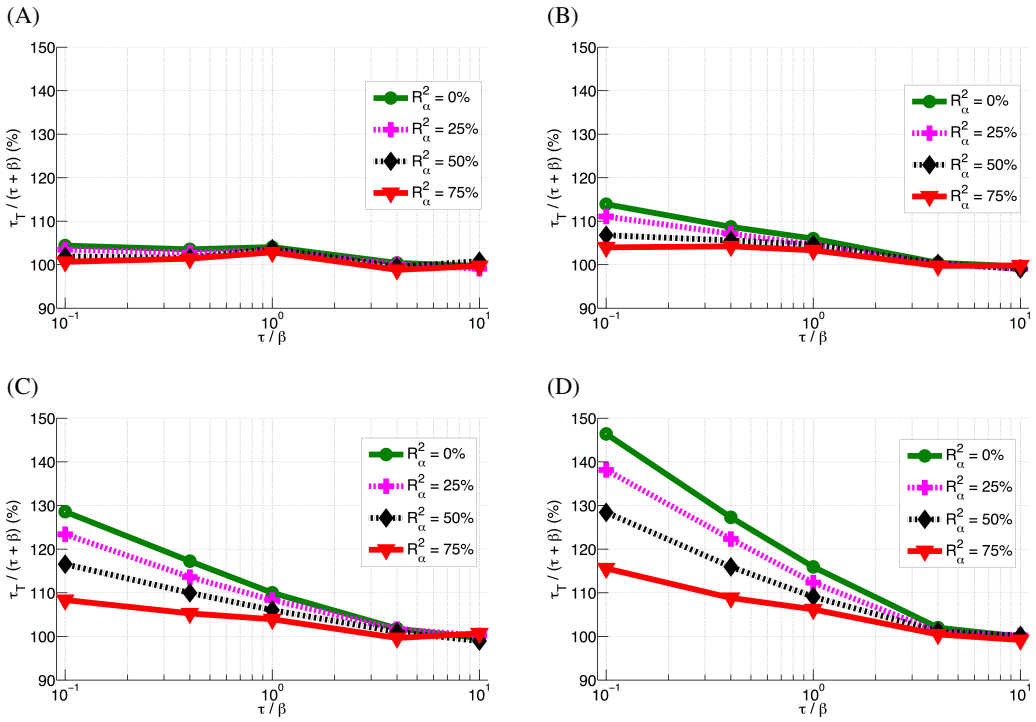

Figure 2.5: Comparison between $\tau + \beta$ and the value estimated for τ_T. Simulated data ($\Delta_{H,L}(t)$) was fit under the same setup as in figure 2.4, and the resulting values of τ_T are compared with the value of $\tau + \beta$. The results are shown in terms of the ratio $\frac{\tau_T}{\tau+\beta}$, expressed as a percentage, for $\sigma_T = 0.3$ (A), $\sigma_T = 0.5$ (B), $\sigma_T = 0.75$ (C), and $\sigma_T = 1$ (D).

inequality in equation 2.21 becomes:

$$\Delta_{H,L}(t) \leq \Delta_{H,L}(0)\ \forall t \tag{2.27}$$

since function $\Delta_{H,L}(t)$ is monotonically non-increasing with time for an exponential decay.

An interesting property arises from the definition of function $\Omega_{H,L}(t)$ in equation 2.23. The random variables representing the untransformed expression levels of cells in the populations of high and low expressors are denoted as $\boldsymbol{x_t^{(H)}}$ and $\boldsymbol{x_t^{(L)}}$, respectively. Assuming that the two isolated populations have lognormal-like distributions in time instant t (ap-

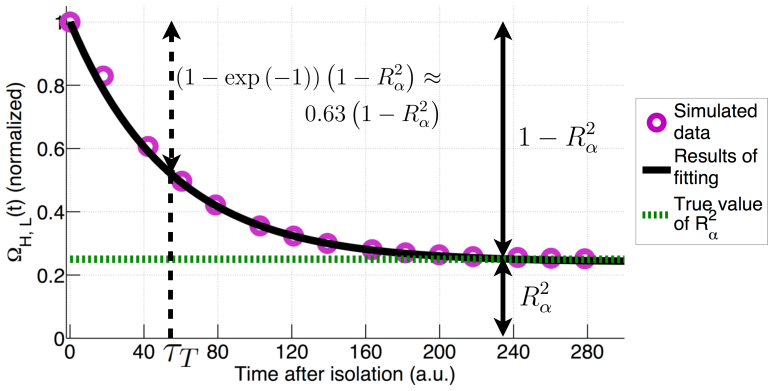

Figure 2.6: Illustration of function $\Omega_{H,L}(t)$. Shown are simulation results (symbols), with $R_\alpha^2 = 25\%$, $\tau = 50$, $\beta = 5$ and $\sigma_T = 0.3$, which were fit to the expression for $\Omega_{H,L}(t)$ in equation 2.23 (continuous line). The figure also includes the true value of R_α^2, shown as the horizontal dotted line, and the value of τ_T, as given by equation 2.26. The function $\Omega_{H,L}(t)$ shown was obtained by fitting $\Delta_{H,L}(t)$ as $k\,\Omega_{H,L}(t)$, where k is a scaling factor, defined as an additional parameter for the fitting.

pendix 2.B), it follows that:

$$\begin{aligned}\Delta_{H,L}(t) &= \mu_H(t) - \mu_L(t) \\ &\approx \log\left(\mathbb{E}\left[\boldsymbol{x_t^{(H)}}\right]\right) - \frac{1}{2}\,\sigma_H^2(t) - \log\left(\mathbb{E}\left[\boldsymbol{x_t^{(L)}}\right]\right) + \frac{1}{2}\,\sigma_L^2(t) \\ &= \log\left(\mathbb{E}\left[\boldsymbol{x_t^{(H)}}\right] \Big/ \mathbb{E}\left[\boldsymbol{x_t^{(L)}}\right]\right) - \frac{1}{2}\left(\sigma_H^2(t) - \sigma_L^2(t)\right)\end{aligned} \tag{2.28}$$

where $\sigma_H(t)$ and $\sigma_L(t)$ are the standard deviation of log-transformed values of high and low expressors, respectively. In the case that high and low expressors are isolated as similar percentiles of the starting population, we would expect that $\sigma_H^2(t) \approx \sigma_L^2(t)$ due to symmetry, and therefore:

$$\Delta_{H,L}(t) \approx \log\left(\mathbb{E}\left[\boldsymbol{x_t^{(H)}}\right] \Big/ \mathbb{E}\left[\boldsymbol{x_t^{(L)}}\right]\right) \tag{2.29}$$

Hence, function $\Delta_{H,L}(t)$ can be approximated by the logarithm of the ratio between the means of untransformed values of the two populations.

A corollary follows for comparing high and low expressors, to test if, after some time t^*, they have become identical, in terms of expression of the molecule of interest. In the present formulation, this will only take place if $R_\alpha^2 = 0$, such that the means of log-transformed values of the two populations should be indistinguishable for $t > t^*$

($\Delta_{H,L}(t) = 0, t > t^*$). In other cases one may be interested in comparing two biological populations, from which high and low expressors are isolated, to ask whether these two biological populations can be described by the same value of R^2_α. In this case, model selection approaches, such as the Akaike Information Criterion (AIC; Burnham and Anderson, 1998), may be used.

Alternatively, by estimating $\Omega_{H,L}(t^\star)$ for a given time instant $t^\star > 0$, it is possible to constrain the range of values for R^2_α. It follows from the definition of function $\Omega_{H,L}(t)$ (equation 2.22) that:

$$\Omega_{H,L}(t^\star) \geq R^2_\alpha \geq 0,\ t^\star > 0 \tag{2.30}$$

and, therefore, $\Omega_{H,L}(t^\star)$ provides an upper bound on R^2_α. The equality is included in the first condition $\Omega_{H,L}(t^\star) \geq R^2_\alpha$ in equation 2.30 as a slightly more conservative bound, accounting for effective numerical convergence of $\Omega_{H,L}(t^\star)$ to R^2_α for $t^\star \gg \tau_T$, given the very fast decay of the exponential term in this case. This analysis using $\Omega_{H,L}(t^\star)$ to constrain the expected values of R^2_α may be particularly useful when analyzing data gathered in a relatively limited window of observation.

Although this section has focused on the case in which high and low expressors are used, all the properties derived also hold for any two isolated populations A and B. The only requirement is that the condition $\Delta_{A,B}(0) \neq 0$ is satisfied. Finally, the variances of the isolated populations can be used to estimate an additional property. As detailed in appendix 2.G, they allow one to estimate the ratio between the values of σ^2_α in the starting and isolated populations. However, this estimate turns out to be biased, tending to under-estimate the true value by up to 20%. Hence, as discussed in appendix 2.G, since this quantity can be inferred via simulations, after estimating the value of R^2_α, the variances of the isolated populations are considered, at least in principle, to be uninformative in practice.

2.6 Discussion

This chapter developed an approach to quantify the contributions and the dynamics shaping variation in expression levels in a cell population. The framework partitions the total variance into contributions due to a stable and an unstable component. These two components can be considered as analogous to what has been referred (Huang, 2009) to as temporal and population noise, respectively. However, we believe the terms stable and unstable components constitute a more intuitive description of each aspect. Potential mechanisms which would be part of the stable component include genetic variation, and epigenetic variation.

In terms of the unstable component, noise in gene expression (Raj and van Oudenaarden, 2008) is one possibility. Stable variants, arising due to the stable component, may be very common, since differentiation stages or cell lineages are hallmarks of cells from multicellular organisms (Orkin, 2000; Hemberger et al., 2009). In contrast, most of the quantitative approaches developed recently (Elowitz et al., 2002; Swain et al., 2002; Sigal et al., 2006; Rausenberger and Kollmann, 2008; Rinott et al., 2011; Komorowski et al., 2013) focus on noise in gene expression as the only underlying mechanism. In the present work, the relative contribution of the stable component to the observed variation is formulated as a single parameter R^2_α, and the dynamics of the variation are represented by the characteristic time of the variation τ_T. This timescale is related to the mean time for relaxation of the unstable component.

The stable and unstable components were initially defined in an informal way, in terms of the impact on the ability of a subset of cells to reconstitute the distribution of a starting population. The first step of this work, studying the partitioning of the contributions to the variance of expression levels in a cell population divided into sub-populations, allowed for a more concrete and precise definition of these components. Based on this, a sub-population of cells was defined such that variation in expression levels within each sub-population is due to the unstable component only, and that the stable component in the cell population arises from variation in the means of the sub-populations. Protein expression levels within a sub-population were described using a stochastic model (Shahrezaei et al., 2008), with a stochastic rate of protein production and deterministic, first-order, protein degradation. This model results in a lognormal distribution of expression levels in a sub-population, and was extended to describe the cell population by assuming, for simplicity, that the means of the various sub-populations also follow a lognormal distribution (Paixão, 2007). In this approach, the analysis based on log-transformed values emerged as the best approach, as the contributions due to the stable and unstable components are additive, greatly simplifying the process of inference. This is particularly relevant for flow cytometry data, which is typically analyzed in a logarithmic scale. It is interesting to note that a recent work (Sisan et al., 2012) also relied on log-transformed values for quantification, a transformation derived from the analysis of properties of time-series of expression levels in individual cells.

The estimation of R^2_α was framed based on a setup of isolating (physically separating) cells and analyzing their expression levels as a function of time. In this formulation, an important conclusion of the present work is that the rigorous quantification of R^2_α can be done based on the difference between the means of isolated populations. The estimates

of this difference in each time interval are normalized by the value just after sorting, and such a time-dependent formulation allowed for the definition of the characteristic time τ_T. The normalization by the value at time zero, but using the squared coefficient of variation of a single population, has been recently used (Singh et al., 2012), under the assumption that all observed variation is due to noise in gene expression. In this way, these authors used (Singh et al., 2012) transcription inhibition to assess whether stochasticity in mRNA production/degradation, or promoter fluctuations contribute to noise in protein expression. The normalization of the differences by the initial value ($t = 0$) in the present work formalizes the definition of how much of the initial difference, which is introduced by the process of sorting, is still present at a particular point in time (function $\Omega(t)$). Hence, a key requirement is that the isolated populations being compared have different means just after sorting. This points to using high and low expressors as the basis for quantification, as a way to maximize resolution, given a particular percentile for isolation of cells (for example, 10%). On the other hand, at least based on the approach developed here, it is not possible to perform inference by isolating cells from the range of intermediate expression levels, having equal mean but different variance (in terms of log-transformed values) from the starting population.

Moreover, the fact that estimation is based on the mean expression levels has interesting implications. A key conclusion of several recent works on variation in cell populations is that many of the standard experimental techniques mask heterogeneity in the response of cells to a given stimulus, since these techniques provide population-averaged readouts (as discussed by Huang, 2009; Spencer and Sorger, 2011). In the scope of the present work, one may combine the isolation of cells, as the only step relying on analysis at the single-cell level, with such population-averaged techniques to quantify the origin and timescale of the variation. This arises from equation 2.29, where function $\Delta(t)$, which is at the core of the estimation process, can be approximated as the logarithm of the fold-ratio between the means of raw values of the two populations. Therefore, this approximation can be used to estimate the values of R^2_α for multiple proteins in cells that have been isolated from a starting population, using genome- (with mRNA levels as proxies of protein levels) or proteome-wide approaches. In terms of methods for analysis of single-cells, we also showed that the variances of the isolated populations can be further informative, allowing the estimation of the ratio between the absolute values of the contribution of the stable component in the isolated and in the starting population. However, the estimate obtained in this way is biased, under-estimating the true value by up to 20%. Consequently, if an estimate of this ratio is needed, a simulation-based approach is suggested.

We note that the quantification of statistics of expression levels of a cell population has been done in the context of time-lapse imaging, in which Sigal et al. (2006) devised a measure of the degree to which a cell attains all expression levels observed in the population (see also Cohen et al., 2009). This measure is determined by ranking the cells based on their expression levels in the beginning of the experiment, and then quantifying the proportion of ranks attained by each cell and its progeny, such that a value equal to one corresponds to a certain cell occupying all ranks in the population. That approach has the advantage of providing information on each cell lineage that grows throughout the experiment and, being non-parametric, is advantageous from a statistical point of view, being robust to the distributions of expression levels of cells in each lineage. However, it is not clear how to relate the value estimated to how the expression level is regulated in a cell and in the population, and this approach requires that individual cells and progeny are tracked as a function of time. Indeed, such a setup of time-lapse imaging is intrinsically constrained in terms of the duration of the analysis, besides requiring the expression of fluorescent reporters for quantification. The analysis of snapshots of a population, which has been considered in this work, does not necessarily require fluorescent reporters for quantification to be done, and provides for analysis on arbitrary time frames, albeit at the cost of not allowing for analysis of each cell lineage in the population.

Although cell division was neglected for simplicity in the analysis of section 2.5, the parameterization in terms of the model of constitutive protein expression allows us to derive some general expectations in this case. The inclusion of cell division would lead to changes in the expression level of each cell, depending on its position in the cell cycle, which would add up to the fluctuations due to the stochastic rate of protein expression. Provided that cell division is completely asynchronous in the cells of each sub-population, the added temporal impact would be averaged out, therefore merely changing the parameter values describing each sub-population and, consequently, the full population. Therefore, the notion of the relative contribution of the stable component, quantified by R_α^2, and the characteristic time of the variation, quantified by τ_T, may be extended to the analysis of cells that are dividing. In this context, it could be interesting to study more precisely the quantification of the impact of cell division on the properties of a cell population.

In order to simplify the subsequent analysis, one assumption that was made in section 2.2 is that of negligible covariances between the frequencies of cells from the cell population that belong to each sub-population and the parameters describing the expression levels in the sub-populations (namely, the means, squared means and variances of expression levels). In this way, the variance of expression levels in a cell population was shown

to be given simply as a function of the means and variances of expression levels of the sub-populations. If any of the covariances are not negligible, additional terms (see appendix 2.A) appear in the equation for the variance of expression levels in a cell population, as a function of the parameters of the sub-populations. In order to investigate how to perform inference in this general case, further studies are necessary.

The framework developed may offer interesting perspectives for quantitative studies using heterogeneous cell populations. In particular, by providing a rigorous approach to quantify the relative contribution of the stable component, this work provides one basis for determining to which degree this component contributes to variation in different experimental systems (Chang et al., 2008; Kalmar et al., 2009; Pina et al., 2012). Furthermore, the model for expression levels considered here may be further extended to incorporate, for example, more elaborated formulations, such as those with positive and feedback loops, as in the case of gene regulatory networks regulating cell differentiation (for example, MacArthur et al., 2012). This would likely imply studying how to perform inference when expression levels within each sub-population are described by a stochastic model that differs from the simplified formulation considered in section 2.3.1 (which leads to a lognormal distribution of expression levels within each sub-population). Another aspect that may be revised is how the stable component comes about, which was considered in section 2.3.2 as being due to variation in the time-averaged rates of protein production of the different sub-populations, with the rates following a lognormal distribution. The quantification of the contributions to variation in expression levels in cell populations may provide an important step towards a detailed understanding of how different molecular mechanisms modulate the stable and unstable components.

In summary, this chapter developed a theoretical quantitative framework for inferring two parameters that describe the variation in expression levels. One of these parameters is R_α^2, the relative contribution of the stable component to the observed variation, which quantifies the degree to which stable differences in the expression levels of two subsets of cells are maintained. The other parameter, τ_T, the characteristic time of the variation, is related to the mean time for transient differences to disappear. In the context of a simple model of constitutive protein expression, we showed that these two parameters can be inferred by isolating cells and quantifying the mean of log-transformed values as a function of time. In this way, this work provides a solid quantitative theoretical framework that grounds an experimental setup that has been used by several authors recently (Chang et al., 2008; Huang, 2009; Kalmar et al., 2009; Pina et al., 2012; Sisan et al., 2012).

Acknowledgements

We greatly acknowledge the support of Jocelyne Demengeot and Henrique Teotónio during the development of this work. TSG was supported by a fellowship from FCT (fellowship number SFRH/BD/33572/2008).

This chapter is an extended version of the theoretical framework presented in the manuscript:

Guzella, T. S., Barreto, V. B., and Carneiro J. (2013). Quantifying the Contributions and Dynamics Underlying Variation in Expression Levels in a Cell Population. *In preparation, under final review by the co-authors*

Chapter 2

Materials and Methods

Notation

The function $\log(\cdot)$ denotes the natural logarithm, and random variables are represented as bold symbols, as in $\boldsymbol{x}$. We use $\mathbb{E}[\boldsymbol{x}]$ to denote the expected value of a random variable $\boldsymbol{x}$, and $\mathbb{V}[\boldsymbol{x}]$ the variance. Moreover, let $\mathbb{K}[\boldsymbol{x}] = \sqrt{\mathbb{V}[\boldsymbol{x}]}/\mathbb{E}[\boldsymbol{x}]$ denote the coefficient of variation of $\boldsymbol{x}$. Being a normalized quantity, the coefficient of variation may be presented as a percentage, for convenience. The notation $\boldsymbol{z} \sim \mathcal{LN}(\mu, \sigma)$ represents a random variable $\boldsymbol{z}$ following a lognormal distribution with parameters μ and σ, having therefore probability density function (Papoulis and Pillai, 2002):

$$f(z) = \frac{1}{\sqrt{2\,\pi}\,\sigma\,z}\,\exp\left(-\frac{1}{2\,\sigma^2}\,(\log(z)-\mu)^2\right) \tag{2.31}$$

Numerical simulations

Simulations of the model describing protein expression levels in the full population, which relies on stochastic differential equations, were conducted using custom software written in C++, based on the GNU Scientific Library (http://www.gnu.org/software/gsl/). The Ornstein-Uhlenbeck process (equation 2.5) was simulated using exact expressions (Gillespie, 1996), while the equation for the dynamics of protein levels (equation 2.4) relied on a 4-th order Runge-Kutta method. In simulations where the value of σ_W was specified, the stochastic model was simulated (for given τ and β), and the Brent-Dekker method (GNU Scientific Library) was used to adjust the value of σ so as to obtain the desired value of σ_W.

Simulations of the isolation of cells were done using an initial population having 1.2×10^6 cells and 2×10^4 sub-populations, with the number of cells per sub-population following a multinomial distribution, and typically with $\sigma_T = 0.3$. From the starting population, 10% of cells were isolated. As a simple approximation of an experimental setting, each isolated population was divided into 3 replicates, and simulated for a given period of time, with snapshots of each replicate being collected in equally spaced time instants.

Analysis of $\Delta_{H,L}(t)$ obtained from simulations

To study the relationship between τ_T and parameters β and τ, simulations were ran for several combinations of values of $(R^2_\alpha, \beta, \tau)$. The values of τ_T were estimated by fitting the exponential model via non-linear least squares in MATLAB (Mathworks).

Appendices

Appendix 2.A Detailed derivation of the mean and variance of the full population

This section presents the detailed derivation of the mean and variance of the full population, given the parameters describing the sub-populations (equations 2.2 and 2.3 in the main text).

2.A.1 Mean and variance given the parameters of each sub-population

We start from the mixture model formulation, where x represents protein expression levels:

$$f(x \mid \theta_1, \theta_2, \ldots, \theta_N) = \sum_{i=1}^{N} w_i \, f_i(x \mid \zeta_i) \tag{2.32}$$

where $\theta_i = (w_i, \zeta_i)$ are parameters describing each of the N sub-populations. The frequency of the cells in the full population that belong to the i-th sub-population is represented by w_i (see equation 2.1 of the main text), while ζ_i parametrizes the probability density function $f_i(x \mid \zeta_i)$ of protein expression levels in that sub-population. For example, in the case that f_i is a normal distribution, ζ_i would be the mean and variance of expression levels in that sub-population.

It follows from equation 2.32 that the mean of the full population is given by:

$$\begin{aligned} \mu_F = \mathbb{E}[\boldsymbol{x}] &= \int_{-\infty}^{\infty} x \, f(x \mid \theta_1, \theta_2, \ldots, \theta_N) \, dx \\ &= \sum_{i=1}^{N} w_i \underbrace{\int_{-\infty}^{\infty} x \, f_i(x \mid \zeta_i) \, dx}_{\mu_i} = \sum_{i=1}^{N} w_i \, \mu_i \end{aligned} \tag{2.33}$$

where μ_i is the mean of the i-th sub-population. Therefore, the mean of the full population (μ_F) is simply the average of the means of the sub-populations, weighted by the frequencies w_i. The variance of expression levels, on the other hand, follows from:

$$v_F = \mathbb{V}[\boldsymbol{x}] = \mathbb{E}\left[\boldsymbol{x}^2\right] - \mu_F^2 \tag{2.34}$$

where:

$$\mathbb{E}\left[\boldsymbol{x}^2\right] = \sum_{i=1}^{N} w_i \underbrace{\int_{-\infty}^{\infty} x^2 \, f_i(x \mid \zeta_i) \, dx}_{v_i + \mu_i^2} \tag{2.35}$$

v_i being the variance of each sub-population. Hence, the variance is given by:

$$v_F = \mathbb{V}[\boldsymbol{x}] = \sum_{i=1}^{N} w_i \left(v_i + \mu_i^2\right) - \mu_F^2 = \sum_{i=1}^{N} w_i \, v_i + \sum_{i=1}^{N} w_i \, \mu_i^2 - \mu_F^2 \tag{2.36}$$

By Jensen's inequality, it follows that:

$$\sum_{i=1}^{N} w_i \, \mu_i^2 - \mu_F^2 \geq 0 \tag{2.37}$$

and, therefore, the variance v_F is always non-negative, as expected.

Therefore, for the "full" population, one has the mean and variance given by:

$$\mu_F = \mathbb{E}[\boldsymbol{x}] = \sum_{i=1}^{N} w_i \, \mu_i \tag{2.38}$$

$$v_F = \mathbb{V}[\boldsymbol{x}] = \sum_{i=1}^{N} w_i \, v_i + \sum_{i=1}^{N} w_i \, \mu_i^2 - \mu_F^2 \tag{2.39}$$

As a remark, these results are independent of the underlying probability density functions f_i describing the expression levels in each sub-population.

2.A.2 Mean and variance in the limit of large number of sub-populations

In the following, we study the asymptotic properties of the equations describing the mean and variance of expression levels in the full population (equations 2.38 and 2.39, respectively). In this case, the parameters of the sub-populations introduced in the previous section become themselves random variables, denoted as $\boldsymbol{w}$, for the frequency, $\boldsymbol{\mu}$ as the mean expression level, and $\boldsymbol{v}$ for the variance of a sub-population. To avoid confusing notation, in this section we will refer to the mean and variance of the random variables $\boldsymbol{w}$, $\boldsymbol{\mu}$ and $\boldsymbol{v}$ solely using the notation $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$. In the general case considered here, these random variables are described by a joint density $h(w, \mu, v)$. Hence, the full population to be studied is constructed from a sample $\mathcal{S}$ given by:

$$\mathcal{S} = \{(\boldsymbol{w_1}, \boldsymbol{\mu_1}, \boldsymbol{v_1}), \ldots, (\boldsymbol{w_i}, \boldsymbol{\mu_i}, \boldsymbol{v_i}) \ldots, (\boldsymbol{w_N}, \boldsymbol{\mu_N}, \boldsymbol{v_N})\} \tag{2.40}$$

consisting of N vector-valued random variables $(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{v})$ sampled from an unknown distribution. A simplifying assumption made hereafter is that $(\boldsymbol{w_i}, \boldsymbol{\mu_i}, \boldsymbol{v_i})$ and $(\boldsymbol{w_j}, \boldsymbol{\mu_j}, \boldsymbol{v_j})$

are independent and identically distributed (iid) for all $i \neq j$. In terms of the frequencies, this is not immediate, since w_i and w_j are dependent due to the constraint of unity sum ($\sum_{i=1}^{N} w_i = 1$). However, this dependency is expected to become negligible, as long as the numbers of cells in each sub-population (see equation 2.1 of the main text) are iid, and N is sufficiently large.

In the following, it is shown that, for a fixed N, the mean and variance of the full population are basically "sample estimates" based on $\mathcal{S}$. Since these estimates are functions of random variables, they are themselves random variables, denoted as $_{\boldsymbol{N}}\boldsymbol{\mu_F}$ and $_{\boldsymbol{N}}\boldsymbol{v_F}$, respectively (as in equations 2.38 and 2.39):

$$_{\boldsymbol{N}}\boldsymbol{\mu_F} = \sum_{i=1}^{N} \boldsymbol{w_i}\, \boldsymbol{\mu_i} \tag{2.41}$$

$$_{\boldsymbol{N}}\boldsymbol{v_F} = \sum_{i=1}^{N} \boldsymbol{w_i}\, \boldsymbol{v_i} + \sum_{i=1}^{N} \boldsymbol{w_i}\, {\boldsymbol{\mu_i}}^2 - {}_{\boldsymbol{N}}{\boldsymbol{\mu_F}}^2 \tag{2.42}$$

in which $\boldsymbol{w_i}$, $\boldsymbol{\mu_i}$ and $\boldsymbol{v_i}$ are random variables.

In this framework, one is interested in the expected value of the mean and variance of the full population. Under the law of large numbers, the sample estimates (equations 2.41 and 2.42) will converge to the expected values of the mean and variance for sufficiently large N. We start by deriving the asymptotic mean of the population:

$$\mu_F = \mathbb{E}\left[{}_{\boldsymbol{N}}\boldsymbol{\mu_F}\right] = N\, \mathbb{E}\left[\boldsymbol{w}\, \boldsymbol{\mu}\right] = N\, \left(\mathbb{E}\left[\boldsymbol{w}\right]\, \mathbb{E}\left[\boldsymbol{\mu}\right] + \mathbb{C}\left[\boldsymbol{w}, \boldsymbol{\mu}\right]\right) \tag{2.43}$$

where $\mathbb{C}\left[\boldsymbol{w}, \boldsymbol{\mu}\right]$ is the covariance between the random variables $\boldsymbol{w}$ and $\boldsymbol{\mu}$. Given that $\mathbb{E}\left[\boldsymbol{w}\right] = 1\,/N$, by definition of the frequencies, one obtains that:

$$\mu_F = \mathbb{E}\left[{}_{\boldsymbol{N}}\boldsymbol{\mu_F}\right] = \mathbb{E}\left[\boldsymbol{\mu}\right] + N\, \mathbb{C}\left[\boldsymbol{w}, \boldsymbol{\mu}\right] \tag{2.44}$$

Therefore, it follows that, when the $\boldsymbol{w}$ and $\boldsymbol{\mu}$ are uncorrelated, the expected mean of the population is simply the expected mean of the sub-populations, corresponding to equation 2.2 of the main text.

Following a similar reasoning, one obtains the variance:

$$v_F = \mathbb{E}\left[{}_{\boldsymbol{N}}\boldsymbol{v_F}\right] = N\, \left(\mathbb{E}\left[\boldsymbol{w}\, \boldsymbol{v}\right] + \mathbb{E}\left[\boldsymbol{w}\, \boldsymbol{\mu^2}\right]\right) - \mathbb{E}\left[{}_{\boldsymbol{N}}{\boldsymbol{\mu_F}}^2\right] \tag{2.45}$$

where:

$$\mathbb{E}\left[\boldsymbol{w}\,\boldsymbol{v}\right] = \frac{1}{N}\,\mathbb{E}\left[\boldsymbol{v}\right] + \mathbb{C}\left[\boldsymbol{w},\boldsymbol{v}\right] \tag{2.46}$$

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{w}\,\boldsymbol{\mu^2}\right] &= \frac{1}{N}\underbrace{\mathbb{E}\left[\boldsymbol{\mu^2}\right]}_{\mathbb{V}[\boldsymbol{\mu}]+(\mathbb{E}[\boldsymbol{\mu}])^2} + \mathbb{C}\left[\boldsymbol{w},\boldsymbol{\mu^2}\right] \\ &= \frac{1}{N}\left(\mathbb{V}\left[\boldsymbol{\mu}\right] + \left(\mathbb{E}\left[\boldsymbol{\mu}\right]\right)^2\right) + \mathbb{C}\left[\boldsymbol{w},\boldsymbol{\mu^2}\right]\end{aligned} \tag{2.47}$$

Note the appearance of the term $\mathbb{C}\left[\boldsymbol{w},\boldsymbol{\mu^2}\right]$, containing the additional random variable $\boldsymbol{\mu^2}$. Furthermore, the last term in equation 2.45 can be written as:

$$\mathbb{E}\left[{}_{\boldsymbol{N}}\boldsymbol{\mu_F}^2\right] = \mathbb{V}\left[{}_{\boldsymbol{N}}\boldsymbol{\mu_F}\right] + \left(\mathbb{E}\left[{}_{\boldsymbol{N}}\boldsymbol{\mu_F}\right]\right)^2 \tag{2.48}$$

which, using equation 2.44, becomes:

$$\begin{aligned}\mathbb{E}\left[{}_{\boldsymbol{N}}\boldsymbol{\mu_F}^2\right] &= \mathbb{V}\left[{}_{\boldsymbol{N}}\boldsymbol{\mu_F}\right] + \left(\mathbb{E}\left[\boldsymbol{\mu}\right] + N\,\mathbb{C}\left[\boldsymbol{w},\boldsymbol{\mu}\right]\right)^2 \\ &= \mathbb{V}\left[{}_{\boldsymbol{N}}\boldsymbol{\mu_F}\right] + \left(\mathbb{E}\left[\boldsymbol{\mu}\right]\right)^2 + 2\,N\,\mathbb{E}\left[\boldsymbol{\mu}\right]\,\mathbb{C}\left[\boldsymbol{w},\boldsymbol{\mu}\right] + N^2\,\left(\mathbb{C}\left[\boldsymbol{w},\boldsymbol{\mu}\right]\right)^2\end{aligned} \tag{2.49}$$

Plugging back equations 2.46, 2.47 and 2.49 into 2.45, the variance is given by:

$$\begin{aligned}v_F = \mathbb{E}\left[{}_{\boldsymbol{N}}\boldsymbol{v_F}\right] &= \mathbb{E}\left[\boldsymbol{v}\right] + N\,\mathbb{C}\left[\boldsymbol{w},\boldsymbol{v}\right] \\ &+ \mathbb{V}\left[\boldsymbol{\mu}\right] + \left(\mathbb{E}\left[\boldsymbol{\mu}\right]\right)^2 + \mathbb{C}\left[\boldsymbol{w},\boldsymbol{\mu^2}\right] - \mathbb{V}\left[{}_{\boldsymbol{N}}\boldsymbol{\mu_F}\right] \\ &- \left\{\left(\mathbb{E}\left[\boldsymbol{\mu}\right]\right)^2 + 2\,N\,\mathbb{E}\left[\boldsymbol{\mu}\right]\,\mathbb{C}\left[\boldsymbol{w},\boldsymbol{\mu}\right] + N^2\,\left(\mathbb{C}\left[\boldsymbol{w},\boldsymbol{\mu}\right]\right)^2\right\}\end{aligned} \tag{2.50}$$

which is reduced to:

$$\begin{aligned}v_F = \mathbb{E}\left[{}_{\boldsymbol{N}}\boldsymbol{v_F}\right] &= \mathbb{E}\left[\boldsymbol{v}\right] + \mathbb{V}\left[\boldsymbol{\mu}\right] \\ &- \mathbb{V}\left[{}_{\boldsymbol{N}}\boldsymbol{\mu_F}\right] + N\left\{\mathbb{C}\left[\boldsymbol{w},\boldsymbol{v}\right] + \mathbb{C}\left[\boldsymbol{w},\boldsymbol{\mu^2}\right]\right. \\ &\left. - 2\,\mathbb{E}\left[\boldsymbol{\mu}\right]\,\mathbb{C}\left[\boldsymbol{w},\boldsymbol{\mu}\right] - N\,\left(\mathbb{C}\left[\boldsymbol{w},\boldsymbol{\mu}\right]\right)^2\right\}\end{aligned} \tag{2.51}$$

The term $\mathbb{V}\left[{}_{\boldsymbol{N}}\boldsymbol{\mu_F}\right]$ represents an additional contribution, due to variance in the sample mean of the full population as a consequence of sampling, and tends to zero as N grows. In this case, provided that there is no correlation between the frequencies ($\boldsymbol{w}$) and either the means ($\boldsymbol{\mu}$), the squared means ($\boldsymbol{\mu^2}$) and the variances ($\boldsymbol{v}$) of the sub-population, one obtains equations 2.2 and 2.3 of the main text.

Appendix 2.B Basic properties of the logarithmic transformation

In this session, we recall some basic properties of the logarithmic transformation (see, for example, Mood et al., 1974). First of all, recall that, a lognormally-distributed random variable $\boldsymbol{x} \sim \mathcal{LN}(\mu, \sigma)$ has expected value, variance and coefficient of variation given, respectively, by:

$$\mathbb{E}[\boldsymbol{x}] = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \tag{2.52}$$

$$\mathbb{V}[\boldsymbol{x}] = \left(\exp\left(\sigma^2\right) - 1\right) \exp\left(2\,\mu + \sigma^2\right) \tag{2.53}$$

$$\mathbb{K}[\boldsymbol{x}] = \sqrt{\exp\left(\sigma^2\right) - 1} \tag{2.54}$$

Conversely, the parameters μ and σ of the lognormal distribution are can be obtained from $\mathbb{E}[\boldsymbol{x}]$ and $\mathbb{V}[\boldsymbol{x}]$ via:

$$\mu = \log\left(\frac{\left(\mathbb{E}[\boldsymbol{x}]\right)^2}{\sqrt{\mathbb{V}[\boldsymbol{x}] + \left(\mathbb{E}[\boldsymbol{x}]\right)^2}}\right) \tag{2.55}$$

$$\sigma = \sqrt{\log\left(1 + \frac{\mathbb{V}[\boldsymbol{x}]}{\left(\mathbb{E}[\boldsymbol{x}]\right)^2}\right)} \tag{2.56}$$

In order to frame the relationship between untransformed and log-transformed values, consider a random variable $\boldsymbol{z}$, and define $\boldsymbol{y} = \log(\boldsymbol{z})$. If $\boldsymbol{y}$ can be well approximated by a normal distribution, then equations 2.52 and 2.53 can be used to relate the mean and variance of $\boldsymbol{z}$ and $\boldsymbol{y}$:

$$\mu_y \approx \log(\mu_z) - \frac{1}{2}\,\sigma_y^2 \tag{2.57}$$

$$\sigma_y^2 \approx \log\left(k_z^2 + 1\right) \tag{2.58}$$

where $k_z = \mathbb{K}[\boldsymbol{z}]$ is the coefficient of variation of $\boldsymbol{z}$.

Appendix 2.C Non-dimensional version of the stochastic model

Let t' denote the rescaled time, $x'_{t'}$ the rescaled protein levels and $y_{t'}$ the rescaled Ornstein-Uhlenbeck process. Note that the later is modified only by the scaling in time. Consider the following approach for non-dimensionalization:

$$t' = t/\beta \tag{2.59}$$

$$x'_{t'} = x_t/(\alpha\,\beta) \tag{2.60}$$

where time is scaled by the parameter representing the mean protein lifetime (β), and protein levels are scaled by the stationary mean of the population. Recalling the property of time scaling for the Wiener process:

$$dW_{t'} = dW_t/\sqrt{\beta} \tag{2.61}$$

one obtains the following non-dimensional model:

$$dx'_{t'} = \left\{\exp\left(y_{t'} - \frac{1}{2}\sigma^2\right) - x'_{t'}\right\}\,dt' \tag{2.62}$$

$$dy_{t'} = -\frac{1}{\tau/\beta}\,y_{t'}\,dt' + \frac{\sigma}{\sqrt{(\tau/\beta)/2}}dW_{t'} \tag{2.63}$$

where variable $x'_{t'}$ has unity stationary mean. In particular, note that equation 2.63 corresponds to the Ornstein-Uhlenbeck process with parameters σ and τ/β. Therefore, as expected, the rescaling of time does not affect the stationary variance of this process, and merely rescales the auto-correlation function. The non-dimensional version proves a key property of the stochastic model defined by equations 2.4 and 2.5: parameters β and τ contribute to the stationary value of every property only through the ratio τ/β. In other words, in the most general case, the stationary value of every property of the model depends only on the parameters α, σ and τ/β. Furthermore, it easily follows that the stationary coefficient of variation depends only on the parameters σ and τ/β. Finally, it can be shown that parameters τ and β will also affect the mean-scaled model (with $x'_{t'} = x_t/(\alpha\,\beta)$) only via the ratio τ/β if time is scaled by τ or by the sum $\tau + \beta$, instead of by β as in equation 2.63.

Appendix 2.D Model of protein expression in a cell population, for untransformed values

2.D.1 Variation within a sub-population

Starting from equation 2.4, it follows that a population of cells with dynamics of protein expression levels governed by equations 2.4 and 2.5 has stationary mean given by:

$$\mu = \mathbb{E}\left[\boldsymbol{x_t}\right] = \alpha\,\beta \tag{2.64}$$

and therefore the stationary mean depends on the average expression rate (therefore, α) and on mean protein lifetime (β). Moreover, the squared stationary coefficient of variation is given by:

$$k_W^2 = \mathbb{K}\left[\boldsymbol{x_t}\right]^2 = g_k(\exp\left(\sigma^2\right) - 1, \tau/\beta) \tag{2.65}$$

where $g_k(\cdot,\cdot)$ is an arbitrary function, which can be estimated via simulation (analogous to $g(\cdot,\cdot)$ in equation 2.12), and the subscript W highlights that the variation is due to the stochastic process influencing the instantaneous rate of protein expression. Hence, the stationary variance is given by:

$$\mathbb{V}\left[\boldsymbol{x_t}\right] = \left(\alpha\,\beta\right)^2\,k_W^2 \tag{2.66}$$

2.D.2 Variation among sub-populations

Following equation 2.66, the i-th sub-population, with parameter α_i, has mean and variance of protein levels (see equations 2.64 and 2.66):

$$\mu_i = \alpha_i\,\beta \tag{2.67}$$

$$v_i = \alpha_i^2\,\left(\beta\,k_W\right)^2 \tag{2.68}$$

where it should be noted that k_W^2 is the same for all sub-populations. Applying equations 2.2 and 2.3 of the main text, one obtains that the squared coefficient of variation of the full population is given by:

$$k_F^2 = k_W^2 + k_\alpha^2 + \left(k_W\,k_\alpha\right)^2 \tag{2.69}$$

Therefore, equation 2.69, based on untransformed values, does not follow the simple additive relationship obtained for the variances of log-transformed values (equation 2.16 of the main text), given the extra term $\left(k_W\,k_\alpha\right)^2$.

Appendix 2.E Dynamics of the mean of log-transformed values

This section studies the dynamics of the log-transformed mean, to provide a rationale for the approximately exponential decay of the function $\Delta_{H,L}(t)$, shown based on simulations in figure 2.4 (section 2.5.2) of the main text. The first step is the derivation of a linearized approximation of the log-transformed stochastic model that describes protein expression in a sub-population (defined by equations 2.4 and 2.5 of the main text). Afterwards, the dynamics of function $\Delta_{H,L}(t)$, which depend on the mean of log-transformed values of high and low expressors, are related to the dynamics of expression levels in the underlying sub-populations.

Since analysis is based on log-transformed values, we define the log-transformed protein level s_t:

$$s_t = \log(x_t) \tag{2.70}$$

such that:

$$x_t = \exp(s_t) \tag{2.71}$$

$$\frac{d\,s_t}{dt} = \frac{1}{x_t}\frac{d\,x_t}{dt} \tag{2.72}$$

where x_t is the protein level and y_t is the Ornstein-Uhlenbeck process in the original stochastic model (section 2.3.1 of the main text) Therefore, it follows from equation 2.4 of the main text that:

$$\begin{aligned}\beta\, ds_t &= \left\{\alpha\,\beta\exp\left(y_t - s_t - \frac{1}{2}\sigma^2\right) - 1\right\} dt \\ &= \left\{\exp\left(\log(\alpha\,\beta) + y_t - s_t - \frac{1}{2}\sigma^2\right) - 1\right\} dt\end{aligned} \tag{2.73}$$

The dynamics of the mean of log-transformed values in a sub-population are then given by:

$$\beta\, d\,\mathbb{E}[s_t] = \left\{\mathbb{E}\left[\exp\left(\log(\alpha\,\beta) + y_t - s_t - \frac{1}{2}\sigma^2\right)\right] - 1\right\} dt \tag{2.74}$$

In order to derive an approximation for the function $\mathbb{E}[s_t]$, it is necessary to simplify the following term:

$$\mathbb{E}\left[\exp\left(\log(\alpha\,\beta) + y_t - s_t - \frac{1}{2}\sigma^2\right)\right] \tag{2.75}$$

Introducing:

$$S_t = \log(\alpha\,\beta) + y_t - s_t - \frac{1}{2}\sigma^2 \tag{2.76}$$

and assuming that it is well-concentrated around a certain instantaneous mean, such that it can be approximated by a normal distribution with mean m_t and variance v_t, it follows that (see equation 2.52):

$$\mathbb{E}[\exp(S_t)] \approx \exp\left(m_t + \frac{1}{2}\,v_t\right) \tag{2.77}$$

$$m_t = \mathbb{E}[S_t] = \log(\alpha\,\beta) + \mathbb{E}[y_t] - \mathbb{E}[s_t] - \frac{1}{2}\sigma^2 \tag{2.78}$$

$$v_t = \mathbb{V}[S_t] = \mathbb{V}[s_t] + \mathbb{V}[y_t] - 2\,\mathbb{C}[s_t, y_t] \tag{2.79}$$

Assuming that $\left|m_t + \frac{1}{2}\,v_t\right|$ is relatively small, the exponential term in the left-hand side of equation 2.77 can be linearized:

$$\mathbb{E}[\exp(S_t)] \approx \mathbb{E}[1 + S_t] = 1 + m_t + \frac{1}{2}\,v_t \tag{2.80}$$

Plugging back into equation 2.74, one obtains the following linear approximation for the dynamics of the mean of log-transformed values:

$$\beta\,\frac{d\,\mathbb{E}[s_t]}{dt} = \log(\alpha\,\beta) + \mathbb{E}[y_t] - \mathbb{E}[s_t] + \frac{1}{2}\left(v_t - \sigma^2\right) \tag{2.81}$$

The equation for the mean of the Ornstein-Uhlenbeck process follows as:

$$\tau\,\frac{d\,\mathbb{E}[y_t]}{dt} = -\mathbb{E}[y_t] \tag{2.82}$$

with solution:

$$\mathbb{E}[y_t] = \mathbb{E}[y_0]\,\exp(-t/\tau) = \mu_{y,0}\,\exp(-t/\tau) \tag{2.83}$$

Therefore, one obtains the following equation for $\mu_t = \mathbb{E}[s_t]$, which denotes the instantaneous mean of log-transformed values of a single sub-population:

$$\beta\,\frac{d\,\mu_t}{dt} = \log(\alpha\,\beta) + \mu_{y,0}\,\exp(-t/\tau) - \mu_t + \frac{1}{2}\left(v_t - \sigma^2\right) \tag{2.84}$$

where recall that v_t (equation 2.79) depends on the variances of log-transformed values of the sub-population (s_t) and the Ornstein-Uhlenbeck process variable (y_t), besides the covariance between these two.

In terms of the function $\Delta_{H,L}(t)$, recall that it is defined as (equation 2.19 of the main text):

$$\Delta_{H,L}(t) = \mu_{H,t} - \mu_{L,t} \tag{2.85}$$

where $\mu_{H,t}$ and $\mu_{L,t}$ are the means of log-transformed values of high and low expressors, respectively, at time t. Using equation 2.44, $\Delta_{H,L}(t)$ can be written as:

$$\begin{aligned}\Delta_{H,L}(t) = \mathbb{E}\left[\boldsymbol{\mu_{H,t}}\right] - \mathbb{E}\left[\boldsymbol{\mu_{L,t}}\right] \\ + \left(N_h\, \mathbb{C}\left[\boldsymbol{w}, \boldsymbol{\mu_{H,t}}\right] - N_l\, \mathbb{C}\left[\boldsymbol{w}, \boldsymbol{\mu_{L,t}}\right]\right)\end{aligned} \tag{2.86}$$

where $\boldsymbol{\mu_{H,t}}$ and $\boldsymbol{\mu_{L,t}}$ are random variables denoting the instantaneous mean of a particular sub-population in the high (N_h sub-populations) and low expressors (N_l sub-populations), respectively. Neglecting the term of weighted difference between the covariances in equation 2.86, it follows that:

$$\Delta_{H,L}(t) \approx \mathbb{E}\left[\boldsymbol{\mu_{H,t}}\right] - \mathbb{E}\left[\boldsymbol{\mu_{L,t}}\right] \tag{2.87}$$

and therefore the dynamics of $\Delta_{H,L}(t)$ can be approximated as:

$$\begin{aligned}\frac{d}{dt}\Delta_{H,L}(t) &\approx \frac{d}{dt}\mathbb{E}\left[\boldsymbol{\mu_{H,t}}\right] - \frac{d}{dt}\mathbb{E}\left[\boldsymbol{\mu_{L,t}}\right] \\ &= \mathbb{E}\left[\frac{d}{dt}\boldsymbol{\mu_{H,t}}\right] - \mathbb{E}\left[\frac{d}{dt}\boldsymbol{\mu_{L,t}}\right]\end{aligned} \tag{2.88}$$

An approximation to the term $\frac{d}{dt}\boldsymbol{\mu_{H,t}}$ has been derived in equation 2.84, such that:

$$\begin{aligned}\mathbb{E}\left[\frac{d}{dt}\boldsymbol{\mu_{H,t}}\right] \approx \frac{1}{\beta}\Bigg\{&\mathbb{E}\left[\log\left(\boldsymbol{\alpha_H}\,\beta\right)\right] + \mathbb{E}\left[\boldsymbol{\mu_{y_H,0}}\right]\,\exp\left(-t/\tau\right) \\ &- \mathbb{E}\left[\boldsymbol{\mu_{H,t}}\right] + \frac{1}{2}\left(\mathbb{E}\left[\boldsymbol{v_{H,t}}\right] - \sigma^2\right)\Bigg\}\end{aligned} \tag{2.89}$$

and analogously for the low expressors. Plugging into equation 2.88, one obtains:

$$\frac{d}{dt}\,\Delta_{H,L}(t) \approx \frac{1}{\beta}\Bigg\{\mathbb{E}\left[\log\left(\boldsymbol{\alpha_H}\right)\right] - \mathbb{E}\left[\log\left(\boldsymbol{\alpha_L}\right)\right] + \left(\mathbb{E}\left[\boldsymbol{\mu_{y_H,0}}\right] - \mathbb{E}\left[\boldsymbol{\mu_{y_L,0}}\right]\right)\,\exp\left(-t/\tau\right) - \left(\mathbb{E}\left[\boldsymbol{\mu_{H,t}}\right] - \mathbb{E}\left[\boldsymbol{\mu_{L,t}}\right]\right) + \frac{1}{2}\left(\mathbb{E}\left[\boldsymbol{v_{H,t}}\right] - \mathbb{E}\left[\boldsymbol{v_{L,t}}\right]\right)\Bigg\} \tag{2.90}$$

By symmetry, the difference $\mathbb{E}\left[\boldsymbol{v_{H,t}}\right] - \mathbb{E}\left[\boldsymbol{v_{L,t}}\right]$ in equation 2.90 is expected to be close to zero. Defining the constants:

$$\Gamma = \mathbb{E}\left[\log\left(\boldsymbol{\alpha_H}\right)\right] - \mathbb{E}\left[\log\left(\boldsymbol{\alpha_L}\right)\right] \tag{2.91}$$

$$\Lambda = \mathbb{E}\left[\boldsymbol{\mu_{y_H,0}}\right] - \mathbb{E}\left[\boldsymbol{\mu_{y_L,0}}\right] \tag{2.92}$$

equation 2.90 is simplified to take the form:

$$\beta\,\frac{d}{dt}\,\Delta_{H,L}(t) \approx \Gamma + \Lambda\,\exp\left(-t/\tau\right) - \left(\mathbb{E}\left[\boldsymbol{\mu_{H,t}}\right] - \mathbb{E}\left[\boldsymbol{\mu_{L,t}}\right]\right) \tag{2.93}$$

Finally, using the original approximation in equation 2.87, it follows that:

$$\beta\,\frac{d}{dt}\,\Delta_{H,L}(t) \approx \Gamma + \Lambda\,\exp\left(-t/\tau\right) - \Delta_{H,L}(t) \tag{2.94}$$

Introducing the auxiliary variable $U(t) = \Lambda\,\exp\left(-t/\tau\right)$, then equation 2.94 can be written as a two-dimensional linear dynamical system:

$$\frac{d}{dt}\begin{bmatrix}\Delta_{H,L}(t)\\ U(t)\end{bmatrix} = \underbrace{\begin{bmatrix}-\frac{1}{\beta} & \frac{1}{\beta}\\ 0 & -\frac{1}{\tau}\end{bmatrix}}_{A}\begin{bmatrix}\Delta_{H,L}(t)\\ U(t)\end{bmatrix} + \begin{bmatrix}\frac{\Gamma}{\beta}\\ 0\end{bmatrix} \tag{2.95}$$

subject to the initial condition:

$$\begin{bmatrix}\Delta_{H,L}(0)\\ U(0)\end{bmatrix} = \begin{bmatrix}\delta_0\\ \Lambda\end{bmatrix} \tag{2.96}$$

Since the matrix A (equation 2.95) has always non-imaginary eigenvalues, it follows that $\Delta_{H,L}(t)$ is the combination of two exponential decays with mean times given by τ and β, with a single dominant exponential for the extreme cases $\beta \gg \tau$ or $\beta \ll \tau$.

Appendix 2.F Detailed simulation study to compare the values of $\lim_{t\to\infty} \Delta_{H,L}(t)$ and R^2_α

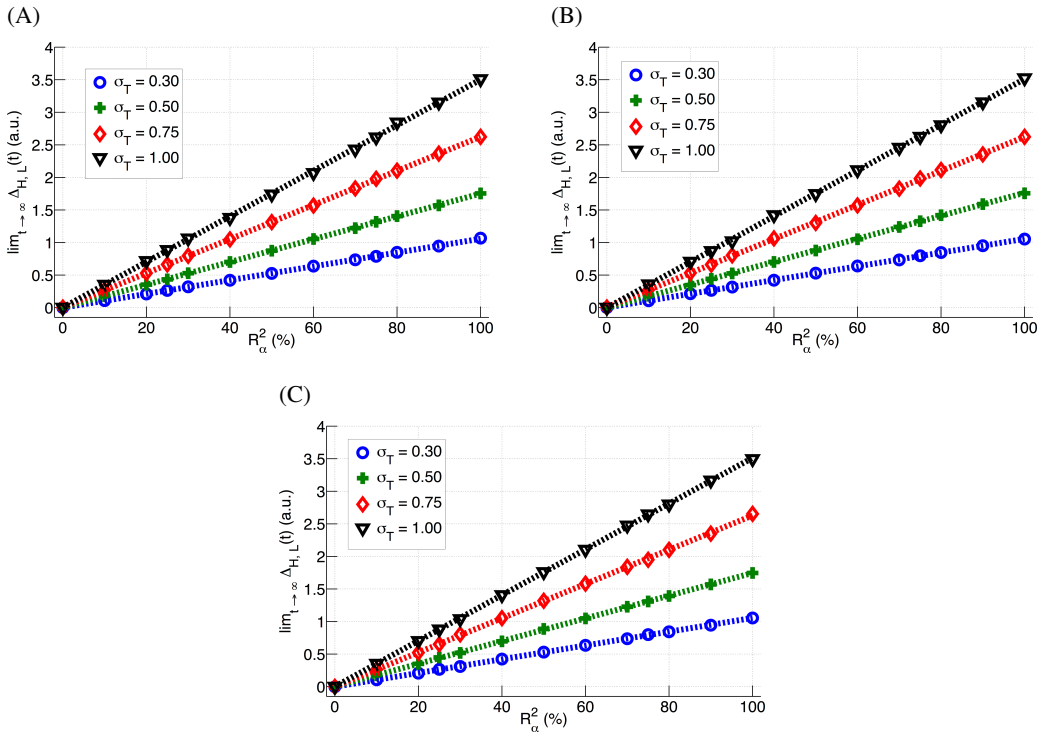

Figure 2.7: Validation of the linear relationship between $\lim_{t\to\infty} \Delta_{H,L}(t)$ and R^2_α for a wide range of parameter values of τ/β, representing whether the changes in the rate of protein production or protein lifetime is the slowest process, governed by parameters τ and β, respectively, and for different values of the total variation σ_T. Each figure shows the results for a given value of τ/β, with $\tau/beta = 0.1$, in which case the mean lifetime of the protein being much greater than the auto-correlation time of the rate of protein production (A), $\tau/\beta = 1$, with both processes having comparable timescales (B), and $\tau/\beta = 10$, with the protein being relatively short-lived (C).

Appendix 2.G Analysis of the variances of isolated populations

In this section, it is shown that analyzing the variances of isolated populations can provide additional information, given the estimate of R^2_α. This analysis is based on the same simulation setting considered in section 2.5. However, it was found that the estimates derived herein are much more sensitive to sampling effects. Therefore, the starting populations considered here had a much larger number of cells (see methods in section 2.G). Further scaling up of the number of cells in the simulations confirmed the overall conclusion obtained in this section, namely that the estimates obtained using the variances are biased. The simulations presented focus on the case of $\sigma_T = 0.3$, with equivalent results being obtained for the values tested up to $\sigma_T = 1$.

Figure 2.8 shows the variance of log-transformed values as a function of time for the "high expressors". As in section 2.5, the "all expressors" are also included, as a reference of the starting population. The variance of high expressors is lower than that of the starting population, and either remains constant or increases as a function of time. The same takes place for the low expressors, since the variance is a moment of even order. Finally, as observed for the mean of log values, the asymptotic (stationary) variance is equal to that of the "all expressors" for $R^2_\alpha = 0$, since in this case the unstable component is the only contribution present.

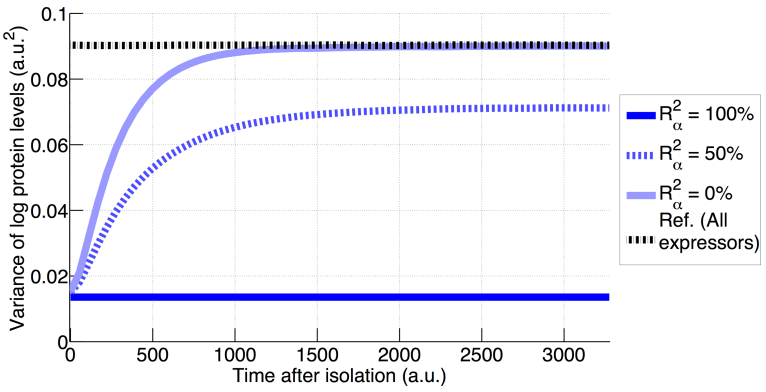

Figure 2.8: Variance (log values) of "high expressors" after isolation as 10% of starting populations with different values of R^2_α, but constant σ^2_T. The results shown correspond to $\tau = 500$, $\beta = 50$ and $\sigma_T = 0.3$.

Focusing on the asymptotic (stationary) variance, figure 2.9 shows that the simple, linear, relationship between the mean and R^2_α does not hold in this case. In particular, for $R^2_\alpha \leq 30\%$, the variance of high and low expressors is very close to that of the all

expressors. In order to understand the basis for the relationship showed in figure 2.9, we consider hereafter the partitioning of the variance of each isolated population, in contrast to the main text, which focused on the starting population. However, the value of R^2_α considered will always refer to that of the starting population.

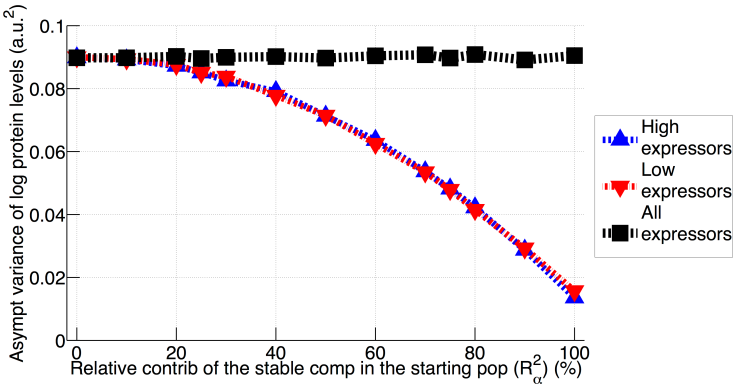

Figure 2.9: Asymptotic (stationary) variance of expression levels (log values) of high and low expressors, isolated in the simulations as 10% of the starting population, and also of all expressors. The symbols represent the values obtained from the simulations, while the lines represent linear interpolation. Results are shown for $\tau = 500$, $\beta = 50$ and $\sigma_T = 0.3$.

Recall that the variance of the starting population is given by:

$$\sigma^2_T = \sigma^2_W + \sigma^2_\alpha \tag{2.97}$$

For a general isolated population D, based on equation 2.51 (appendix 2.A), the variance is partitioned as:

$$\sigma^2_{T,D}(t) = \sigma^2_{W,D}(t) + \sigma^2_{\alpha,D} + \delta_{T,D} \tag{2.98}$$

The subscripts H, L and A in place of D will be used to refer to the isolated populations corresponding to high, low and all expressors, respectively. The notation $\sigma^2_{W,D}(t)$ highlights that the variance due to the unstable component becomes a function of time, which will increase until the population becomes stationary. The variance due to the stable component in the isolated population is represented by $\sigma^2_{\alpha,D}$, to highlight the fact that it may be different from that of the starting population (σ^2_α) as a consequence of isolating only some sub-populations (see discussion in section 2.4.2 of the main text). Finally, the term $\delta_{T,D}$ in equation 2.98 represents a "residual contribution", which may be introduced by the process of isolating cells. It arises from the covariance terms, which may become non-negligible

even for a starting population satisfying equation 2.97. Note that, by definition, $\delta_{T,A} = 0$ (since the "all expressors" satisfy equation 2.97).

In analogy with the standard F-statistic used for comparing the variances of two samples, we will denote by F_D the ratio between the asymptotic variance of the isolated population (equation 2.97) and the variance of the starting population (2.98):

$$F_D = \lim_{t \to \infty} \frac{\sigma^2_{T,D}(t)}{\sigma^2_T} \tag{2.99}$$

highlighting that this ratio depends only on measurable properties of the two populations. Moreover, define Φ_D as the ratio between the absolute variances of the stable component in the isolated and starting populations:

$$\Phi_D = \frac{\sigma^2_{\alpha,D}}{\sigma^2_\alpha},\ R^2_\alpha \neq 0 \tag{2.100}$$

to denote the relative change, as a consequence of isolating cells, in the variance of the stable component in the "new" (isolated) population. In the following, it is shown that F_D and R^2_α can be used to construct an estimator for Φ_D, denoted as $\widehat{\Phi}_D$. The requirement for $R^2_\alpha \neq 0$ stems from the constraint of $\sigma^2_\alpha \neq 0$.

It follows from taking the ratio between equations 2.98 and 2.97 that:

$$\begin{aligned} F_D &= \frac{\sigma^2_W}{\sigma^2_W + \sigma^2_\alpha} + \frac{\sigma^2_{\alpha,D} + \delta_{T,D}}{\sigma^2_W + \sigma^2_\alpha} \\ &= \frac{\sigma^2_W}{\sigma^2_W + \sigma^2_\alpha} + \frac{\left(\sigma^2_{\alpha,D} + \delta_{T,D}\right)/\sigma^2_\alpha}{1 + \left(\sigma^2_W/\sigma^2_\alpha\right)} \\ &= \frac{\sigma^2_W}{\sigma^2_W + \sigma^2_\alpha} + \frac{1}{1 + \left(\sigma^2_W/\sigma^2_\alpha\right)} \left(\Phi_D + \epsilon_{V,D}\right),\ R^2_\alpha \neq 0 \end{aligned} \tag{2.101}$$

where $\epsilon_{V,D}$ is defined as:

$$\epsilon_{V,D} = \frac{\delta_{T,D}}{\sigma^2_\alpha} \tag{2.102}$$

Using the definition of R^2_α:

$$\Phi_D = \underbrace{1 - \frac{1}{R^2_\alpha}\left(1 - F_D\right)}_{\widehat{\Phi}_D} - \epsilon_{V,D},\ R^2_\alpha \neq 0 \tag{2.103}$$

it follows that one estimator for Φ_D, denoted as $\widehat{\Phi}_D$, can be obtained via:

$$\widehat{\Phi}_D = 1 - \frac{1}{R_\alpha^2}\,(1 - F_D) \tag{2.104}$$

Hence, the "true" and estimated values are related via:

$$\widehat{\Phi}_D = \Phi_D + \epsilon_{V,D},\ R_\alpha^2 \neq 0 \tag{2.105}$$

in which $\epsilon_{V,D}$ becomes the bias in the estimation of Φ_D.

Hereafter, we conduct a more detailed analysis of the contributions to the variance in the isolated populations, to evaluate the use of the estimator $\widehat{\Phi}_D$ in quantifying Φ_D. This analysis is based on isolating the populations of interest, and simulating until they become stationary. At this point, using the underlying structure of each isolated population, the expression levels and number of cells in each sub-population, were determined. Using equation 2.51, the different terms were then calculated.

To understand the basis of the residual contribution ($\delta_{T,D}$), figures 2.10A–2.10C depict the contributions to the asymptotic variance ($\sigma_{T,D}^2$, $\sigma_{W,D}^2$, $\sigma_{\alpha,D}^2$ and $\sigma_{W,D}^2 + \sigma_{\alpha,D}^2$) of each of the isolated populations, as a function of the value of R_α^2. For each isolated population, the total variance ($\sigma_{T,D}^2$) corresponds exactly to the data shown in figure 2.9, while the term due to the unstable component ($\sigma_{W,D}^2$), being equal to that in the starting population (σ_W^2), is simply $\sigma_T^2\,\left(1 - R_\alpha^2\right)$. The converse holds for the term arising due to the stable component in the population of all expressors ($\sigma_{\alpha,A}^2 = \sigma_T^2\,R_\alpha^2$), since this population is equivalent to the starting one. On the other hand, for high and low expressors, the value of $\sigma_{\alpha,D}^2$, follows a more complicated dependence on R_α^2, reaching a maximum for values of R_α^2 around 60%. Moreover, in these two isolated populations, the sum $\sigma_{W,D}^2 + \sigma_{\alpha,D}^2$ is greater than the total variance $\sigma_{T,D}^2$, especially for intermediate values of R_α^2. This difference corresponds exactly to the residual component $\delta_{T,D}$. Moreover, as shown in figure 2.10D, it constitutes up to 15% of the total variance, having negative values for all R_α^2. The occurrence of negative values is not unexpected, given that there are both positive and negative terms in equation 2.51.

However, it is important to recall that the bias $\epsilon_{V,D}$ in the estimation of Φ_D corresponds to the residual component divided by σ_α^2. The bias is shown in figure 2.11 as a function of R_α^2, keeping in mind that it is only defined for $R_\alpha^2 \neq 0$. While the residual variance has an absolute value of up to 5% (figure 2.10D), it follows that the bias $\epsilon_{V,D}$ varies from -0.18 to 0, vanishing only for $R_\alpha^2 \to 100\%$. Hence, it is expected that $\widehat{\Phi}_D$ under-estimates Φ_D.

Finally, the "true" and the estimated values of Φ_D are shown in figure 2.12. In this

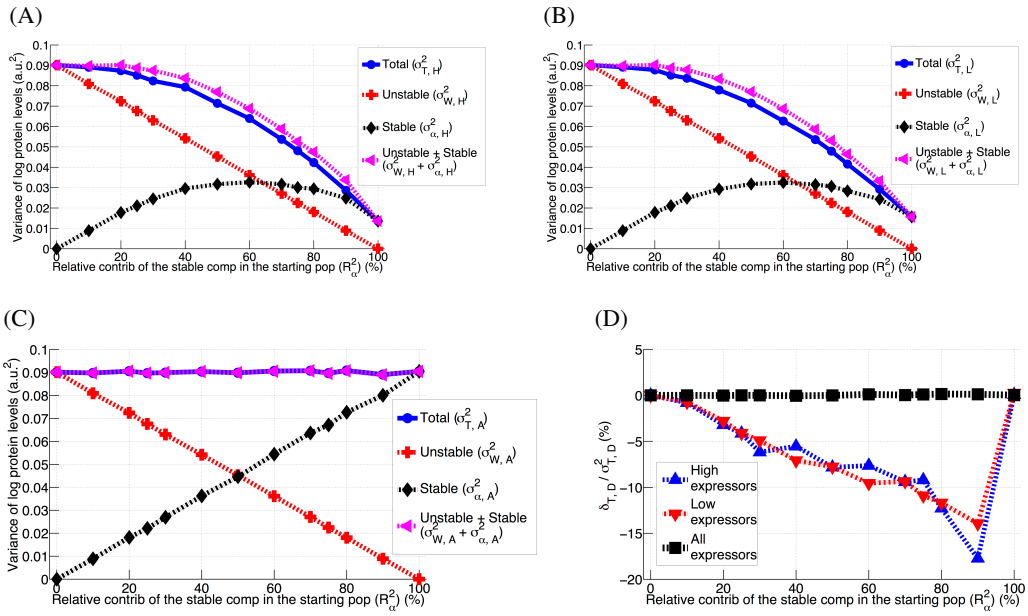

Figure 2.10: Properties of the isolated populations for various values of R^2_α, always considering log-transformed expression levels. Shown here are the variance components in the various isolated populations, along with the residual component ($\delta_{T,D}$). The latter has been calculated based on equation 2.51, and the values shown are normalized by the total variance. The results shown correspond to $\tau = 500$, $\beta = 50$ and $\sigma_T = 0.3$.

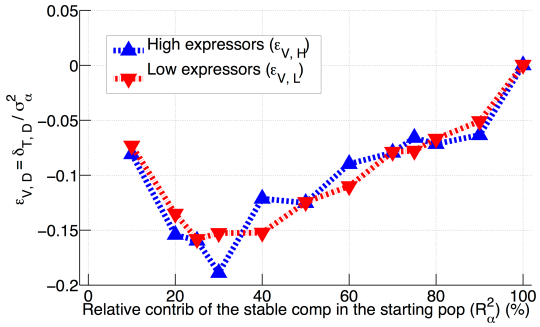

Figure 2.11: Bias term $\epsilon_{V,D}$ as a function of R^2_α (for $R^2_\alpha \neq 0$), considering the analysis based on the pairs (high, all) and (low, all), determined based on the residual variance (equation 2.51) and σ^2_α. Results are shown for $\tau = 500$, $\beta = 50$ and $\sigma_T = 0.3$.

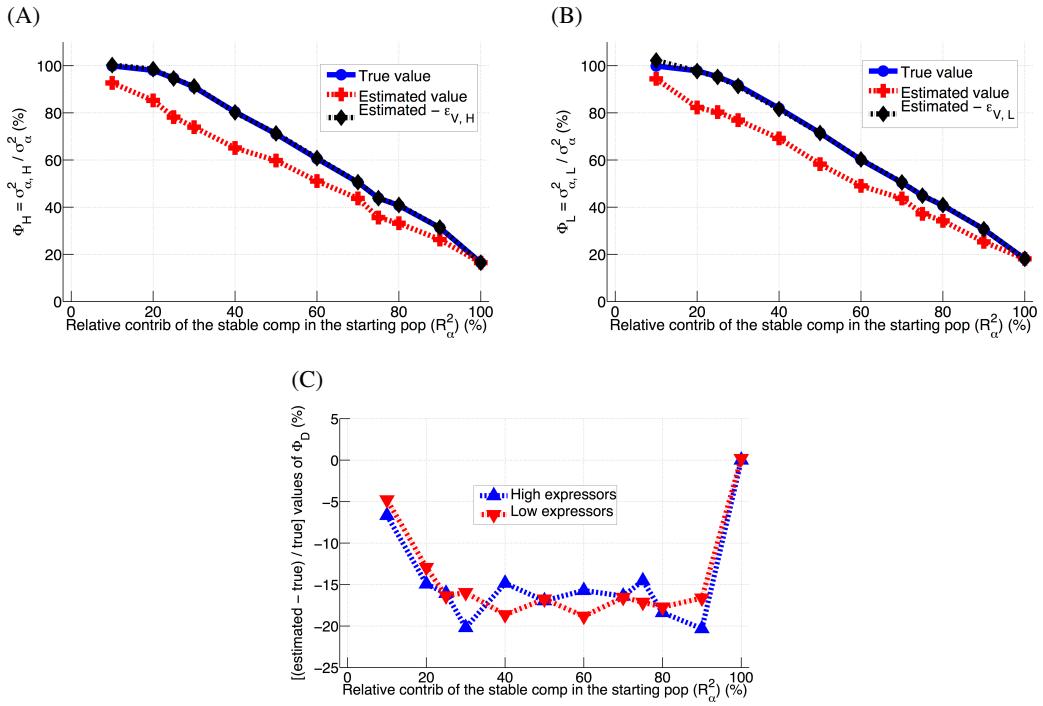

Figure 2.12: Comparison between the "true" value of Φ_D and the estimated value $\widehat{\Phi}_D$ (obtained based on equation 2.104). Each figure also includes the results of subtracting the bias (obtained as in figure 2.51) from the estimated value, to show that it explains the discrepancy between Φ_D and $\widehat{\Phi}_D$. Results are shown for $\tau = 500$, $\beta = 50$ and $\sigma_T = 0.3$.

figure, Φ_D was calculated based on data on the sub-populations composing each isolated population, while the estimate $\widehat{\Phi}_D$ was obtained based on equation 2.104. These two values are clearly different for values of R^2_α lower than 70%. Figure 2.12 also shows that, when the bias $\epsilon_{V,D}$ is accounted for (using equation 2.51, to obtain the residual component and σ^2_α), subtracting it from the estimate results in the "true" value. However, given that the bias cannot be estimated in practice, since it depends on the underlying structure of each cell population, it follows that the estimation of Φ_D via $\widehat{\Phi}_D$ is, indeed, biased in most of the cases.

Therefore, we conclude that the variance can provide additional information, allowing the estimation of the ratio between the variances due to the stable component in an isolated population, such as high or low expressors, and that of the all expressors (as a proxy for the starting population). This estimate depends on the value of R^2_α, which can be estimated using the approach outlined in section 2.5 of the main text, and the ratio between the total variances of the two populations being compared (either high and all expressors, or low and

all expressors). However, it was shown here that this estimate is biased, due to introduction of a residual component as a consequence of isolating cells based on the expression levels. This bias results typically in an under-estimation of the "true" value of Φ_D by up to 20% of the true value. Hence, we interpret these results to imply that the asymptotic (stationary) variance is, at least in principle, uninformative. Further highlighting the approach based on the means to estimate R^2_α, in the case that an estimate of Φ_D is needed, a simulation-based approach is suggested:

1. estimate R^2_α using the analysis of the means
2. using R^2_α, simulate the actual isolation of cells from the starting population, and determine Φ_D

Materials and methods

Simulations were done in exactly the same setup as those for section 2.5 of the main text, but with a much larger number of cells in the starting population (30×10^6), and the same number (2×10^4) of sub-populations. Statistics on the sub-populations were calculated based on the asymptotic mean and variance defined by the parameters $\{\alpha_i, \beta, \sigma^2_W\}$ describing each sub-population.

Chapter 2

Bibliography

Arnold, L. (1974). *Stochastic differential equations: theory and applications*. Wiley, 1 edition.

Burnham, K. P. and Anderson, D. R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York, NY.

Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E., and Huang, S. (2008). Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, 453(7194):544–7.

Cohen, A. A., Kalisky, T., Mayo, A., Geva-Zatorsky, N., Danon, T., Issaeva, I., Kopito, R. B., Perzov, N., Milo, R., Sigal, A., and Alon, U. (2009). Protein dynamics in individual human cells: experiment and theory. *PLoS ONE*, 4(4):e4901.

Dunlop, M. J., Cox, R. S., Levine, J. H., Murray, R. M., and Elowitz, M. B. (2008). Regulatory activity revealed by dynamic correlations in gene expression noise. *Nature Genetics*, 40(12):1493–8.

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584):1183–6.

Feinerman, O., Veiga, J., Dorfman, J. R., Germain, R. N., and Altan-Bonnet, G. (2008). Variability and robustness in T cell activation from regulated heterogeneity in protein levels. *Science*, 321(5892):1081–4.

Gillespie, D. T. (1996). Exact numerical simulation of the Ornstein-Uhlenbeck process and its integral. *Physical Review E*, 54(2):2084–2091.

Hemberger, M., Dean, W., and Reik, W. (2009). Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. *Nature Reviews Molecular Cell Biology*, 10(8):526–37.

Hill, W. G. and Kirkpatrick, M. (2010). What Animal Breeding Has Taught Us about Evolution. *Annual Review of Ecology, Evolution, and Systematics*, 41(1):1–19.

Huang, S. (2009). Non-genetic heterogeneity of cells in development: more than just noise. *Development*, 136(23):3853–62.

Kalmar, T., Lim, C., Hayward, P., Muñoz Descalzo, S., Nichols, J., Garcia-Ojalvo, J., and Martinez Arias, A. (2009). Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biology*, 7(7):e1000149.

Komorowski, M., Miekisz, J., and Stumpf, M. P. (2013). Decomposing Noise in Biochemical Signaling Systems Highlights the Role of Protein Degradation. *Biophysical Journal*, 104(8):1783–1793.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, 1 edition.

MacArthur, B. D., Sevilla, A., Lenz, M., Müller, F.-J., Schuldt, B. M., Schuppert, A. A., Ridden, S. J., Stumpf, P. S., Fidalgo, M., Ma'ayan, A., Wang, J., and Lemischka, I. R. (2012). Nanog-dependent feedback loops regulate murine embryonic stem cell heterogeneity. *Nature Cell Biology*, 14(11):1–11.

Market, E. and Papavasiliou, F. N. (2003). V(D)J recombination and the evolution of the adaptive immune system. *PLoS Biology*, 1(1):E16.

Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the Theory of Statistics*. McGraw Hill, 3 edition.

Munsky, B., Trinh, B., and Khammash, M. (2009). Listening to the noise: random fluctuations reveal gene network parameters. *Molecular Systems Biology*, 5(318):318.

Orkin, S. H. (2000). Diversification of haematopoietic stem cells to specific lineages. *Nature Reviews Genetics*, 1(1):57–64.

Paixão, T. (2007). *The Stochastic Basis of Somatic Variation*. PhD thesis, University of Porto.

Papoulis, A. and Pillai, S. U. (2002). *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 4 edition.

Pedraza, J. M. and van Oudenaarden, A. (2005). Noise propagation in gene networks. *Science*, 307(5717):1965–9.

Pina, C., Fugazza, C., Tipping, A. J., Brown, J., Soneji, S., Teles, J., Peterson, C., and Enver, T. (2012). Inferring rules of lineage commitment in haematopoiesis. *Nature Cell Biology*, 14(3):287–94.

Raj, A., Rifkin, S. A., Andersen, E., and van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. *Nature*, 463(7283):913–8.

Raj, A. and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–26.

Raser, J. M. and O'Shea, E. K. (2004). Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678):1811–4.

Rausenberger, J. and Kollmann, M. (2008). Quantifying origins of cell-to-cell variations in gene expression. *Biophysical Journal*, 95(10):4523–8.

Rinott, R., Jaimovich, A., and Friedman, N. (2011). Exploring transcription regulation through cell-to-cell variability. *PNAS*, 108(15):6329–34.

Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S., and Elowitz, M. B. (2005). Gene regulation at the single-cell level. *Science*, 307(5717):1962–5.

Shahrezaei, V., Ollivier, J. F., and Swain, P. S. (2008). Colored extrinsic fluctuations and stochastic gene expression. *Molecular Systems Biology*, 4(196):196.

Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Liron, Y., Rosenfeld, N., Danon, T., Perzov, N., and Alon, U. (2006). Variability and memory of protein levels in human cells. *Nature*, 444(7119):643–6.

Singh, A., Razooky, B. S., Dar, R. D., and Weinberger, L. S. (2012). Dynamics of protein noise can distinguish between alternate sources of gene-expression variability. *Molecular Systems Biology*, 8(607):1–9.

Sisan, D. R., Halter, M., Hubbard, J. B., and Plant, A. L. (2012). Predicting rates of cell state change caused by stochastic fluctuations using a data-driven landscape model. *PNAS*, 109(47):19262–7.

Spencer, S. L. and Sorger, P. K. (2011). Measuring and modeling apoptosis in single cells. *Cell*, 144(6):926–39.

Swain, P. S., Elowitz, M. B., and Siggia, E. D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. *PNAS*, 99(20):12795–800.

Yates, L. R. and Campbell, P. J. (2012). Evolution of the cancer genome. *Nature Reviews Genetics*, 13(11):795–806.

Chapter 3

THE T CELL RECEPTOR IN $CD4^+$ T CELLS AS A MODEL SYSTEM FOR THE STABLE AND UNSTABLE COMPONENTS

Author contributions

The author of the thesis designed and planned all the experiments, together with the supervisors Dr. Jorge Carneiro and Dr. Vasco Barreto. The author performed all the experiments, with the help of the supervisor Dr. Vasco Barreto in injecting the recipients for the adoptive transfer experiments (section 3.5). The results were analyzed and interpreted by the author and the supervisors Dr. Vasco Barreto and Dr. Jorge Carneiro.

This chapter is an extended version of the analysis of experimental data presented in the following manuscript:

Guzella, T. S., Barreto, V. B., and Carneiro J. (2013). Quantifying the Contributions and Dynamics Underlying Variation in Expression Levels in a Cell Population. *In preparation, under final review by the co-authors*

Abstract

A remarkable property of the adaptive immune system in jawed vertebrates is the phenotypic variation in the various lymphocyte populations, in particular with B and T lymphocytes being genetically diverse. In the case of the expression levels of a particular molecule in a lymphocyte population, variation may come about due to the genetic diversity. It may also come about due to epigenetic variation, in the general context of mechanisms that result in phenotypic differences that are not explained by differences in DNA sequence, but that persist throughout time. Finally, fluctuations in the expression levels of each cell throughout time is an additional possibility. This chapter addresses how these three mechanisms shape the variation in expression levels of the T Cell Receptor (TCR) in $CD4^+$ T cells. Analysis centers on the concepts of the stable and unstable components, developed in chapter 2. These two components lump mechanisms that result in permanent (stable component) and transient (unstable component) differences in the expression levels of two subsets of cells. By focusing on various T cell populations, one being polyclonal (genetically diverse) and two being isogenic, we quantify the relative contribution of the stable component (R^2_α) and the characteristic time of the variation (τ_T) in expression levels of the TCR. The two isogenic populations considered, Marilyn and OT-II, are TCR-transgenic on a *Rag2*$^{-/-}$ background. In the polyclonal population, both genetic and epigenetic variation have the potential to mold the stable component, while in the TCR-transgenic populations, the stable component would be due, by definition, to epigenetic variation. By evaluating the value of R^2_α in each of these populations, we assess the impact of genetic and epige-

netic variation on the stable component in the polyclonal population. We find that the best description of these populations in an *in vitro* setup, without stimulation, is one in which the stable component is the main contribution to the variation in the polyclonal population ($R^2_\alpha \approx 70\%$), and that the unstable component preponderates in the two TCR-transgenic ($R^2_\alpha \approx 10\%$ for Marilyn, and $R^2_\alpha \approx 45\%$ for OT-II). This description is based on all populations having a characteristic time of the variation equal to 37 hours. Moreover, in an *in vivo* setup, using adoptive transfers to *Rag2*$^{-/-}$ mice, we find that differences in TCR expression levels between subsets of cells can indeed persist for several weeks, consistent with the notion of the stable component of variation. Altogether, these results provide strong evidence for the stable component in the polyclonal population, and preliminary evidence for the stable component in the TCR-transgenic populations, and indicate genetic variation as the main explanatory factor for the stable component in the polyclonal population, with epigenetic variation having only a marginal impact. Therefore, this analysis establishes the TCR in a polyclonal population of CD4$^+$ T cells as a model system to study how the stable and unstable components contribute to variation in expression levels.

3.1 Introduction

One of the hallmarks of the adaptive immune system in jawed vertebrates is the generation of populations of cells that are genetically diverse. This is established by a process of somatic DNA rearrangement, termed V(D)J recombination, in which particular loci of the genome are targeted. These loci code for some of the sub-units of the antigen receptors in developing lymphocytes (Tonegawa, 1983; Market and Papavasiliou, 2003). This is the case for both developing B and T cells, whose antigen receptors are named, respectively, B cell receptor (BCR) and T cell receptor (TCR). The antigen receptors interact with antigens available throughout the body, and the specificity of the receptor influences the nature of the particular signal that will be received by the cell upon encountering a particular antigen. Consequently, populations of lymphocytes are inherently heterogeneous, composed of clones.

Another general mechanism that may lead to phenotypic variation in lymphocyte populations is, in a broad context, epigenetic. Such a denomination (see section 1.2.2, page 12) is used to refer to mechanisms that result in phenotypic differences that persist throughout time, but are not explained by differences in DNA sequence. One example is in terms of the ability of these populations to give rise to distinct cell types under certain conditions. Such a process of cell differentiation has been extensively studied in the case of a particular population of T cells, known as helper T cells (Th cells) or $CD4^+$ T cells (since they express the CD4 co-receptor). Depending on stimuli that are provided, these cells can differentiate into Th1 or Th2 cells, which have different functional properties, as shown in the seminal work of Mosmann and Coffman (Mosmann et al., 1986). This is indeed epigenetic, as even cells expressing the same TCR can differentiate into Th1 and Th2 cells (Liew, 2002). Recent works have described additional cell types, such as Th17 and Th9 (Zhu et al., 2010; Kaplan, 2013; Weaver et al., 2013). These cell types are characterized by distinct patterns of gene expression, which can be maintained once established (Wilson et al., 2009; Zhu et al., 2010). One example is in terms of what are often referred to as "master regulators" associated with each cell type, such as the transcription factors GATA3 (Th2), TBX21 (Th1), also known as T-bet, and RORγt (Th17). Hereafter, we refer to permanent differences in an isogenic population as being due to epigenetic variation.

When focusing on the expression level of a particular molecule, stochastic fluctuations are also expected to contribute to the variation. One of the processes that is thought to influence these fluctuations is noise in gene expression (reviewed in Raj and van Oudenaarden, 2008), due to the small copy number of molecules involved in the reactions governing

expression of a gene.

Variation in the expression levels in a snapshot of a cell population is a widespread observed in lymphocyte populations (Paixão, 2007; Feinerman et al., 2008), especially given the extensive use of flow cytometry for the analysis of these populations. In fact, T cells constitute an interesting model system for studying variation in expression levels. It has been described that some molecules in these cells have a component of stability in their expression levels in different cells, such as CD5 (Smith et al., 2001; Palmer et al., 2011; Mandl et al., 2013). On the other hand, studies centered on the cytokines IL-4 (Mariani et al., 2010) and IL-10 (Calado et al., 2006; Paixão et al., 2007) in T cells found that cells initially expressing these cytokines may not do so after a certain amount of time. In other words, the decision of a single cell to express these cytokines is stochastic, and is randomized after some time. For a particular molecule in a population of cells, the variation is shaped by a particular combination of the three aforementioned mechanisms, namely genetic variation, epigenetic variation, and stochastic fluctuations. However, it remains unclear to which degree each of these mechanisms contributes to the variation that is observed.

As defined in detail in chapter 2, in the context of a simplified model of constitutive protein expression, the stable component of variation leads to permanent differences in the expression levels of two subsets of cells in a population, while the unstable component implies transient differences. Therefore, both genetic and epigenetic variation mold the former component (stable), while stochastic fluctuations represent the latter (unstable). To address the contribution of the various mechanisms, this chapter focuses on the TCR in mouse $CD4^+$ T cells. As a protein complex composed of multiple sub-units, some of which are genetically distinct among clones, the expression level of the TCR may be different in each clone. This would result in a stable component of variation in such a polyclonal (genetically heterogeneous) population. To address whether epigenetic variation may also have an effect, we rely on two isogenic (genetically homogeneous) T cell populations, as being representative, in a general sense, of each clone in a genetically heterogeneous T cell population. These isogenic populations are obtained from genetically modified mice (Barnden et al., 1998; Lantz et al., 2000), in which the somatic rearrangement is ablated ($Rag2^{-/-}$ background), and the T cells express a single receptor encoded by transgenes.

This chapter is organized in the following way: section 3.2 provides a brief review of the biology of $CD4^+$ T cells, focusing on the TCR, T-cell development and population dynamics. Section 3.3 then describes the genetically homogeneous populations that will be used as approximations of the clones in a genetically diverse (polyclonal) T cell popu-

lations. Section 3.4 quantifies parameters describing the stable and unstable components of variation in expression levels of the TCR in the T cell populations considered under *in vitro* conditions, while section 3.5 further addresses the stable component in the polyclonal population in an *in vivo* setting. Finally, the conclusions of this chapter are presented in section 3.6.

3.2 A Brief Review of T Cell Biology

This section presents a review of the biology of T cells, with particular focus on $CD4^+$ T cells, which are the model system used in this work. In T cells, the TCR, a multi-subunit receptor complex expressed in the cellular membrane, interacts with ligands and, depending on the particular ligand, mediates the transduction of signals that ultimately determine the fate of the cell. Various types of T cells are known in vertebrates, and are distinguished based on the structure of the TCR and their functional properties. In animals that have not been immunized, the main T cell population in most lymphoid tissues is composed of cells that express a TCR having α and β chains, and hence sometimes referred to as $\alpha\beta$ T cells. These cells are further divided into helper or cytotoxic T cells, reflecting their different functional properties, and can be identified depending on the expression of either of two co-receptors, with helper cells being $CD4^+$, while cytotoxic cells are $CD8^+$. Other types of T cells have been described (such as $\gamma\delta$ T cells and natural killer T cells; see for example, Hayday, 2000; Bendelac et al., 2007), but will not be further mentioned, since they are not the subject of this work. Hereafter, the denomination of T cells will be used exclusively to refer to $\alpha\beta$ T cells.

T cells undergo development in the thymus, and are fundamental components of the immune system. During an immune response, pathogen-derived molecules are processed into peptides by antigen-presenting cells (APCs), and presented in the context of MHC (Major Histocompatibility Complex) molecules to T cells. This results in the "activation" of T cells that have TCRs specific for a particular peptide, a state characterized by the activation of several signaling pathways downstream of the TCR when cells receive a sufficiently strong stimulus (Smith-Garvin et al., 2009). Once activated, T cells proliferate and become capable of acquiring effector functions: $CD8^+$ T cells eliminate infected cells and produce cytokines (soluble factors), while $CD4^+$ T cells provide further stimulation signals to both other T and non-T cells, via cell-cell interactions and the production of cytokines.

The TCR is first introduced in more detail in section 3.2.1. Section 3.2.2 then reviews the processes that take place during T cell development in the thymus, with section 3.2.3

focusing on the process of V(D)J recombination, which establishes the diverse repertoire. Afterwards, the processes acting on mature T cells once they reach the periphery are briefly mentioned (section 3.2.4).

3.2.1 The T cell receptor (TCR)

The ($\alpha\beta$) TCR is composed of four modules. The $\alpha\beta$ heterodimer formed by the α and β chains is non-covalently associated with three other heterodimers, CD3γ-CD3ϵ and CD3δ-CD3ϵ, referred to as CD3$\gamma\epsilon$ and CD3$\delta\epsilon$, respectively, and $\zeta\zeta$ (Schrum et al., 2003; Kuhns et al., 2006; Wucherpfennig et al., 2010). Moreover the η chain, an isoform resulting from alternative splicing of ζ (Clayton et al., 1991), may be present, in which case the $\zeta\zeta$ homodimer is replaced by either $\zeta\eta$ or $\eta\eta$ (Bauer et al., 1991). The classification into these four modules ($\alpha\beta$, CD3$\gamma\epsilon$, CD3$\delta\epsilon$ and $\zeta\zeta$) does not necessarily reflect the order of assembly, as will be discussed below. Hereafter, we will refer to the TCR as the fully assembled complex, composed of all four modules.

Despite years of investigation, several questions concerning the TCR still remain. In particular, no high-resolution structure of an intact complex has been obtained so far. The currently held organization of the TCR is illustrated in figure 3.1, with the clonotypic $\alpha\beta$ heterodimer, the CD3 heterodimers CD3$\gamma\epsilon$ and CD3$\delta\epsilon$, and the $\zeta\zeta$ chain homodimer. In this view, the CD3 heterodimers would occupy opposite sides of the fully assembled complex (Schrum et al., 2003; Sun et al., 2004). The $\alpha\beta$ dimer constitutes the clonotypic module, which interacts with ligands, with the particular combination of $\alpha\beta$ determining the specificity of the TCR. This dimer has a very short intracellular domain, such that signaling in response to ligands is mediated by the CD3$\gamma\epsilon$, CD3$\delta\epsilon$ and $\zeta\zeta$ modules (Wucherpfennig et al., 2010). It should be noted that an alternative to the model shown in figure 3.1 has recently been proposed (Kuhns et al., 2010), where the the CD3$\gamma\epsilon$ and CD3$\delta\epsilon$ heterodimers would be side-by-side to one another (see also Kuhns and Badgandi, 2012). Both models are, nevertheless, based on the prevailing view of the TCR as monovalent in resting cells (Punt et al., 1994; Call et al., 2002, 2004; Wucherpfennig et al., 2010; Kuhns and Davis, 2012), as the occurrence of bivalent TCRs, with two $\alpha\beta$ dimers per complex, instead of a single one (Fernández-Miguel et al., 1999; Schrum et al., 2011) remains highly controversial.

The assembly of the six sub-units that are present in the TCR is highly regulated, such that only fully assembled complexes are expressed on the membrane (Klausner et al., 1990). Assembly takes place in the endoplasmic reticulum, with the quick degradation of nascent

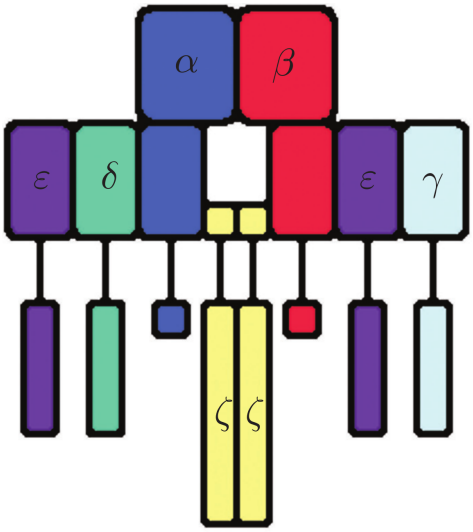

Figure 3.1: Organization of the TCR, highlighting the components of the fully assembled complex: the clonotypic $\alpha\beta$ heterodimer, besides the invariant modules CD3$\gamma\epsilon$, CD3$\delta\epsilon$ and $\zeta\zeta$. Figure adapted from Schrum et al. (2003) (see also Sun et al., 2004).

sub-units and intermediates (Call and Wucherpfennig, 2005). It is still unclear if assembly takes place via an ordered set of steps, or whether alternative intermediates may be generated. The process has been investigated with a multitude of experimental systems, ranging from an *in vitro* translation system with ER microsomes (Call et al., 2002, 2004), to non-T cell lines (Manolios et al., 1991), T cell lines and hybridomas (Minami et al., 1987; Geisler, 1992), and primary T cells (Kearse et al., 1995). The present data suggest that assembly may occur as indicated in figure 3.2, highlighting the unassembled sub-units, the intermediates that have been reported, and the fully assembled complex. The main assembly sequence is thought to be the formation of the two CD3 heterodimers CD3$\gamma\epsilon$ and CD3$\delta\epsilon$, followed by their preferential association, respectively, with the TCRα and TCRβ monomers, resulting in the intermediate TCR$\alpha\beta$-CD3$\gamma\epsilon$-CD3$\delta\epsilon$, which would then pair with $\zeta\zeta$ to form the fully assembled complex (Dave, 2009). This particular sequence is based, in part, on a work using primary T cells (Kearse et al., 1995), which reported that most of the intermediates containing $\alpha\beta$ dimers would be formed via the association of TCRα-CD3$\delta\epsilon$ and TCRβ-CD3$\gamma\epsilon$. However, this sequence may not be the only one, as studies based on others experimental systems have found evidence of other intermediates. In particular, early works reported that the $\alpha\beta$ dimer may be formed in the absence of CD3 components in hybridomas and transfected non-T cell lines (Bonifacino et al., 1988; Mano-

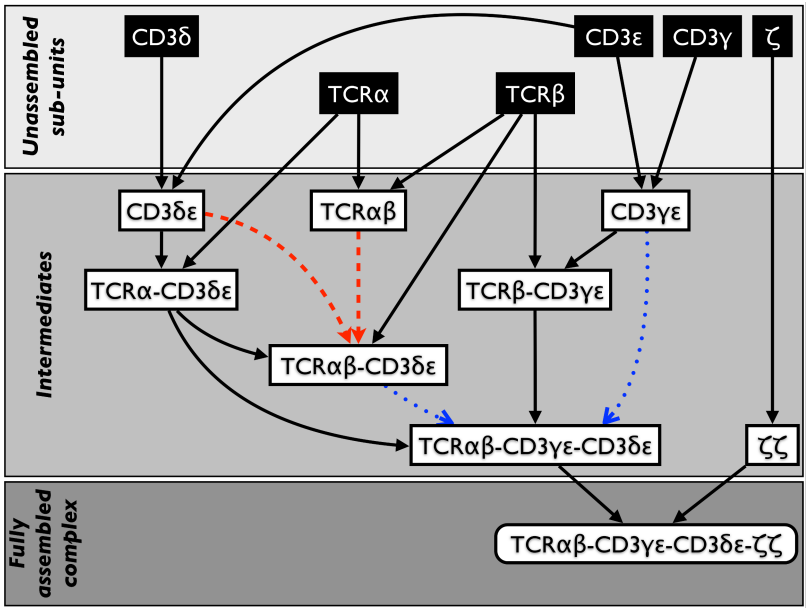

Figure 3.2: Diagram depicting the steps in assembly of the TCR, starting from the unassembled chains (TCRα, TCRβ; filled blocks), with the intermediates that have been reported, until the fully assembled complex (TCR$\alpha\beta$-CD3$\gamma\epsilon$-CD3$\delta\epsilon$-$\zeta\zeta$). Each intermediate and the fully assembled complex are formed by pairing two components. The pair of matching arrows pointing to each block denote the association between the two source components that produce the resulting intermediate or the fully assembled complex. Two intermediates, namely TCR$\alpha\beta$-CD3$\delta\epsilon$ and TCR$\alpha\beta$-CD3$\gamma\epsilon$-CD3$\delta\epsilon$, may be assembled in two different ways, each way being indicated by a pair of matching arrows. For example, TCR$\alpha\beta$-CD3$\delta\epsilon$ may be assembled via the association of TCR$\alpha\beta$ and CD3$\delta\epsilon$, or TCRα-CD3$\delta\epsilon$ and TCRβ (see text for details).

lios et al., 1991). Upon association with the CD3$\gamma\epsilon$, CD3$\delta\epsilon$ and $\zeta\zeta$ dimers, this would also produce the fully assembled complex. Furthermore, the intermediate TCR$\alpha\beta$-CD3$\delta\epsilon$ has also been reported in a T cell line lacking CD3γ (Geisler, 1992). It is possible that this intermediate, would, in CD3γ-sufficient cells, associate with CD3$\gamma\epsilon$, and subsequently with $\zeta\zeta$, thereby also producing the fully assembled complex. It remains unclear what is the actual contribution, if any, of the intermediates $\alpha\beta$ and TCR$\alpha\beta$-CD3$\delta\epsilon$ to assembly of the TCR in primary T cells.

Since assembly of the TCR depends on a complex multi-step process, where all subunits are indispensable for final assembly, any single one, or a combination of them, may be rate-limiting for formation of the fully assembled complex (Schrum et al., 2003). An

early work relied on pulse-chase with metabolic labeling to analyze the turnover of sub-units and various intermediates in hybridoma cells, and concluded that the ζ chain would be the limiting sub-unit for assembly of the full complex (Minami et al., 1987). In support of this notion, later studies observed that reconstitution of $\zeta^{-/-}$ mice with transgenes over-expressing the ζ chain resulted in total T cells with upregulated surface TCR levels, in comparison with cells from wild-type mice (Love et al., 1993; Shores et al., 1994). Furthermore, Azzam et al. (1998) reported that in $\zeta^{+/-}$ mice, cells in a particular stage of T cell development, known as double-positive (DP; see section 3.2.2), have reduced TCR levels in comparison with cells in the same stage from $\zeta^{+/+}$ mice, although such a difference may not hold for (mature) T cells, since the DP stage is characterized by TCR levels around 10-fold lower than that of T cells (Schrum et al., 2003). Consequently, it has been proposed that the ζ chain would be the rate-limiting sub-unit for assembly of the TCR (Baniyash, 2004). On the other hand, pairing constraints between α and β chains have long been considered to be an important factor in TCR expression, especially during T cell development (Brady et al., 2010), when these chains are rearranged. In particular, if a particular pair of β and α chains do not efficiently pair (Saito et al., 1989), expression of these two chains may become the rate-limiting step for assembling the complex.

Once the full complex has been assembled, it is then transported to the Golgi compartment, before finally reaching the cellular membrane (Hayes et al., 2003). As such, TCRs are very stable, constantly internalized from the cellular membrane to intracellular compartments and re-exported back. Finally, most of the total TCR (70–75%) in resting $CD4^+$ T cells is expressed on the membrane (Liu et al., 2000).

Therefore, the TCR, as expressed on the membrane of cells, is a protein complex resulting from the multi-step assembly of sub-units. While receptors composed of multiple sub-units are a widespread observation in various cell types (Hynes, 2002; Paoletti et al., 2013), a particular feature of the antigen receptors is that some of its sub-units are genetically diverse, as established by a process of somatic rearrangement taking place during lymphocyte development. T cells undergo development in the thymus, as will be reviewed in the next section.

3.2.2 T cell development in the thymus

This section presents a brief overview of the development of T cells in the thymus. One of the outcomes of this process is the generation of a genetically diverse cell population, due to the somatic rearrangement of the α and β chains. Detailed discussions of the molecular and

cellular events taking place during this process can be found in recent reviews (Germain, 2002; Starr et al., 2003; Singer et al., 2008; Koch and Radtke, 2011).

Signaling via the TCR depends on the interaction with peptide-MHC complexes presented by APCs. This interaction shapes the repertoire of T cells that complete development in the thymus, through the processes of positive and negative selection, which are reviewed in the following.

Development of $\alpha\beta$ T cells starts with thymocytes that are in the double-negative (DN) stage, in which neither the CD4 nor CD8 co-receptors are expressed. This stage can be further divided into four sub-stages, based on the expression of the surface markers CD25 and CD44, which are, in sequence: DN1 ($CD44^{+}CD25^{-}$), DN2 ($CD44^{+}CD25^{+}$), DN3 ($CD44^{-}CD25^{+}$) and DN4 ($CD44^{-}CD25^{-}$). The DN stage is marked by rearrangement of the β chain (DN2–DN3), which must successfully pair with the invariant pTα chain, to form a functional pre-TCR and rescue the corresponding cell from programmed cell death. This process is also known as β-selection, since it selects for cells that have productively rearranged and paired the β chain. After β-selection, the surviving cells undergo proliferation and up-regulation of both CD4 and CD8, in what is known as the double-positive (DP) stage. In this stage, rearrangement of the α chain takes place, along with pairing with the previously rearranged β. At this point, positive selection skews the repertoire for cells that express a functional TCR and that can receive TCR-derived signals upon interaction with APCs, since the cells that fail to receive signals undergo apoptosis. This is followed by commitment to either the CD4 or CD8 lineages, also referred to, respectively, as single-positive (SP) CD4 or CD8 stages. Afterwards, an additional process, known as negative selection, is marked by the elimination (deletion) of the cells that receive strong TCR-derived signals upon interaction with APCs. As a consequence of positive and negative selection, the $CD4^{+}$ and $CD8^{+}$ T cell populations that develop have an intermediate "strength" of interaction with peptides presented by the APCs, and it is estimated that less than 5% of the developing cells survive both processes (Starr et al., 2003). Finally, the surviving SP thymocytes have completed development, being henceforth referred to as T cells, and migrate to the periphery.

3.2.3 V(D)J recombination

One distinguishing feature of T lymphocytes is that they constitute genetically diverse cell populations, consisting of sets of clones. Hence, such a population of T cells is referred to as being polyclonal. The genetic variation is established by the somatic rearrangement of

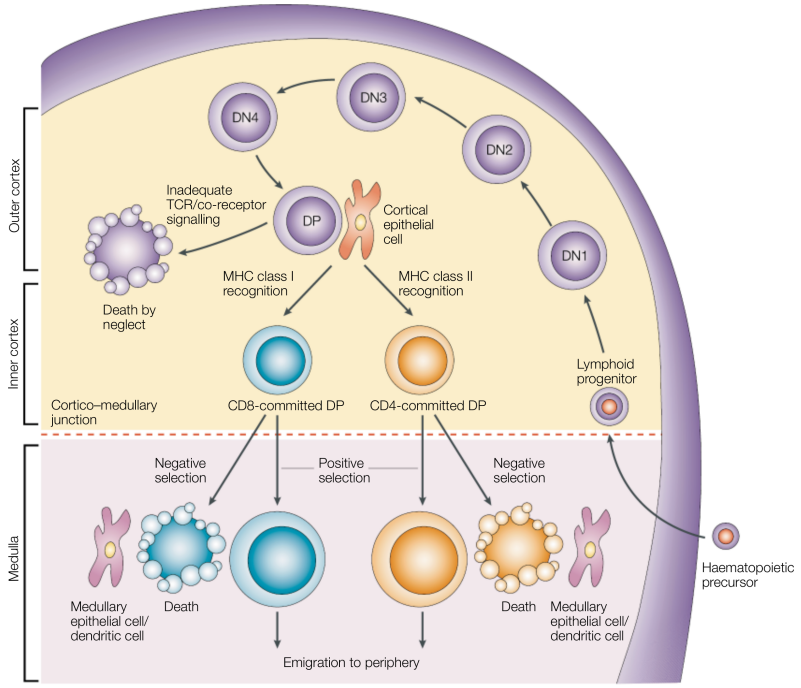

Figure 3.3: Overview of T cell development in the thymus, highlighting the DN and DP stages, CD4/CD8 lineage choice, positive and negative selection, until export to the periphery. Figure adapted from Germain (2002).

the loci coding for the α and β chains of the TCR, consisting of the rearrangement of V (variable), D (diversity) and J (joining) regions into a new gene.

The germline (non-recombined) genomic organization of the TCRα and TCRβ loci are shown in figure 3.4. The TCRα locus is intermixed with TCRδ (which is rearranged in $\gamma\delta$ T cells), and is located in mouse chromosome 14, spanning around 2 Mbp (Bosc and Lefranc, 2003). The α chain is formed via the recombination of one Vα gene, out of the 70–80 present (the actual number depending on the particular mouse haplotype), with one of the 44 functional Jα genes, and the single Cα (Jouvin-Marche et al., 2009). Furthermore, the Eα enhancer is located 3' of the single Cα. Upon recombination of the TCRα locus, the D-J-C region of the TCRδ locus is deleted.

In turn, the TCRβ locus spans approximately 700 kbp in mouse chromosome 6. It has a total of 22 functional Vβ segments, and includes the D regions, which are not present in the TCRα locus. The locus has undergone duplication of the D-J-C region, which are therefore named Dβ1-Jβ1-Cβ1 and Dβ2-Jβ2-Cβ2, each containing a single D region, and,

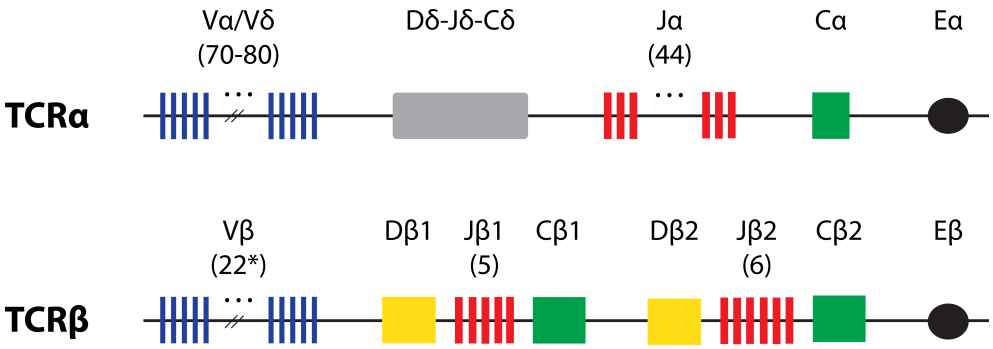

Figure 3.4: Genomic organization of the germline mouse TCRα and TCRβ loci. Each locus is represented by a set of V segments (blue rectangles), D segments (yellow rectangles; in the case of the TCRβ locus only), J segments (red rectangles), C regions (green rectangles) and enhancers (ovals). The number of functional segments, if greater than one, is shown in brackets; one of the Vβ segments, which is located 3' of the Eβ enhancer, has been omitted for simplicity, but accounted for in the total number of segments (as denoted by the asterisk). In the TCRα locus, the segments corresponding to TCRδ (Dδ, Jδ and Cδ) are summarized as a single cluster (gray rounded rectangle). The figure is based on references Paul (2003); Brady et al. (2010), updated with information from the international ImMunoGeneTics information system® (IMGT) (Bosc and Lefranc, 2003; Bosc et al., 2011), and is not drawn to scale.

respectively, 5 and 6 functional J genes. Due to this duplication, the constant β region may be encoded by either of two "regions" (Cβ1 and Cβ2), but that are virtually identical in terms of protein sequence (Gascoigne et al., 1984). The enhancer for the TCRβ locus, referred to as Eβ, is located 3' of Cβ2. In both the TCRα and TCRβ loci, each V segment is associated with its promoter, which upon recombination is brought in close proximity of the respective enhancer. The variation in the regulatory elements present in the promoters within the Vα and Vβ genes (Chen et al., 2001; Glusman et al., 2001) is an additional mechanism, besides inefficient pairing of the α and β chains (Saito et al., 1989), with the potential to result in differences in expression levels among cells that have rearranged a particular Vα-Vβ combination. Using semi-quantitative RT-PCR, one study (Chen et al., 2001) focusing on the Vβ genes observed differences in the germline expression levels in the absence of recombination (*Rag2*$^{-/-}$ thymocytes), but claimed that many of these differences seemed to be canceled in total thymocytes of wild-type animals, although without a quantitative analysis. This effect was attributed to the strong effect of the Eβ enhancer, as revealed by the much increased expression level of each Vβ gene when comparing *Rag2*$^{-/-}$ and total wild-type thymocytes (Chen et al., 2001).

The rearrangement of the TCRα and TCRβ loci is dependent on the RAG complex, formed by the association between the RAG1 and RAG2 proteins (Schatz et al., 1989; Oettinger et al., 1990). A recombination event takes place upon binding of the RAG complex to RSSs (recombination signal sequences; see, for example, Swanson, 2004), flanking the gene segments, leading to the introduction of a double strand break (DSB). The DSB is then processed by components of the non-homologous end joining (NHEJ) repair pathway (Gellert, 2002). At this point, a distinguishing feature is the random addition of nucleotides, mediated by the lymphocyte-specific enzyme TdT (terminal deoxynucleotidyl transferase; Alt and Baltimore, 1982), in the junction between the two segments being assembled. During the resolution of the DSB, some nucleotides are also lost in this junctional region (Gellert, 2002).

Since it involves the introduction of a DSB, V(D)J recombination is subject to stringent control. Various regulatory mechanisms acting to restrict recombination to particular cell-types (whereby developing B and T cells rearrange, respectively, the immunoglobulin and TCR loci), loci (rearrangement of TCRα occurs only after TCRβ has been successfully rearranged), besides a well-defined order of assembly of gene segments in the case of TCRβ (Dβ $\rightarrow$ Jβ, followed by Vβ $\rightarrow$ DβJβ) (Cobb et al., 2006). An additional mechanism operating during the recombination process is allelic exclusion, strongly favoring the rearrangement of a single allele of each locus (Brady et al., 2010), such that each cell expresses a single TCR specificity. In this context, a well-documented mechanism that contributes to allelic exclusion is feedback inhibition (see, for example, Mostoslavsky et al., 2004), in which rearrangement of a functional allele, resulting in protein expression and assembly of the receptor, transduces a signal that shuts down expression of the RAG1 and RAG2 proteins, thereby reducing the likelihood of rearranging the other allele.

Based solely on the combinatorial assembly of gene segments, the maximum number of distinct $\alpha\beta$ heterodimers that could be formed is estimated on the order of 10^6 (Glusman et al., 2001; Davis and Bjorkman, 1988). However, this number increases to 10^{15} upon accounting for the diversity of the junctional region (Davis and Bjorkman, 1988). Even when considering the estimated reduction of two orders of magnitude as a consequence of thymic selection, the expected 10^{13} distinct TCRs (Nikolich-Zugich et al., 2004) indicates the considerable potential genetic diversity of a T cell population produced by the thymus. However, the actual diversity of the TCR repertoire is shaped by the population dynamics of mature T cells in the "periphery" (peripheral lymphoid organs), upon thymic export, as reviewed in the next section.

3.2.4 T cell population dynamics

Upon export from the thymus, T cells circulate throughout the body, continuously interacting with APCs. The majority of cells that have recently left the thymus display a "naive" phenotype, in the sense that they have not been activated upon interaction with peptide-MHC complexes that are continuously presented by APCs. Naive cells are operationally defined based on the expression of a particular set of markers (see, for example Lee et al., 1990; Farber et al., 1995; Caramalho et al., 2003). The size and composition of mature T cell populations is regulated by several homeostatic mechanisms, with signals such as peptide-MHC complexes and cytokines, for example IL-7, inducing the survival and under some circumstances the activation and proliferation of cells (Surh and Sprent, 2008).

3.3 TCR-transgenic T Cell Populations as Approximations of Monoclonal Populations

The variation in expression levels that is observed in a given cell population can be partitioned into contributions from two components, one stable and the other unstable. The stable component may come about due to two mechanisms shaping the expression levels, genetic variation and also, in a broad sense, epigenetic variation. As discussed in chapter 2, the stable component is a consequence of different average rates of protein production. In the case of the membrane-expressed TCR, the rate of protein production lumps together expression of the sub-units, assembly of the sub-units into the complex, and export to the cellular membrane. In considering that the TCR is a protein complex composed of multiple sub-units, one mechanism dependent on genetic variation that has the potential to result in a stable component is variation in the ability of the different α and β chains to pair, and hence affecting assembly of the complex (Saito et al., 1989). Therefore, as both genetic and epigenetic variation may contribute to the stable component in a polyclonal population, the analysis of isogenic T cell populations, allows to assess whether genetic variation would be the only explanation for a putative stable component. For this analysis, we rely on TCR-transgenic mouse strains on a *Rag2*$^{-/-}$ background as sources of isogenic T cell populations, as an initial approximation of the clones in a polyclonal population. In the following, we present an initial characterization of these populations.

The use of TCR-transgenic mice has been instrumental in immunology (see, for example, Lafaille, 2004), as these mice constitute a source of a genetically homogeneous T cell population, in most times of specificity to ligands that are known. In this work, we

rely on two well-known $CD4^+$ TCR-transgenic strains, Marilyn (Lantz et al., 2000) and OT-II (Barnden et al., 1998). This allows the quantification of TCR expression levels in genetically homogeneous T cell populations. The Marilyn TCR-transgenic strain was generated by the group of James Di Santo (Lantz et al., 2000), using α and β chains derived from a $CD4^+$ Th1 clone (Gallucci et al., 1999). This clone is specific for a peptide (Dby) derived from the male antigen H-Y. For this reason, all T cells are deleted during negative selection in male mice, such that only females have mature T cells (Lantz et al., 2000). The OT-II strain is specific for a peptide from ovalbumin (OVA), and was generated by Francis Carbone and co-workers, using a T cell clone derived from immunization of a wild-type mouse with OVA. Both Marilyn and OT-II are widely used (see, for example, Grandjean et al., 2003; Moon et al., 2007; Guimond et al., 2009; Tubo et al., 2013). A basic characterization of these two strains is shown in table 3.1. Both rely on shuttle vectors (also known as cassette vectors), which are general constructs, often used to drive expression of the α and β chains in cell lines. These mouse strains were generated via standard transgenic approaches, whereby the constructs are co-injected into fertilized eggs. In this approach, the typical outcome is co-integration of the α and β transgenes, as multiple copies in tandem in a particular region of the genome (Greenberg et al., 1991; Babinet, 2000).

In the analysis of the polyclonal population, we focused on cells that have a naive phenotype, as operationally defined by high levels of the CD45RB marker (Lee et al., 1990) (hereafter referred to as $CD45RB^{high}$), and also negative for CD25. This results in a phenotypically more homogeneous population, avoiding potentially confounding effects of certain specialized T cell types, such as regulatory T cells (Sakaguchi et al., 2008). We hereafter refer to this population of naive cells as simply polyclonal population. For the two TCR-transgenic strains, always used in a *Rag2*$^{-/-}$ background, thereby ensuring that the resultant T cell population is, indeed, genetically homogeneous. Under these conditions, virtually all cells from these two mouse strains are described as having a naive phenotype (Lantz et al., 2000; Moon et al., 2007). For consistency with the setup used for cell sorting in the next section (section 3.4), analysis was done in terms of $CD62L^+Lineage^-$ cells (see methods for details), based on the identification of naive cells in terms of expression of the marker CD62L (Lantz et al., 2000), and to avoid potential contaminants and/or unspecific staining, To quantify (surface) TCR levels, we relied on surface staining using an anti-TCRCβ antibody (hereafter referred to as anti-TCRβ), which binds to the constant region of the β chain (see, for example, Ghendler et al., 1998), being therefore independent of the actual TCR specificity.

Mouse line	Ref.	Vα-Vβ family	α chain transgene		β chain transgene	
			Vector used	# copies	Vector used	# copies
Marilyn	Lantz et al. (2000); [1]	Vα1.1 Vβ6	Shuttle expression vector	6–8	Shuttle expression vector	NA
OTII	Barnden et al. (1998); [2]	Vα2 Vβ5	Shuttle expression vector	NA	Shuttle expression vector	NA

Table 3.1: Characterization of the TCR-transgenic mouse lines considered in this work. The information on the promoters and number of copies was obtained by contacting the respective authors. NA: not available.

We first sought to compare TCR expression levels in these three T cell populations (figure 3.5). Consistent with reports comparing various TCR-transgenic populations (Grandjean et al., 2003; Kassiotis et al., 2003), Marilyn and OT-II have distinct histograms of TCR expression levels in comparison with the (naive) polyclonal population (figure 3.5A). An analysis of TCR levels based on staining for CD3ϵ led to equivalent results (figure 3.11, appendix 3.A, page 133). We also do not find a relationship between the differences in TCR expression levels and the forward-scatter, often used in flow cytometry as an initial approximation of cell size, on any of the three populations, indicating that the different expression levels do not seem to be associated with differences in cell size (figure 3.12). The quantification of the median TCR levels, based on untransformed values, reveals that both TCR-transgenic populations have statistically significant higher median TCR levels when compared with the polyclonal population, Marilyn and OT-II having, respectively, 20–35% and 15–25% higher medians (figure 3.5B). In other words, the average cell from Marilyn and OT-II expresses higher TCR levels than the average cell from the polyclonal population. These data are consistent with expression of functional TCR not being limited by excessively low expression of the transgene-driven α and β chains, but by other sub-units, such as the ζ chain (Baniyash, 2004).

Based on the theoretical framework developed in chapter 2, the total variation in a population (σ^2_T) is given by the variance of log-transformed values (figure 3.5C). In the case

[1] Dr. Olivier Lantz, personal communication
[2] Dr. Francis Carbone, personal communication

of lognormal distributions, as considered here, σ_T^2 can be readily related to the coefficient of variation (CV), a commonly used measure of dispersion, using equation 2.54 (page 65). Based on the data from the three populations, the CV is around 19% for Marilyn and OT-II, and around 26% for the polyclonal population. In considering the total variation σ_T^2, both Marilyn and OT-II have statistically significant lower values of σ_T^2 than the polyclonal population. Neglecting measurement noise, the estimated ratio between the values of σ_T^2 of the polyclonal population and Marilyn ranges around 2.2–2.6-fold, while between the polyclonal and OT-II ranges around 1.9–2.2-fold, for OT-II. These estimates are based on the average values obtained for σ_T^2 when staining using anti-TCRβ, and also anti-CD3ϵ (figure 3.11) in the two independent experiments done.

Using the estimates of ratios between the values of σ_T^2, it is possible to obtain an initial estimate of the value of R_α^2 of the polyclonal population, by making three assumptions. First, assuming that genetic variation is the only mechanism underlying the stable component in the polyclonal population, every clone in this population would have $R_\alpha^2 = 0$. Second, assuming that the TCR-transgenic populations are representative of the clones in the polyclonal population, in the sense that expression of a functional TCR is not altered due to the transgenes driving the α and β chains. This would imply that the TCR-transgenic populations would also have $R_\alpha^2 = 0$. The additional assumption is that measurement noise, which influences the above estimates of σ_T^2, is negligible. These assumptions imply that the TCR-transgenic populations are equivalent to sub-populations, a concept originally introduced in section 2.2 (page 36) to denote a set of cells where the unstable component is the only contribution to the variation in expression levels that is observed, having therefore $R_\alpha^2 = 0$. Consequently, equation 2.18 (page 43) can be used to obtain an estimate of the value of R_α^2 of the polyclonal population, denoted by $R_{\alpha,P}^2$:

$$R_{\alpha,P}^2 = 1 - \frac{1}{\sigma_{T,P}^2 \Big/ \sigma_{T,tg}^2} \tag{3.1}$$

where $\sigma_{T,P}^2$ and $\sigma_{T,tg}^2$ correspond to the variances of log-transformed values of the polyclonal and TCR-transgenic populations, respectively. Plugging in the values of the ratio $\sigma_{T,P}^2 \Big/ \sigma_{T,tg}^2$, one finds that $R_{\alpha,P}^2$ would range around $55\% - 62\%$, based on Marilyn, and around $48\% - 55\%$, based on OT-II. Hence, in the scope of the aforementioned assumptions, the stable component is expected to be present in the (naive) polyclonal population, contributing with around 50–60% of the total variation observed in that population. By maintaining the assumptions that the TCR-transgenic are representative of the clones in the

polyclonal population and that measurement noise is negligible, it follows that epigenetic variation would result in increased estimates of R^2_α for the polyclonal population, when compared with the initial estimate of 50–60%. Besides, such a scenario would imply that the TCR-transgenic populations and the clones are described by positive values of R^2_α.

Therefore, based on this preliminary analysis, the stable component is expected to be present in the polyclonal population. However, the snapshots of TCR expression levels in these three cell populations do not provide direct estimates of the contribution of the stable and unstable components of variation, quantified by R^2_α, or the characteristic time of the variation (τ_T). Therefore, the next section focuses on directly estimating R^2_α and τ_T for each of the three T cell populations, based on the setup of isolating cells.

3.4 Quantifying the Origin and Timescale of Variation in Levels of the T Cell Receptor

The analysis done in the previous section provided initial evidence for a stable component of variation in the polyclonal population. Hence, this section directly addresses the question of the origin and timescale of variation in TCR expression levels. This is based on estimating R^2_α and τ_T for the (naive) polyclonal population, along with the two TCR-transgenic *Rag2*$^{-/-}$ populations Marilyn and OT-II. In this analysis, a simplified setup is adopted, in which high and low expressors, defined as around 10% of cells expressing respectively the highest and lowest expression levels in the starting population, are sorted and then maintained *in vitro*. Under these well-established conditions, neither stimulation or cell division are expected, and cells slowly die off (Deenick et al., 2003), such that after 3 to 4 days no live cells are left.

In this setup, we are interested in comparing the values of R^2_α and τ_T estimated for the polyclonal and the two TCR-transgenic populations Marilyn and OT-II. Based on the initial analysis done in the previous section, we would expect $R^2_\alpha > 0$ for the polyclonal population. On the other hand, by focusing on an isogenic population (TCR-transgenic), we sought to address whether genetic variation may be the sole explaining factor for the stable component. In the affirmative case, one would obtain $R^2_\alpha = 0$ for a TCR-transgenic population. Otherwise, epigenetic variation would result in $R^2_\alpha > 0$.

The experimental data, shown separately for each population in terms of the fold-ratio between the median intensities of untransformed values, is presented in appendix 3.B. Importantly, figure 3.16 (appendix 3.B, page 138) shows that the staining for the sorting does not lead to changes in TCR expression levels, as verified by comparing "all expressors"

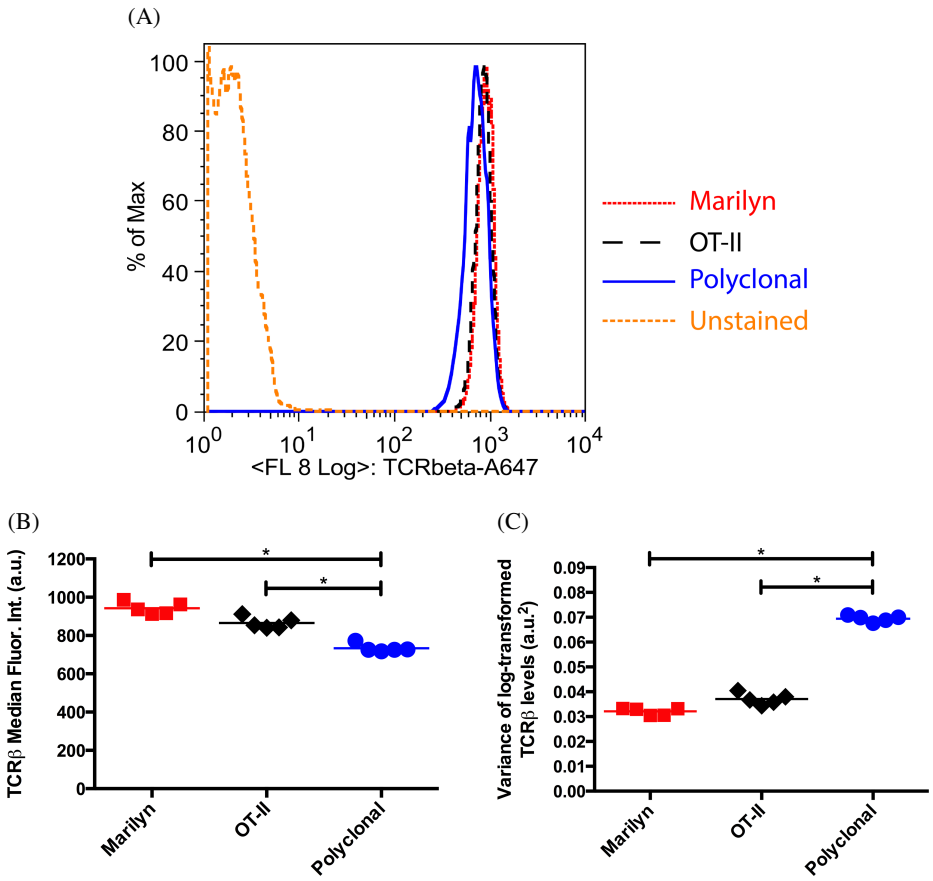

Figure 3.5: Comparison of TCR expression levels between the TCR-transgenic *Rag2*$^{-/-}$ populations Marilyn and OT-II, along with a naive polyclonal population, based on staining for TCRβ. For consistency with the setup in the section to follow (section 3.4), analysis of all populations was done based on the same gating strategy used for cell sorting (see methods), followed by gating for TCRβ^{+}CD4^{+} events. **(A)** Illustrative histograms of TCR levels of the populations. **(B)** Quantification of the median TCR levels, considering untransformed values. **(C)** Quantification of the variance of log-transformed values (σ_T^2). Data correspond to the first of 2 independent experiments, with 5 mice per group. The data from the second experiment are included in appendix 3.A. $* \; p < 0.05$.

with a control population sorted without staining for the TCR, and hence that cells are not perturbed by this staining. Moreover, the maximum values of the fold-ratio range, for the three T cell populations analyzed, from 1.6–1.8-fold (polyclonal), to around 1.5-fold (Marilyn) and 1.4 (OT-II) (figure 3.16).

An overview of the data in the form of function $\Delta_{H,L}(t)$ that was used for estimating R^2_α and τ_T is shown in figure 3.6. For the polyclonal and Marilyn populations, the data consist of three independent experiments for each, while, for OT-II, data on only two experiments are available, due to limitations in the number of mice obtained when breeding this line. For this reason, in principle the parameter estimates for this population are expected to be more uncertain, due to small data sample size. In spite of this, including an additional TCR-transgenic populations allows for a broader analysis, since it is possible that the isogenic populations may be better described by different values of R^2_α and τ_T. Therefore, the ensemble of the data used for analysis here consists of a total of eight independent experiments. However, the data on the OT-II TCR-transgenic shows considerable experimental variation in the value of $\Delta_{H,L}(t)$ in the time-points of 48 and 60 hours of one of the experiments (in figure 3.6, the two data-points with lower values of $\Delta_{H,L}(t)$). This is attributed to experimental variation because, in these two time-points, one of the replicates (wells) of high expressors has an estimated mean of log-transformed values that is considerably lower than that for the other two replicates. These replicates are, nevertheless, included in the data for analysis, as there is no evidence that they constitute problematic estimates, but simply a consequence of experimental variation to which the setup is subject to. Of note, this variation in the values of $\Delta_{H,L}(t)$ is in contrast to the remaining time-points, which are in good quantitative agreement between the two experiments.

Two initial insights can be gained from the data in figure 3.6. First, $\Delta_{H,L}(0)$ is clearly greater for the polyclonal population, as expected given that the two TCR-transgenic lines have lower total variation (σ^2_T) when compared with that population (section 3.3). Second, the data on function $\Delta_{H,L}(t)$ in figure 3.6 show that the condition $\Delta_{H,L}(t) \leq \Delta_{H,L}(0)\ \forall t$ (equation 2.27, page 50) for inference is overall well-satisfied for these data. This validates the use of this experimental setup for inferring the parameters that quantify the notions of the origin (R^2_α) and timescale (τ_T) of the variation under these *in vitro* conditions.

In order to estimate R^2_α and τ_T, fitting is done using an alternative formulation combining equations 2.22 and 2.23 (section 2.5.2):

$$\Delta_{H,L}(t) = \delta_0 \left(R^2_\alpha + \left(1 - R^2_\alpha\right) \exp\left(-t/\tau_T\right)\right) \tag{3.2}$$

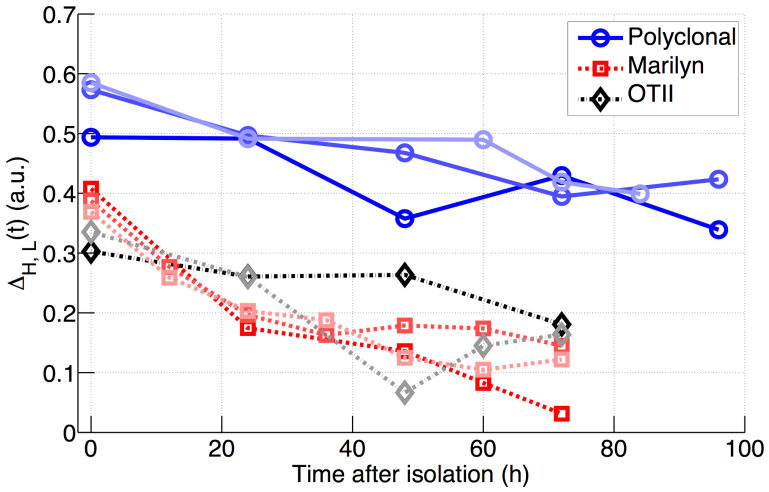

Figure 3.6: Overview of the experimental data, in terms of point estimates for $\Delta_{H,L}(t)$ in different time instants after isolation. Symbols represent point estimates (Marilyn: squares; OT-II: diamonds; polyclonal: circles), while broken lines are a linear interpolation. Each set of symbols connected by lines represents an experiment with the respective population (Marilyn, OT-II or polyclonal).

where δ_0, introduced as an additional parameter, represents an estimate obtained via fitting of the "true" initial value $\Delta_{H,L}(0)$. Equation 3.2 has the important property of preserving the statistical independence between data points used as input for the fitting, a key requirement for most statistical analyses. Given parameters R^2_α and τ_T, the function $\Omega_{H,L}(t)$ is then estimated using the original definition from equation 2.23 (section 2.5.2):

$$\Omega_{H,L}(t) = R^2_\alpha + \left(1 - R^2_\alpha\right) \exp\left(-t/\tau_T\right) \tag{3.3}$$

Analysis is based on fitting the three-parameter exponential model (equation 3.2) via non-linear least squares (Seber and Wild, 2003) to the ensemble of the data. In this way, the different models being tested are obtained by specifying each of the three parameters for each biological population as being shared or not between the three biological populations. Since experimental variation is assumed to affect parameter δ_0 only, due to small variations in defining the percentages for sorting high and low expressors in different experiments, this parameter was always fit separately for each experiment. Therefore, the models are obtained by defining whether R^2_α and τ_T are constrained or not to be the same in the different biological populations.

The models considered are listed in table 3.2, ordered based on the number of param-

eters that are estimated in each one of them. Additional models that were also considered are briefly mentioned afterwards. Model 1 represents the null model, according to which all three biological populations are described by the same values of R^2_α and τ_T, having therefore the smallest number of parameters considered. Model 2 considers that the three biological populations differ only in R^2_α, having therefore equal τ_T, with model 3 corresponding to the converse case, in which τ_T is different in each of the populations, but R^2_α is the same, while in model 4, the populations differ in both R^2_α and τ_T. Finally, model 5 provides a lower bound in the discrepancy of the fitting, where data from each experiment is fit separately, and has the largest number of parameters. To compare the different models, the Akaike Information Criterion (AIC; Burnham and Anderson, 1998), a standard approach for comparing different models fit to the same data, is used. The AIC has a strong theoretical basis, based on information theory, representing a compromise between the discrepancy in the fitting and the number of parameters in the model. For analysis, the small-sample-size version of the AIC, termed AIC$_c$ (Burnham and Anderson, 1998) is used, which further penalizes models that have an increased number of parameters. The results are presented in terms of the difference ΔAIC$_c$ between the AIC for each model and the minimum value of the AIC obtained for the models in table 3.2. In comparing different models, the one with the smallest value of the AIC (and therefore, the smallest value of ΔAIC$_c$) provides the most parsimonious approximation to the data, with $\Delta\text{AIC}_c > 10$ typically indicating strong evidence against a particular model (Burnham and Anderson, 1998).

The results of model fitting, including the sum of squared residuals, which quantifies the discrepancy in the fitting, point estimates of the parameters, and the value of ΔAIC$_c$ are shown for each model in table 3.3. Model 3 attains the lowest value of the AIC, having therefore ΔAIC$_c$ equal to zero. Model 2 also has a low value of the AIC, which is essentially indistinguishable from that of model 3, especially given the small sample size of the data. On the other hand, all other models have $\Delta\text{AIC}_c > 10$, being therefore effectively unsupported by the present data. Models 2 and 3 also have lower ΔAIC$_c$ when compared with others, not shown in tables 3.2 and 3.3, according to which both TCR-transgenic populations are taken to be identical, in the sense of having the same values of values of R^2_α and τ_T, such that the only distinction considered is that between the polyclonal population and the TCR-transgenic populations (having $\Delta\text{AIC}_c = 9.4$, for only R^2_α as different, $\Delta\text{AIC}_c = 9.1$ for only τ_T, and $\Delta\text{AIC}_c = 14.4$, for both R^2_α and τ_T as different). This suggests that, in models 2 and 3, the addition of an extra parameter, to describe each TCR-transgenic population separately, is justified by the decrease in the sum of the squared

Model	Description	# Parameters Fitted
1	Both R^2_α and τ_T are the same in all biological populations	10
2	R^2_α may be different for each biological population, but τ_T is the same in all	12
3	R^2_α is the same, but τ_T may be different for each biological population	12
4	Both R^2_α and τ_T may be different for each biological population	14
5	Each experiment is fitted independently from the others	24

Table 3.2: Overview of the models tested, with a description of how parameters R^2_α and τ_T are set in the three biological populations, and the resulting number of parameters that were fit. As discussed in the text, parameter δ_0 was fit separately for each experiment

residuals compared with these other models. Therefore, the experimental data in figure 3.6 can be most parsimoniously explained by the three biological populations differing in R^2_α (model 2) or τ_T (model 3). In the following paragraphs, we analyze the relationship between these two models in more detail.

To compare models 2 and 3, figure 3.7 presents the values of $\Omega_{H,L}(t)$ estimated for each of the three populations, for time instants concordant with the duration of the experiments done. Also shown are the values of R^2_α for each model, to which $\Omega_{H,L}(t)$ converges as t increases. For simplicity, we will consider that convergence of $\Omega_{H,L}(t)$ to R^2_α has taken place for $t = 3\,\tau_T$, corresponding to the exponential term in equation 3.3 having decayed to 5% of its initial value. In the comparison shown in figure 3.7, $\Omega_{H,L}(t)$ for one model is indistinguishable from that of the other model in every one of the three populations. Therefore, the two models are equivalent in terms of the dynamics of function $\Omega_{H,L}(t)$ in the time span corresponding to the lifetime of the cells in these experiments. In the case of Marilyn, $\Omega_{H,L}(t)$ has decayed considerably by 72 hours, while the polyclonal population shows only a marginal decrease in $\Omega_{H,L}(t)$ even after 96 hours. While the point estimates for the OT-II population are in between the values of $\Omega_{H,L}(t)$ for Marilyn and the polyclonal population, the considerable experimental variation in the data on this population, and consequently in the confidence intervals for $\Omega_{H,L}(t)$ in figure 3.7B, precludes a de-

Fitting $\Delta_{H,L}(t)$							
Model	SS Residuals	Population	Exp. #	δ_0	R^2_α (%)	τ_T (h)	ΔAIC$_c$
1	0.075	Polyclonal	1	0.56	59	33	27.2
			2	0.63			
			3	0.64			
		Marilyn	1	0.24			
			2	0.29			
			3	0.27			
		OTII	1	0.32			
			2	0.27			
2	0.033	Polyclonal	1	0.52	70	37	2.2
			2	0.58			
			3	0.58			
		Marilyn	1	0.37	12		
			2	0.39			
			3	0.37			
		OTII	1	0.34	45		
			2	0.30			
3	0.031	Polyclonal	1	0.50	24	187	0.0
			2	0.56			
			3	0.57			
		Marilyn	1	0.37		23	
			2	0.41			
			3	0.38			
		OTII	1	0.34		69	
			2	0.29			
4	0.031	Polyclonal	1	0.51	59	75	11.7
			2	0.57			
			3	0.58			
		Marilyn	1	0.37	24	24	
			2	0.41			
			3	0.38			
		OTII	1	0.34	28	63	
			2	0.29			
5	0.019	Polyclonal	1	0.50	22	199	87.1
			2	0.57	67	46	
			3	0.57	0	249	
		Marilyn	1	0.40	4	32	
			2	0.39	40	16	
			3	0.37	25	27	
		OTII	1	0.31	0	172	
			2	0.34	30	30	

Table 3.3: Estimates for the parameters of the populations obtained by fitting the data on $\Delta_{H,L}(t)$, based on the different models being considered. The results are presented in terms of ΔAIC$_c$, the difference between the value of the AIC (corrected for small sample size; see methods) of each model and the minimum value of the AIC. Models with lower values of ΔAIC$_c$ provide a more parsimonious explanation for the data.

tailed quantitative analysis. In particular, the question of whether the dynamics of $\Omega_{H,L}(t)$ for the OT-II population more closely resemble that of Marilyn or the polyclonal population remains unclear.

The parameter estimates in table 3.3 and the data in figure 3.7 imply that these two models differ essentially only in how they extrapolate on the dynamics of the polyclonal and OT-II populations for time instants longer than the experiments done. This occurs because, given data on a relatively short window of observation, up to a time instant t_{max}, there are two extreme scenarios that may be considered: either $3\,\tau_T \approx t_{max}$, such that $\Omega_{H,L}(t)$ has converged to R^2_α by the end of the experiment, or $3\,\tau_T \gg t_{max}$, and it would take much longer than the duration of the experiment for $\Omega_{H,L}(t)$ to decay to R^2_α. The former scenario corresponds to model 2, and the latter to model 3. The data on Marilyn, which shows the most pronounced decay, further shapes this distinction, by constraining, to some degree, the maximum value of τ_T in model 2 and of R^2_α in model 3, so that the dynamics of the three populations can be fitted by these two models. In a related way, statistical variation in the estimates of $\Delta_{H,L}(t)$ will introduce a negative correlation in the values estimated for R^2_α and τ_T, such that lower values of R^2_α will tend to be paired with larger values of τ_T, while larger values of R^2_α will be paired with lower values of τ_T. Such a negative correlation is apparent, for example, in the estimates of R^2_α and τ_T of the individual experiments in model 5 (table 3.3). To the extreme, a special case of model 3 where R^2_α is constrained to be equal to zero further increases the estimated values of τ_T, thereby maximizing the extrapolation done, can also provide a parsimonious explanation for the data, in terms of the AIC.

To refer to the parameters estimated for each population, we will use the subscripts P, M and O to refer to the polyclonal, Marilyn and OT-II populations, respectively. Even though models 2 and 3 are equivalent in the time span considered, given that $\Omega_{H,L}(t)$ provides an upper bound on R^2_α (as discussed in section 2.5.2, page 48), the results on the Marilyn population in figure 3.7A necessarily imply that $R^2_{\alpha,M} \leq 40\%$, and therefore that the unstable component preponderates in this population. This is further confirmed by analyzing the 95% confidence intervals of R^2_α for this population according to both models, with 95% confidence interval (CI) of $R^2_{\alpha,M}$ from 0% to 35% in model 2, and from 0% to 39% in model 3. For the OT-II population, although the point estimates of R^2_α given by the two models differ to some degree, they altogether suggest that the unstable component also contributes considerably to the variation, with $R^2_{\alpha,O}$ around 45% in model 2, and $R^2_{\alpha,O}$ around 25% for model 3. As in the case of Marilyn, the possibility that the unstable component is the only contribution is also included. The CI for R^2_α given by model 2

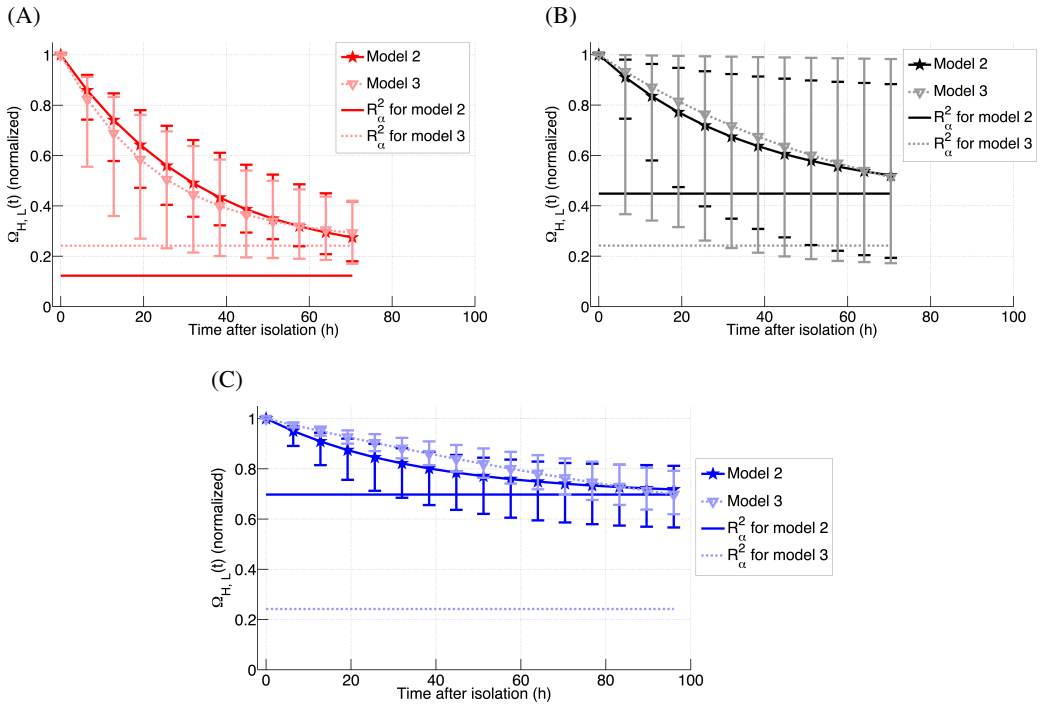

Figure 3.7: Comparison between the values of function $\Omega_{H,L}(t)$ derived from models 2 and 3, for Marilyn (A), OT-II (B) and the polyclonal population (C). The values of $\Omega_{H,L}(t)$ are shown only for time instants that are compatible with the duration of the experiments for each population. Error bars represent joint 95% confidence intervals (based on the Bonferroni correction) for the 3 biological populations in each model, as estimated by bootstrapping, while the horizontal lines are the values of R^2_α estimated by each models for each biological population. The original data are not included in the figures, to facilitate visualization, and also as each model implies a different normalization of the data (dependent on the values of δ_0).

includes essentially all possible values of R^2_α (ranging from 2% to 100%), the considerable uncertainty being due to the limitations of the data on this population, and that given by model 3 ranges from 0% to 39%. For the polyclonal population, on the other hand, the predictions of the two models are fundamentally different, with model 2 implying that the stable component is the main contribution in this population (CI for $R^2_{\alpha,P}$ from 51% to 81%), while model 3 implicates the unstable component as the main contribution (CI for $R^2_{\alpha,P}$ from 0% to 39%). Even though $\Omega_{H,L}(t)$ for the polyclonal population is centered around 70% after 96 hours for both models, it would take 460 hours more for $\Omega_{H,L}(t)$ to converge to $R^2_{\alpha,P}$ in model 3.

However, in terms of providing a good description for the data, given that this exper-

imental setup limits the viability of cells to 3–4 days, any value of τ_T corresponding to a longer time period is unrealistic in this setting. Therefore, model 2, by constraining τ_T to a value that is compatible with the lifetime of the cells under these conditions, provides a more appropriate description for the populations in this setting, resulting in a value of R^2_α that is close to that of function $\Omega_{H,L}(t)$ by the time each population has essentially disappeared. Even though the "true" characteristic time τ_T may be considerably longer, as estimated by model 3, it will never be observed under these conditions. For this reason, in order to provide a good description of these data, we rely on model 2, with function $\Omega_{H,L}(t)$ being shown, along with the original data, in figure 3.8, highlighting the values of R^2_α estimated for each population. Based on the description provided by this model, and in this simplified *in vitro* setup, all three populations may be characterized as being composed of a set of stable variants, with varying contributions to the total variation that is observed, ranging from $R^2_\alpha \approx 70\%$ in the polyclonal population, $R^2_\alpha \approx 45\%$ for OT-II, notwithstanding the uncertainty and the small sample size of the data, and $R^2_\alpha \approx 10\%$ for the Marilyn population. Moreover, in the case of Marilyn, given that it can be described by such a small value of R^2_α, at present it provides the best empirical approximation to the concept of a sub-population, introduced in chapter 2, as a population of cells in which the unstable component is the only contribution to variation in expression levels. These results are in agreement with the indirect analysis done in the previous section, which suggested that the stable component would be present in the polyclonal population. In the next section, we sought to extend this observation in a setup allowing for long-term analysis of high and low expressors.

3.5 Assessing the Stable Component in the Polyclonal Population under Long-term Conditions upon Adoptive Transfer

The setup in which cells were maintained *in vitro* without explicit stimulation, used in the previous section, indicated that all T cell populations could be described as a set of stable variants, consistent with different contributions of the stable component, as judged based on the values of R^2_α. However, that particular setup constrained measurements to a limited time window, as cells would eventually die off after 3–4 days. Therefore, to address whether the differences in expression levels indeed persist for longer periods of time, we sought to use an experimental condition in which the viability of the cells is maintained. While long-term cultures would perhaps be the closest analog of the setup used in section 3.4, they tend to require extensive optimization for periodic stimulation of the cells

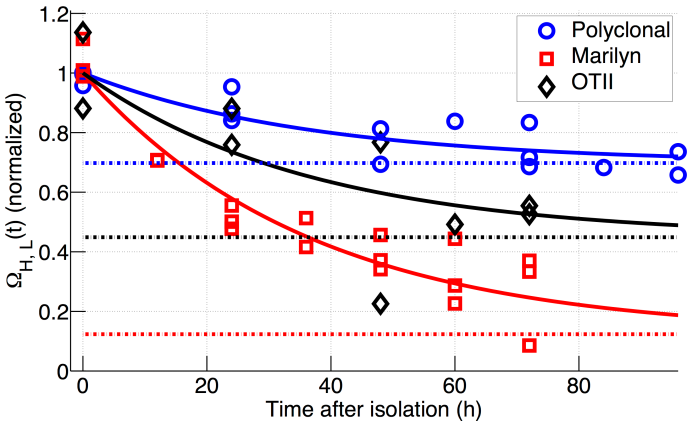

Figure 3.8: Results of fitting the experimental data (symbols) with model 2 (continuous lines), in terms of function $\Omega_{H,L}(t)$, highlighting the estimated values of R_α^2 for the Marilyn, OT-II and polyclonal populations (horizontal dotted lines). The data points corresponding to the experimental data are normalized by the value of δ_0 (equation 3.2) obtained for each curve.

(Levine et al., 1997). Therefore, this section instead presents a proof-of-concept for the analysis of the stable component under *in vivo* conditions. These conditions are based on adoptive transfers to lymphopenic recipients, which tend to induce extensive proliferation and activation of the T cells transferred (Kieper et al., 2005; Min et al., 2005; Surh and Sprent, 2008). We chose to use *Rag2*$^{-/-}$ mice as recipients, given their lack of endogenous B and T cells. This constitutes a more refined system, in the sense that the transferred cells should constitute a higher percentage of the cells recovered in the peripheral lymphoid organs for analysis. It also avoids having to stain a large number of total recovered cells in order to analyze a minimally reliable number of transferred cells, and is therefore expected to minimize experimental variation upon the staining. Due to the requirement of an increased number of cells for transfer (2×10^5 per recipient), and experimental limitations in the numbers of high and low expressors than can be isolated in a reasonable amount of time, we opted to use only the polyclonal population, sorting high and low expressors as 25–30% of the starting population. Moreover, the extensive proliferation induced under these conditions, along with the lack of competition with endogenous T cell populations in the recipient mice, should guarantee that a sufficient number of transferred cells can be analyzed several weeks after transfer.

As a consequence of extensive stimulation and activation of the transferred cells, and hence possible changes in the regulation of expression of the TCR, a simple qualitative

analysis is done, in terms of whether or not the stable component is present. This question is addressed by analyzing whether the function $\Delta_{H,L}(t)$ is equal to zero in different time-points after transfer (see section 2.5).

The transfer of T cells to a lymphopenic environment leads to considerable proliferation, due to the availability of abundant proliferation and survival factors, such as peptide-MHC ligands and cytokines like IL-7 (Surh and Sprent, 2008; Guimond et al., 2009). This provides for an estimate of the ability of transferred cells to reconstitute the peripheral pool in the recipient mice. Moreover, as the transfer of purified $CD25^{-}CD45RB^{high}$ $CD4^{+}$ T cells to *Rag2*$^{-/-}$ recipients has been used by some authors as a protocol to induce colitis (for example, Wang et al., 2008; Durant et al., 2010), we took advantage of this setup to also analyze the pathogenic potential of high and low expressors, as assessed by the weight loss induced in the recipients upon transfer.

Hence, high and low expressors, isolated from the polyclonal population, were transferred to separate *Rag2*$^{-/-}$ recipients via intravenous injection, with each animal receiving 2×10^5 cells of either isolated population. In the two experiments done, cells were analyzed in different time-points: in the first experiment, the time-points of analysis were 14, 28 and 110 days after transfer, while in the second, 48, 55 and 67 days after transfer. In the following, we show the results of this second experiment, which was conducted with a larger number of animals. The conclusions drawn here, in terms of the function $\Delta_{H,L}(t)$, and the comparison between the number of cells recovered and the weight of the mice are overall concordant between the two experiments. The results of the first experiment are included in appendix 3.C.

In different time-points after transfer, around 7, 8 and 10 weeks, some of the recipients were sacrificed, and cells analyzed to quantify TCR expression levels. In all cases, the estimates of the fold-ratio between the median fluorescence intensities (untransformed values) of high and low expressors remained lower than 2-fold. The value of function $\Delta_{H,L}(t)$ remained overall positive, in terms of 95% confidence intervals (CIs), in both lymph nodes (figure 3.9A) and spleen (figure 3.9B), showing that high and low expressors indeed remain different in all time-points analyzed. This conclusion was further validated by a two-way ANOVA. Equivalent results were obtained by staining for CD3ϵ, confirming that the differences detected between high and low expressors reflect, indeed, differences in TCR expression levels. Of note however, is the fact that in a single time-point of each of the two experiments done, the 95% CIs of function $\Delta_{H,L}(t)$ in the spleen, when staining for CD3ϵ (in figure 3.9B, 7 weeks after transfer), include the value zero. This most likely represents merely experimental variation, due to the limitations in the number of animals of

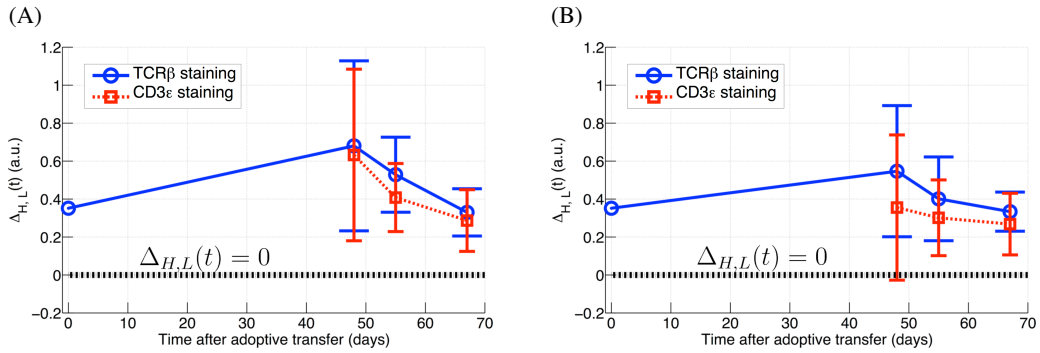

Figure 3.9: The stable component in a polyclonal population is robust to the highly stimulatory conditions provided by the lymphopenic mice upon adoptive transfer (second experiment). The function $\Delta_{H,L}(t)$ was estimated by staining cells for TCRβ or, alternatively, for CD3ϵ. Shown is data for lymph nodes (A) and spleen (B). Error bars denote 95% confidence intervals for each time-point, while the dotted black line highlights the threshold corresponding to $\Delta_{H,L}(t)$. No error bars are shown for the initial time-point ($t = 0$), as that estimate is based on a single replicate of high and low expressors, just before performing the adoptive transfer. Data correspond to the second of 2 independent experiments, with TCR levels quantified in different time-points, each having 3–5 animals per group.

each group used for analysis, since this was not observed in the other timepoints of each experiment. These data show that there is a stable component of variation in TCR expression levels, as the differences between high and low expressors are maintained ($\Delta_{H,L}(t) > 0$), even in these long-term conditions, in the strongly stimulatory environment of lymphopenic mice.

To examine the ability of high and low expressors to reconstitute the peripheral pool in the recipients, we quantified the number of cells in lymph nodes and spleen. The average number of cells recovered per animal ranged from 2×10^6 to 5×10^6 across the time-points analyzed, indicating extensive proliferation. In comparing high and low expressors, in face of the considerable experimental variation, we find that the number of cells recovered in lymph nodes (figure 3.10A) and spleen (figure 3.10B) are indistinguishable between animals that received high or low expressors, suggesting that these two populations have equivalent capacities to expand. Finally, in terms of the weight of recipients, we find that relatively few animals show noticeable loss of weight up to the last time-point tracked, and given the experimental variation observed, no difference between the groups of mice that received high or low expressors. Therefore, high and low expressors do not seem to have different pathogenic potential in this setup, as assessed based on the weight loss of the recipients.

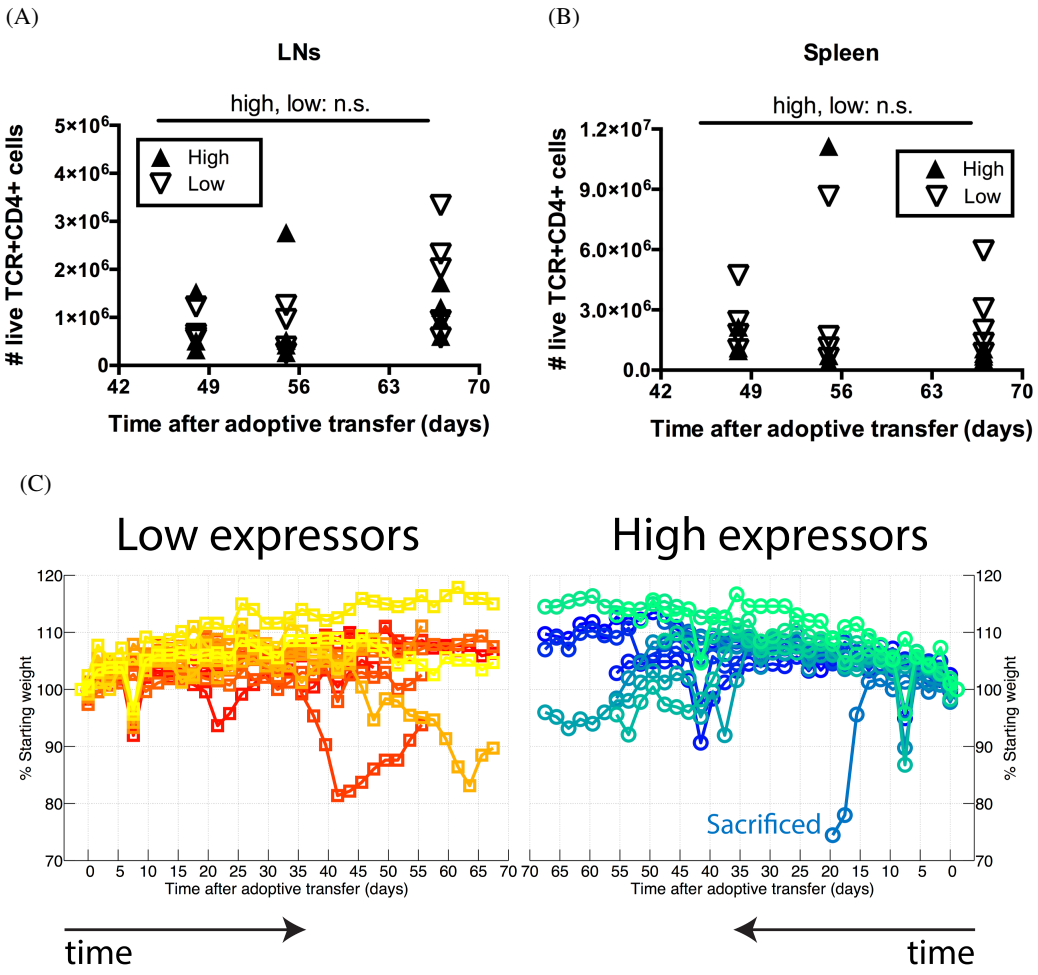

Figure 3.10: High and low expressors have indistinguishable abilities to reconstitute the peripheral pool (lymph nodes (A) and spleen (B); filled triangles pointing up denote animals that received high expressors, while open triangles pointing down denote those that received low expressors), and to induce weight loss (C) upon adoptive transfer to lymphopenic (*Rag2*$^{-/-}$) recipients (second experiment). Animals that had lost 20% or more of the initial weight were sacrificed, in accordance with standard operating procedures for animal welfare. Animals from the control group, injected with PBS alone, slowly gain weight, reaching, by the end of the experiment, $\sim$110% of the starting weight (not shown). Data correspond to the second of 2 independent experiments, with 12 animals per group at the start; n.s.: differences between the number of cells in animals receiving high or low expressors are not significant, based on a two-way ANOVA.

These results provide further support for the existence of a stable component in TCR expression, as the differences between high and low expressors persist even *in vivo*, on a time frame of several weeks, upon extensive proliferation. Furthermore, these data indicate that high and low expressors, at least under the conditions assessed here, have equivalent abilities to expand in such an *in vivo* environment and induce weight loss.

3.6 Discussion

In this chapter, we analyzed the components that shape variation in expression levels of the TCR. The theoretical framework put forward in chapter 2 defined the stable and unstable components, which lump mechanisms that result in, respectively, permanent and transient differences in the expression levels of two subsets of cells. If the stable component is present, the biological population can be described as a set of stable variants. These variants are termed sub-populations, with each sub-population consisting of a set of cells in which all variation in expression levels is due to the unstable component. The impact of the stable and unstable components on the expression levels in a population is formalized by decomposing the variance of expression levels, and quantified by parameter R^2_α, defined as the relative contribution of the stable component. This parameter ranges from 0% (only the unstable component is present) to 100% (only the stable component is present). Given that the unstable component results in stochastic fluctuations in the expression levels of each cell throughout time, an additional parameter of interest is the characteristic time of the variation, τ_T, related to the expected time for the transient differences between the expression levels of two subsets of cells to disappear. By focusing on various T cell populations, one being polyclonal (genetically diverse) and two being isogenic (TCR-transgenic on a *Rag2*$^{-/-}$ background), we quantified the relative contribution of the stable and unstable components (R^2_α) and the characteristic time of the variation (τ_T) in expression levels of the TCR. In the polyclonal population, both genetic and epigenetic variation have the potential to mold the stable component, while in the isogenic populations, if the stable component is present, it is due, by definition, to epigenetic variation. The denomination of epigenetic variation is used in a general sense, to refer to mechanisms that result in phenotypic differences that are not explained by differences in DNA sequence, but that persist throughout time. Therefore, by evaluating the relative contribution of the two components in these various populations, one may assess the impact of genetic and epigenetic variation on the stable component in the polyclonal population.

In the polyclonal population, we always considered non-activated cells, referred to as

naive cells, which constitute a phenotypically more homogeneous population, so as to avoid the potentially confounding effects of, for example, regulatory T cells (Sakaguchi et al., 2008). The two isogenic T cell populations, used as approximations of the clones in a polyclonal population, were obtained from two mouse strains, Marilyn (Lantz et al., 2000) and OT-II (Barnden et al., 1998). In these mouse strains, the somatic rearrangement is ablated (*Rag2*$^{-/-}$ background), and a functional TCR is obtained by transgenes driving the expression of rearranged α and β chains. These two cell populations, often referred to as TCR-transgenic T cell populations, are described as being composed essentially only of naive cells (Lantz et al., 2000; Moon et al., 2007). In section 3.3, consistent with data from a previous report for Marilyn (Grandjean et al., 2003) and also other TCR-transgenic populations (Kassiotis et al., 2003), a comparison between TCR levels in the polyclonal, Marilyn and OT-II populations revealed that these two TCR-transgenic populations have lower total variation (σ^2_T) than the polyclonal. The analysis, done in section 3.3, based on the values of σ^2_T, suggested that the stable component would be present in the polyclonal population, whose value of R^2_α was indirectly estimated as between 50% and 60%. This result is based on the assumptions that i) genetic variation is the only mechanism influencing the stable component in the polyclonal population, ii) that the TCR-transgenic populations are representative, in terms of the statistics of TCR expression levels, of the clones in the polyclonal population, and iii) that measurement noise is negligible. In particular, these assumptions imply that $R^2_\alpha = 0$ for the TCR-transgenic populations and for the clones in the polyclonal population. If, on the other hand, the assumption of genetic variation as the only mechanism impacting on the stable component is relaxed, admitting that epigenetic variation may also contribute to this component, one would obtain an increased value of R^2_α for the polyclonal and positive values for the TCR-transgenic populations.

To directly estimate the value of R^2_α for the polyclonal and TCR-transgenic populations, along with τ_T, we then relied in section 3.4 on the isolation of high and low expressors, using cell sorting. In this setup, the sorted cells were maintained *in vitro* for up to 4 days, in the absence of contact with non-T cells or stimulation, a well-defined condition under which all cells eventually die off, and no cell division is expected (Deenick et al., 2003). We asked whether the ensemble of the data, composed of multiple experiments with each of the populations, could be most parsimoniously described by the populations having equal or different values of R^2_α and τ_T. In this analysis, we relied on the Akaike Information Criterion (AIC; Burnham and Anderson, 1998) to compare models with different relationships between R^2_α and τ_T in the three T cell populations. We found that the most parsimonious explanation for the data would be that the three populations differ only in the values of R^2_α,

following model 2, or only in the values of τ_T, according to model 3. These two models are indistinguishable in terms of the function $\Omega_{H,L}(t)$, differing only in the way in which they extrapolate on the dynamics for longer time frames. Despite being indistinguishable in this initial time window, given that later time instants cannot be observed for these populations under these conditions, model 2 provides the most adequate description of these data. By constraining the characteristic time τ_T to a value that is compatible with the lifetime of the cells under these conditions, this model provides a more appropriate description for the populations in this setting, with R^2_α being close to $\Omega_{H,L}(t)$ by the time each population has essentially died out. In this particular description, all three T cell populations are seen as a set of stable variants, with various values of the relative contributions of the stable component. This model describes the polyclonal population as having the greatest value of R^2_α among the populations tested, equal to 70%, with the stable component as the main contribution to the variation observed, while the unstable component would be the main contribution in the Marilyn population (R^2_α equal to 10%) and in OT-II (R^2_α equal to 45%). In this analysis, the OT-II population, with R^2_α equal to 45%, appears as an intermediate scenario in between the polyclonal and Marilyn populations. This provides initial evidence that TCR-transgenic populations may be better described by distinct sets of values of R^2_α and τ_T, since models considering Marilyn and OT-II with identical values of R^2_α and τ_T had greater values of the AIC. However, given the considerable experimental variation in the data on OT-II, the question of whether it more closely resembles Marilyn or the polyclonal population cannot be answered at present. Furthermore, it should be emphasized that the estimates obtained are specific for this *in vitro* setup, in which we focus on cell-intrinsic components only, since there are no other cells present that could provide external signals throughout the experiment. Consequently, it is possible that other signals, such as those arising from the intermittent contact with antigen-presenting cells in the *in vivo* environment result in different values of R^2_α for these populations.

In the indirect analysis of section 3.3 and the direct quantification done in section 3.4, we obtained independent estimates of R^2_α for the polyclonal population above 50%, altogether providing considerable evidence that the stable component contributes to the variation in TCR levels in this population. Moreover, given that the point estimate for R^2_α of 70% from section 3.4 is marginally greater than the range of 50% to 60% indirectly estimated in section 3.3, we find an initial suggestion that epigenetic variation may have an impact on the stable component. Concordant with this notion, the point estimates of R^2_α for the two TCR-transgenic populations considered here were also greater than zero. However, the data on the effect of such epigenetic variation is, at best, suggestive, being

based on the point estimates, which represent the most likely values given the present data. When it comes to the polyclonal population, the comparison between the indirect estimate of section 3.3, considering genetic variation only, and the point estimate of section 3.4, indicates that, if indeed present, these mechanisms would make a small contribution, such that genetic variation is likely to be the main contribution to the stable component. In fact, the confidence interval for R_α^2 in this population (from 50% to 80%) includes the range of 50%–60% estimated in section 3.3. Furthermore, the confidence intervals for Marilyn and OT-II also envisage the case of $R_\alpha^2 = 0$, and hence that the unstable component would be the only contribution to the variation that is observed. However, such epigenetic variation, if indeed present, may be a feature only of TCR-transgenic populations, as their relationship to the actual clones in a polyclonal T cell population is not known. However, since this particular description of the data is conditioned on the limited time span for analysis of the populations, this possibility remains as mere speculation at present. The differences in the values of R_α^2 estimated for the Marilyn and OT-II populations suggest that the underlying epigenetic mechanism may differ among isogenic populations of T cells, such as the different clones in a wild-type animal, or the populations from different TCR-transgenic (*Rag2*$^{-/-}$) mouse strains. While a more detailed investigation of the nature of this epigenetic variation could start by a detailed comparison between the Marilyn and OT-II populations, we note that the considerable experimental variation of the data on the latter population suggests caution in this direction. It is clear, therefore, that further studies are needed in order to substantiate the evidence of the impact of such epigenetic variation.

Besides R_α^2, the particular description of the population afforded by model 2 in section 3.4 also provided an estimate of the characteristic time of the variation, τ_T, equal to 37 hours. This parameter is related to the expected time for differences in the expression levels of two subsets of cells to decrease, and describes the dynamics of the unstable component, representing a form of transient memory of expression levels (Sigal et al., 2006). Under the model of protein expression in chapter 2, since $\sigma_T < 0.3$ for the populations considered, τ_T is relatively well approximated by the sum of the expected time of changes in the rate of protein production (parameter τ) and the average protein lifetime (β), the latter being related to the half-life of the protein. Early studies relying on metabolic labeling estimated the half-life of the TCR in hybridomas as between 10 and 20 hours (Klausner et al., 1990), corresponding to a value of β between 14 and 28 hours. In primary CD4$^+$ T cells, it has been reported (Liu et al., 2000) that the TCR is very stable, as treatment with protein synthesis inhibitor for up to 12 hours led only to modest changes in expression levels, suggesting that the half-life is greater than 12 hours, and hence that β is greater than 17

hours. However, this estimate is potentially problematic, since it has been reported that this treatment results in up-regulation of ζ chain mRNA levels (Bronstein-Sitton et al., 1999), such that the regulation of the TCR could be altered under these conditions. In human T cell clones, Sousa and Carneiro (2000) estimated the baseline turnover of the TCR by fitting the dynamics of the mean TCR levels upon short-term stimulation, and found a value for β of 15 hours. All three values (Klausner et al., 1990; Liu et al., 2000; Sousa and Carneiro, 2000) are compatible with the value of τ_T estimated here, and suggest that β has the same order of magnitude as τ_T according to this particular description of the data. Therefore, in the context of the analysis done in chapter 2, these results indicate that regulation of expression of the TCR is characterized by $\tau/\beta > 1$, and hence that the considerable stability of the protein is a relevant factor in the dynamics of the fluctuations in expression levels.

In studies quantifying the dynamics of the percentage of T cells that express cytokines when stimulated (Paixão et al., 2007; Mariani et al., 2010), the characteristic time τ_T is estimated as around 70 hours for the cytokines IL-10 (Paixão et al., 2007) and IL-4 (Mariani et al., 2010), under conditions in which the cells divide. In terms of molecular mechanism, this property has been linked to slow dynamics of chromatin remodeling (Paixão et al., 2007; Mariani et al., 2010), the most likely explanation for the increased value compared with the one estimated for the TCR using model 2 with no cell division expected.

Interestingly, Bonnet et al. (2009) compared the dynamics of expression levels of the TCR in two clonal populations of hybridomas. One population was generated out of wild-type (WT) $CD4^+$ cells, and the other out of $CD4^+$ cells expressing a reduced version of the Eβ enhancer (Eβ^{169} hybridoma), hereafter referred to as simply WT and mutant hybridomas, respectively. The data on these two populations, shown in detail in appendix 3.D, correspond to high and low expressors, analyzed after 2 and 5 weeks in culture. Being a transformed cell line, these cells undergo extensive cell division in the period (M. Bonnet, personal communication). Although the limited amount of data available precludes a quantitative analysis, it takes around 5 weeks for $\Delta_{H,L}(t)$ to change considerably in the WT hybridoma. Intriguingly, in the mutant hybridoma, $\Delta_{H,L}(t)$ changes only marginally throughout the timespan of analysis, pointing to an impact of the Eβ enhancer in shaping the origin and/or the timescale of the variation in the hybridomas. Moreover, based on the values of $\Delta_{H,L}(t)$ for the WT hybridoma, the characteristic time of the variation in these cells would be around 300–400 hours, which is markedly greater than the value estimated for the primary cells in section 3.4. In the case of the mutant hybridoma, the data is consistent with an even longer characteristic time. The total variation in both hybridomas

(estimated as $\sigma_T^2 \approx 0.7$) is much larger than in the primary cells (sections 3.3 and 3.4), in which σ_T^2 ranges from 0.03 to 0.07. That hybridomas have greater total variation (σ_T^2) is consistent with data from a study (Mandl et al., 2013) in which hybridomas were derived from two different TCR-transgenic populations (on a Rag-deficient background). This correlation between the total variation and the characteristic time in the primary cells analyzed in this study and the hybridoma in Bonnet et al. (2009) is in line with the experimental observations of Sigal et al. (2006) on a correlation between the coefficient of variation and the auto-correlation time of the expression levels of various molecules in an isogenic mammalian cell line. While it could be interesting to quantify R_α^2 and τ_T in primary Eβ^{169} CD4$^+$ T cells, the severely reduced number of mature T cells in these animals (around 7-fold less CD4$^+$ T cells in lymph nodes, as compared with WT mice; Bonnet et al., 2009) essentially precludes such analysis.

Although hybridomas may not be comparable with *ex vivo* T cell populations, the results of Bonnet et al. (2009) may provide a possible basis for different T cell populations to have different characteristic times τ_T, in part due to the dependence of Marilyn and OT-II on transgenic expression of the α and β chains. In light of this evidence, and considering that the analysis of the *ex vivo* T cells (section 3.4) was limited to up to 72–96 hours after the isolation, it remains possible that the populations differ in the values of τ_T, and that the "true" value of R_α^2 is equal to zero for some or all of these populations, as predicted by model 3. Indeed, the different statistics of TCR expression levels in comparing the TCR-transgenic and the polyclonal, reported in section 3.3, and previously interpreted to suggest that the polyclonal population would be described by R_α^2 between 50% and 60%, could be alternatively taken as evidence of disturbed expression of the TCR due to the dependence on the transgenes. In this case, the different values of τ_T could be a reflection of particular features of the transgenes in each isogenic population, such as the integration site, number of copies, specific regulatory elements used, among others, while the polyclonal population would have a characteristic time reflecting the regulation based on the endogenous TCRα and TCRβ loci. However, we consider this possibility to be unlikely, since, as reviewed in section 3.2, it is thought that the ζ chain is the rate-limiting sub-unit for assembly of the TCR (Baniyash, 2004). In this case, the only requirement for effective regulation of the TCR to recapitulate that of cells from the polyclonal population is that expression of the α and β chains is sufficiently large, such that the ζ chain would remain as rate limiting. Interestingly, we verified that both TCR-transgenic populations express higher median TCR levels when compared with the polyclonal population, suggesting that expression of the rearranged α and β chains is not limiting in these cells. However, since it is not known

whether the ζ chain is indeed limiting for expression of the TCR in every T cell clone, it remains possible that expression of the TCR is disturbed in the TCR-transgenic cells.

Finally, to address whether the stable component indeed contributes to the variation that is observed in TCR levels, by analyzing cells under longer time-periods, we relied on the adoptive transfer of cells to *Rag2*$^{-/-}$ recipients, which do not have lymphocytes. For technical reasons, we focused only on the polyclonal population, due to the limiting number of Marilyn and OT-II cells that can be isolated. Under these conditions of extensive proliferation, the values of $\Delta_{H,L}(t)$ remained greater than zero for all timepoints analyzed, up to around 10 weeks. This demonstrates that at least a portion of the difference in expression levels between high and low expressors is maintained, confirming that there is, indeed, a stable component in TCR expression. However, in this setting, the likely changes in TCR expression levels due to the activation of cells constitutes a scenario for inference of R^2_α that deviates from the setting considered in chapter 2. This is consistent with a trend of the values of function $\Delta_{H,L}(t)$ in these data being greater than the initial value $\Delta_{H,L}(0)$ (figure 3.9). Therefore, much like the description of the data in the *in vitro* (section 3.4) is, at least in principle, specific to that condition, the detectable contribution of the stable component *in vivo* may be consequence of the activation of the cells in this particular setup.

This adoptive transfer setup has distinguishing features, one of them being that the recipients are immunodeficient, lacking functional B and T cells. This has been described as a chronically lymphopenic environment, in contrast to that of lymphocyte-replete mice rendered transiently lymphopenic upon irradiation (Surh and Sprent, 2008). There is evidence (Kieper et al., 2005; Min et al., 2005) that adoptive transfer of naive polyclonal T cells to immunodeficient mice induces two types of proliferative responses, one of them being relatively slow, referred to as homeostatic proliferation, and another massive, very rapid. Besides the different kinetics, the two responses depend differently on IL-7, the former (slow) being essentially abolished in the absence of IL-7. On the other hand, the rapid proliferative response is reduced, but not completely blocked, upon transfer to hosts that lack IL-7 or are germ-free (Kieper et al., 2005). Moreover, this response has been reported recently to be influenced by the stimulation of dendritic cells to produce IL-6 (Feng et al., 2010), and hence not only due to antigen-dependent T cell activation. The progeny of cells undergoing the rapid proliferative response constitute the majority of the population up to 10 days after transfer, the typical time frame analyzed in these studies (Kieper et al., 2005; Min et al., 2005; Feng et al., 2010), but this is not known on a longer time frame. Therefore, it is possible that the cells we recovered several weeks after transfer are mainly the progeny of clones that have undergone or are still undergoing this rapid proliferation.

Further studies would be necessary to address this possibility.

We also used this *in vivo* setup as a first step towards addressing potential functional correlates of the differences in TCR expression levels, by comparing the ability of high and low expressors to reconstitute the peripheral pool and the weight loss in the mice receiving these cells. In particular, the latter was analyzed given that the transfer of polyclonal naive $CD4^+$ T cells ($CD25^-CD45RB^{high}$) to *Rag2*$^{-/-}$ mice has been used, under some conditions, as a protocol to induce colitis (for example, Wang et al., 2008; Durant et al., 2010). In terms of the reconstitution of the peripheral pool, as assessed by the number of cells recovered in the lymph nodes and spleen, we obtained altogether equivalent results for high and low expressors in the time-points analyzed. This may reflect particular features of this *in vivo* setup, as it is possible that TCR levels are not limiting, such that all cells receive essentially the same TCR-derived signals, and the ability to reconstitute the peripheral pool in the recipient mice is mainly determined by other factors. In the analysis of the weight loss, we also did not find a distinction between the mice that received high or low expressors, despite suggestions of an initial experiment, done with a small number of animals, that some of those receiving high expressors would show pronounced loss of weight around 40 days after transfer. These initial suggestions were, however, not reproduced in the subsequent experiment. It has been shown, however, that the induction of colitis in this setup can be exacerbated by the colonization of particular commensal bacterial strains (Wang et al., 2008), which, being likely absent under our conditions, could explain the relatively low frequency of animals that lost a considerable amount of body mass. Therefore, the functional correlates and consequences of the variation in TCR expression levels remain unclear at present, with further studies needed to address these questions.

Given the demonstration that the stable component indeed contributes to variation in expression levels of the TCR, at least in the polyclonal population, we believe that the most acute continuation would be the comparison of the values of R^2_α and τ_T estimated from T cell populations maintained under long-term *in vitro* conditions. Under these conditions, even cells from TCR-transgenic animals can be propagated for long periods of time, upon periodic stimulation with mitogenic stimuli such as immobilized antibodies that crosslink the TCR. These conditions also allow for the analysis of isogenic populations, grown out of clones from the polyclonal population upon several cycles of stimulation (see, for example, Carneiro et al., 2009), circumventing the limitations of the TCR-transgenic populations. This setup provides a particularly appealing system to study the impact of differentiation in terms of variation in expression levels. However, this tends to require extensive optimization for determining the optimum periodic stimulation of the cells (Levine et al., 1997),

so as to avoid excessive cell death due to chronic stimulation (Jelley-Gibbs et al., 2005; Schrum et al., 2005).

In summary, in this chapter, we provide evidence for the stable component in the polyclonal population in an *in vitro* setup without stimulation of the cells. In this same setup, we also obtained initial suggestions for this component in the TCR-transgenic ($Rag2^{-/-}$) populations analyzed, indicating that epigenetic variation may impinge on the stable component. Moreover, in this setup we estimated the characteristic time of the variation *in vitro* as equal to 37 hours. By analyzing cells from the polyclonal population upon adoptive transfer to $Rag2^{-/-}$ mice, we have found that differences in TCR expression levels can indeed persist for several weeks, consistent with the idea of a stable component of variation. Therefore, these results establish the TCR in a polyclonal population of $CD4^+$ T cells as a model system to study how the stable and unstable components contribute to variation in expression levels. However, further studies are needed to validate the initial data supporting the stable component in the TCR-transgenic populations and to address the potential functional correlates and consequences of the differences in TCR expression levels.

Acknowledgements

We thank Rui Gardner, Telma Lopes and Cláudia Bispo for cell sorting and assistance on flow cytometric analysis; Manuel Rebelo for animal house management; Rosa Santos and Ana Regalado for antibody production. We thank Alexandre Varela and Luis Graça for OT-II.*Rag2*$^{-/-}$ animals for breedings, and Marie-Louise Bergman for some of the Marilyn.*Rag2*$^{-/-}$ animals used for initial breedings. We greatly acknowledge the support of Jocelyne Demengeot and Henrique Teotónio during the development of this work. We also thank Jocelyne Demengeot, Marie-Louise Bergman, Íris Caramalho, Andreia Lino, Ana Catarina Martins, Ricardo Paiva for discussions and suggestions on various aspects of the experimental setup. We thank Dr. Olivier Lantz (Institut Curie, France) and Dr. Francis Carbone (University of Melbourne, Australia) for providing additional information on the TCR-transgenic strains used in this work. This work was supported by a grant from the Fundação para a Ciência e Tecnologia (FCT) (grant number PTDC/BIA-BCM/108020/2008). TSG was supported by a fellowship from FCT (fellowship number SFRH/BD/33572/2008).

This chapter is an extended version of the analysis of experimental data presented in the following manuscript:

Guzella, T. S., Barreto, V. B., and Carneiro J. (2013). Quantifying the Contributions and Dynamics Underlying Variation in Expression Levels in a Cell Population. *In preparation, under final review by the co-authors*

Chapter 3

Materials and Methods

Mice

C57BL6/J (B6) mice were obtained from the Jackson Laboratory. Marilyn TCR-transgenic mice (Lantz et al., 2000) were kindly provided by Olivier Lantz (Institut Curie, France), and bred with B6.*Rag2*$^{-/-}$ (Taconic) to produce Marilyn.*Rag2*$^{-/-}$. OT-II.*Rag2*$^{-/-}$ (Barnden et al., 1998) animals (originally OT-II TCR-transgenic (Jackson) and *Rag2*$^{-/-}$ (CDTA)) were further bred with B6.*Rag2*$^{-/-}$ (Taconic) to produce OT-II.*Rag2*$^{-/-}$ animals. Mice were bred and maintained under specific pathogen-free conditions at the animal house of the Instituto Gulbenkian de Ciência, and used for experiments with ages between 8 and 16 weeks. Genotyping of animals used for initial breedings was performed on mouse tail DNA by PCR. Marilyn.*Rag2*$^{-/-}$ were bred and used for experiments as homozygous for the transgenes (*Marilyn*$^{+/+}$), while OT-II.*Rag2*$^{-/-}$ were used for experiments as heterozygous for the transgenes (*OTII*$^{+/-}$). In the case of Marilyn.*Rag2*$^{-/-}$, only females were used for experiments (Lantz et al., 2000). All animal procedures were approved by the ethics committee of the Instituto Gulbenkian de Ciência.

Antibodies and Flow Cytometry

Flow cytometry was performed using a Beckman-Coulter CyAN ADP. Fc receptors were always blocked prior to staining, by incubation with FcBlock (2.4G2, produced in-house). Cells were stained at 4°C, in ice-cold buffer with 1x PBS, 5% Fetal Bovine Serum (PAA), and, except in the case of sorting, with 0.1% Sodium Azide.

The following monoclonal antibodies produced in-house were used: anti-TCR-Cβ (H57-597), anti-CD4 (GK1.5), anti-CD8 (YTS169.4), anti-CD25 (PC61), anti-CD45RB (16A), anti-CD62L (MEL-14), anti-B220 (RA3-6B2), anti-MHC-II (M5/114), anti-Mac1 M1/70), anti-CD3ϵ (2C11). Commercial antibodies used were: anti-CD49b (pan-NK, DX5, BD), anti-CD4 (RM4-5, BD), anti-TCR$\gamma\delta$ (GL3; BD). Biotinylated antibodies were further labeled with PE-Streptavidin (BD).

Cell sorting and *in vitro* cultures

Single-cell suspensions were prepared from lymph nodes, and also spleens in the case of TCR-transgenic *Rag2*$^{-/-}$ animals (due to limited number of cells), by passing cells through a nylon mesh. Cell sorting was done using a FACSAria (BD), using a strategy based on negative selection of CD4^{+} T cells. Briefly, cells were stained with lineage markers

not expressed by naive $CD4^+$ T cells, and then $Lineage^-$ cells were sorted. A polyclonal naive population was sorted as $CD45RB^{high}$, $Lineage^-$ (CD8, pan-NK, B220, TCR$\gamma\delta$ and CD25) cells, while TCR-transgenic *Rag2*$^{-/-}$ cells were sorted as $CD62L^+$, $Lineage^-$ (B220, CD11c, pan-NK, Mac1, MHC-II). The use of CD62L as an alternative marker of naive cells allows for a more efficient sorting (due to a slow loss in the CD45RB signal throughout the sorting), given the limited number of cells, based on the fact that it has been described that the vast majority of these TCR-transgenic *Rag2*$^{-/-}$ cells have a naive phenotype (Lantz et al., 2000; Moon et al., 2007). Before each sorting for TCR-transgenic *Rag2*$^{-/-}$ cells, the gating for $CD62L^+Lineage^-$ cells, when analyzed in a control sample also labeled for CD4, includes more than 80% of TCR^+CD4^+ cells. For the *in vitro* analysis of T cell populations, high and low expressors, were sorted as 10% of the starting population. Purities of the sorted populations were assessed by staining aliquots of the sorted populations for CD4 expression, were typically greater than 96%. After sorting, T cells were cultured in flat-bottom 96-well plates (50×10^3 cells per well), in RPMI (Invitrogen), 10% Fetal Bovine Serum (PAA), 1% Sodium Pyruvate (Gibco), Penicillin/Streptomycin (Gibco), Gentamycin (Sigma), 0.1% 2-ME (Gibco), in an incubator at 37°C, with 5% CO_2.

TCR levels were quantified by staining, with anti-TCRCβ antibody, which binds to the constant region of one of the sub-units of the TCR (see, for example, Ghendler et al., 1998). Cells just after sorting and those that were maintained in culture were analyzed by re-staining the TCR, under optimal saturating conditions, using the same antibody anti-TCRCβ (clone and fluorochrome) as used for the sorting. In each time-point, 3 replicates (wells) of each sorted population were analyzed. In each experiment, an additional population (control) was sorted, using the same gates used for "all expressors", but without staining for the TCR, as a control for the impact of this staining.

Data was analyzed using FlowJo 8.8.7 (Tree Star Inc.). Cells were analyzed by gating on forward-scatter and side-scatter, live cells (propidium iodide negative) and $CD4^+$ cells. For the analysis of TCR levels, cells were further gated on $CD62L^+$ cells, to reduce experimental variation in TCR levels, especially in later time-points. Percentages of $CD62L^-$ cells were always lower than 20% in early time-points (up to 48 hours), and similar to those from control cells, arguing against an impact of staining for the TCR in order to sort cells. In the last time-point of each experiment, typically more than 100 events from each replicate of each sorted population were used for the quantification of TCR expression levels. In each time-point, TCR levels of the control population were compared against those of "all expressors", to confirm that the staining for the sorting does not induce changes in TCR expression levels (detailed in appendix 3.B). Gated data was exported as text files and

analyzed in MATLAB (Mathworks) using custom code.

Statistical analysis

Statistics of TCR expression levels (median of untransformed values, and variance of log-transformed values) were compared between the two TCR-transgenics Marilyn and OT-II, and the polyclonal naive population, using GraphPad Prism 6.0 (GraphPad Software, La Jolla, CA, USA). One-way ANOVA with Dunnett's test was used for the multiple comparisons between (Marilyn, polyclonal) and (OT-II, polyclonal), with a significance level of 5%.

Fitting the function $\Delta_{H,L}(t)$ and model selection

Numerical analysis was conducted using MATLAB. The exponential model was fit to the data by non-linear least squares. Fitting the ensemble of the experimental data was done by equally weighting each experiment, based on the number of data points per experiment and the number of experiments for each population. Values of the Akaike Information Criterion (AIC) were corrected for small sample size, as highlighted in section 2.4 of Burnham and Anderson (1998), and includes the residual variance as an additional effective parameter being estimated for each model. Confidence intervals (95%) for parameter values were obtained by bootstrapping each experiment separately, then fitting the ensemble of the data (10000 replicates).

Adoptive cell transfers

High and low expressors were sorted as for *in vitro* analysis, but as around 30% of the starting population, to increase the number of cells obtained. After sorting, cells were thoroughly washed in 1x PBS, counted (always using Beckman-Coulter Flow-Count beads) and resuspended in 1x PBS. B6.*Rag2*$^{-/-}$ recipients received 2×10^5 cells each via retro-orbital injection. Recipients were sacrificed in various time-points after transfer, with lymph nodes (pooled brachial, inguinal, axillary and mesenteric) and spleens harvested, cells counted and stained to quantify TCR levels. Confidence intervals (95%) for the function $\Delta_{H,L}(t)$ were determined using the t-Student approximation (Sokal and Rohlf, 1981, section 9.4), assuming that the two random variables have equal variances. Data were further analyzed with a two-way ANOVA, with the Holm-Sidak multiple comparison test with a group-wise (high and low expressors) significance level of 5%, using GraphPad Prism 6.0.

Extracting data on the hybridoma populations

Histograms of high and low expressors isolated from the clonal WT and E$\beta^{169/169}$ hybridoma populations were extracted from figure S6B of Bonnet et al. (2009). The populations just after isolation were inferred based on the gates shown in the original data, assuming that cell sorting is ideal. The data were then quantified in terms of the function $\Delta_{H,L}(t)$ for the WT and E$\beta^{169/169}$ hybridomas, based on a logarithmic scale (with base $\exp(1)$, rather than 10).

Appendices

Appendix 3.A Supplementary data for the analysis of TCR-transgenic strains

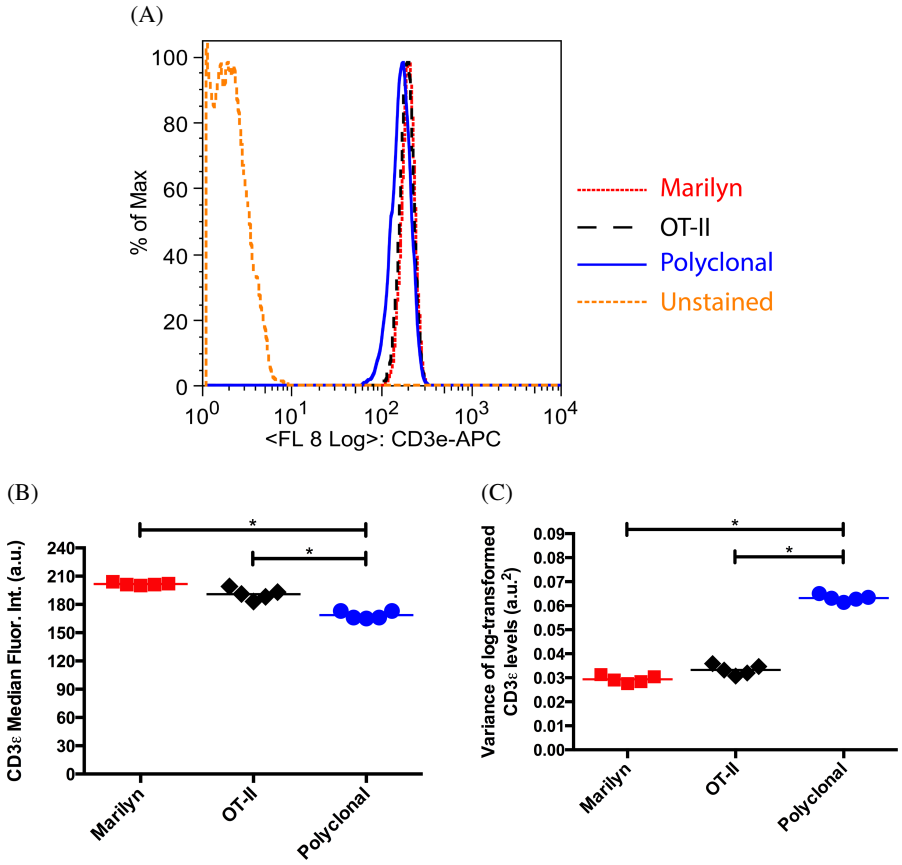

Figure 3.11: Comparison of TCR expression levels between the TCR-transgenic *Rag2*$^{-/-}$ populations Marilyn and OT-II, along with a naive polyclonal population, based on staining for CD3ϵ. For consistency with the setup in section 3.4, analysis of all populations was done based on the same gating strategy used for cell sorting (see methods), followed by gating for CD3ϵ^{+}CD4^{+} events. **(A)** Illustrative histograms of TCR levels of the populations. **(B)** Quantification of the median TCR levels, considering untransformed values. **(C)** Quantification of the variance of log-transformed values (σ^2_T). Data correspond to the first of 2 independent experiments, with 5 mice per group. $* \; p < 0.05$.

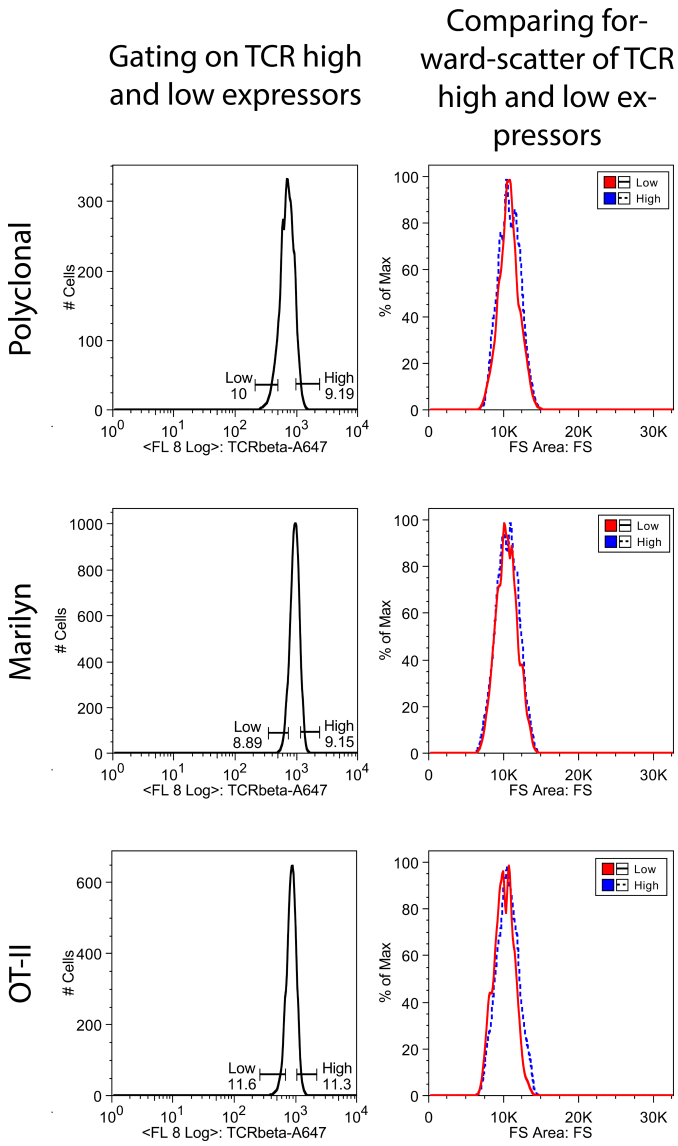

Figure 3.12: Comparison of the relationship between TCR expression levels and forward-scatter on the TCR-transgenic *Rag2*$^{-/-}$ populations Marilyn and OT-II, along with a naive polyclonal population, based on staining for TCRβ. Cells from each population, as in figure 3.5, were gated as around 10% into high and low expressors based on TCR levels, and compared in terms of the forward-scatter (often used in flow cytometry as an initial approximation of cell size). For consistency with the setup in section 3.4, analysis of all populations was done based on the same gating strategy used for cell sorting (see methods), followed by gating for TCRβ^{+}CD4^{+} events. Data correspond to the first of 2 independent experiments.

3.A.1 Data from the second experiment

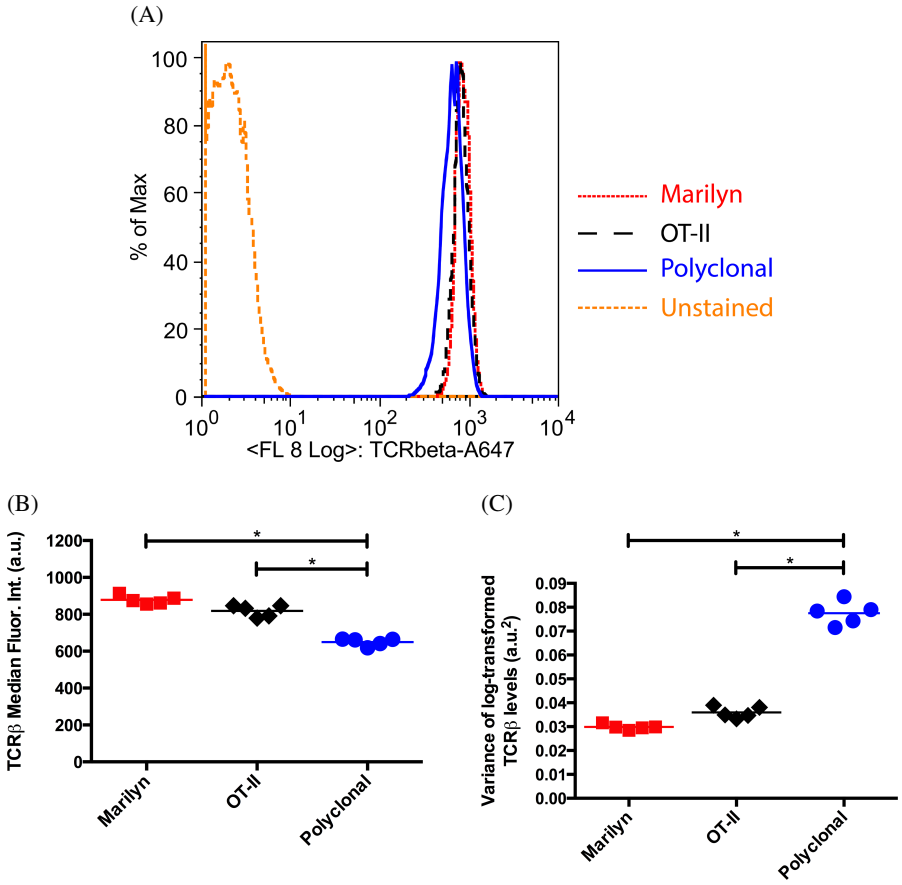

Figure 3.13: Comparison of TCR expression levels between the TCR-transgenic *Rag2*$^{-/-}$ populations Marilyn and OT-II, along with a naive polyclonal population, based on staining for TCRβ (second experiment). For consistency with the setup in section 3.4, analysis of all populations was done based on the same gating strategy used for cell sorting (see methods), followed by gating for TCRβ^{+}CD4^{+} events. **(A)** Illustrative histograms of TCR levels of the populations. **(B)** Quantification of the median TCR levels, considering untransformed values. **(C)** Quantification of the variance of log-transformed values (σ_T^2). Data correspond to the second of 2 independent experiments with 5 mice per group. $* \; p < 0.05$.

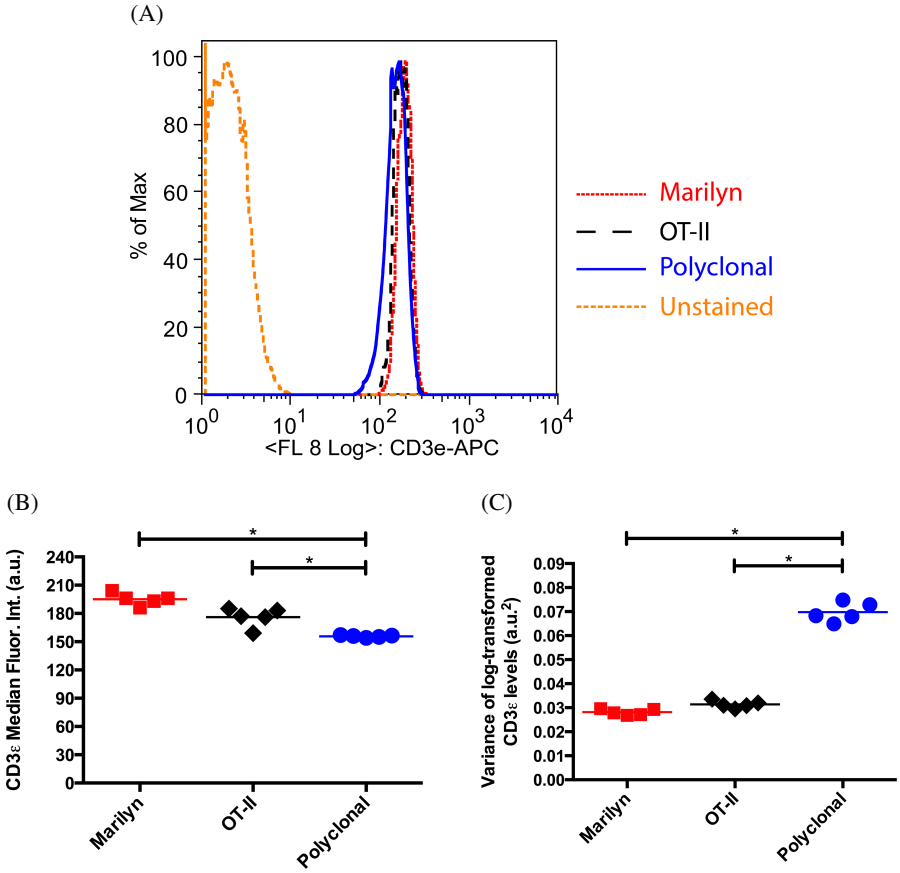

Figure 3.14: Comparison of TCR expression levels between the TCR-transgenic *Rag2*$^{-/-}$ populations Marilyn and OT-II, along with a naive polyclonal population, based on staining for CD3ϵ (second experiment). For consistency with the setup in section 3.4, analysis of all populations was done based on the same gating strategy used for cell sorting (see methods), followed by gating for CD3ϵ^{+}CD4^{+} events. **(A)** Illustrative histograms of TCR levels of the populations. **(B)** Quantification of the median TCR levels, considering untransformed values. **(C)** Quantification of the variance of log-transformed values (σ^2_T). Data correspond to the second of 2 independent experiments with 5 mice per group. $* \; p < 0.05$.

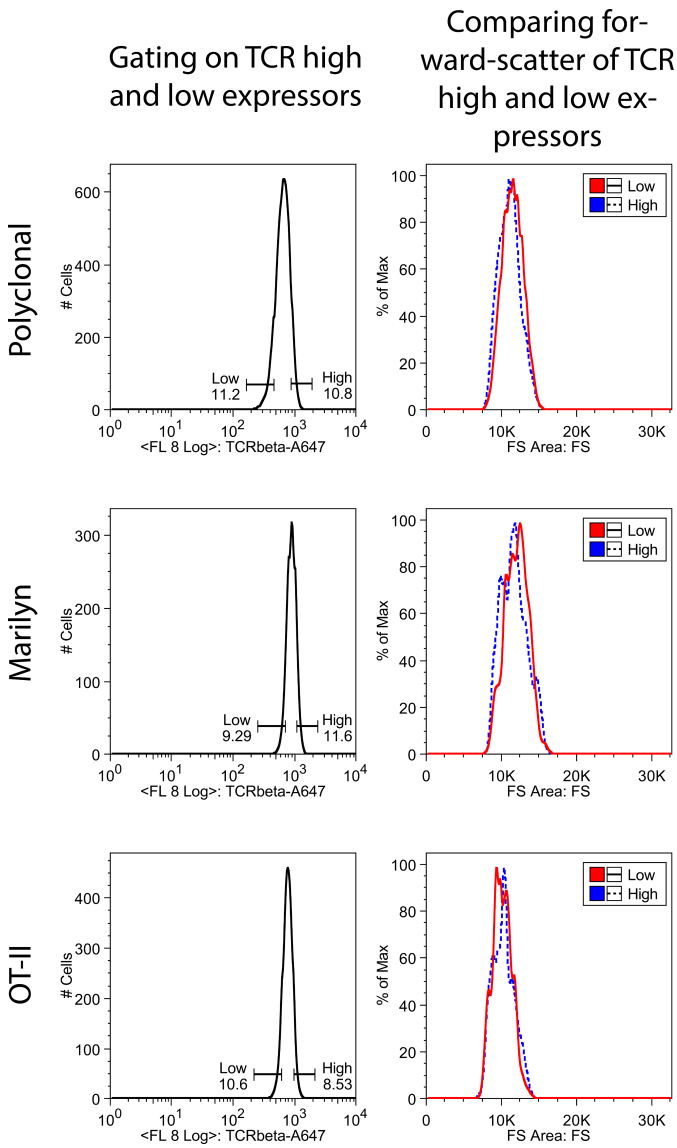

Figure 3.15: Comparison of the relationship between TCR expression levels and forward-scatter on the TCR-transgenic *Rag2*$^{-/-}$ populations Marilyn and OT-II, along with a naive polyclonal population, based on staining for TCRβ (second experiment). For consistency with the setup in section 3.4, analysis of all populations was done based on the same gating strategy used for cell sorting (see methods), followed by gating for TCRβ^{+}CD4^{+} events. Cells from each population, as in figure 3.13, were gated as around 10% into high and low expressors based on TCR levels, and compared in terms of the forward-scatter (often used in flow cytometry as an initial approximation of cell size). Data correspond to the second of 2 independent experiments.

Appendix 3.B Overview of the *in vitro* data on the polyclonal and TCR-transgenic populations

This section presents additional details of the experimental data analyzed in section 3.4. Figure 3.16 presents the fold-ratios between the medians of TCR expression levels (untransformed values) of the different populations. In particular, it shows the ratio between the medians of the all expressors and the control population (see methods), which is always very close to unity, demonstrating that the staining for sorting does not induce changes in TCR expression levels, and hence that the approach for quantification of TCR in the sorted cells is reliable. As discussed in section 2.5.2 (page 48), equation 2.29 implies that the value of $\Delta_{H,L}(t)$ can be approximated by the logarithm of the fold-ratio between the means of high and low expressors. As a consequence, the overall shapes of the curves for high and low expressors in the figure resemble that of the data in figure 3.6 (section 3.4).

The dynamics of high and low sorted from the Marilyn, OT-II (both of which are on a *Rag2*$^{-/-}$ background) and polyclonal populations, in one typical experiment with each biological population, are shown in figure 3.17. In accordance with the unstable component being the main contribution in the Marilyn and OT-II TCR-transgenic populations, as reflected by the relatively small values of R^2_α in section 3.4, high and low expressors from these two populations become very similar, as a function of the time after sorting. On the other hand, since in the most appropriate description of the data (section 3.4) the polyclonal population the stable component is the main contribution ($R^2_\alpha \approx 70\%$), high and low expressors sorted from this population remain clearly different, even 96 hours after sorting.

Materials and methods

The values of the median fluorescence intensity were obtained as reported by FlowJo 8.8.7 (Tree Star Inc.). Confidence intervals for the ratio between the median fluorescence intensities were calculated by uncertainty propagation (Taylor, 1997, chap. 3), based on the standard error of the mean.

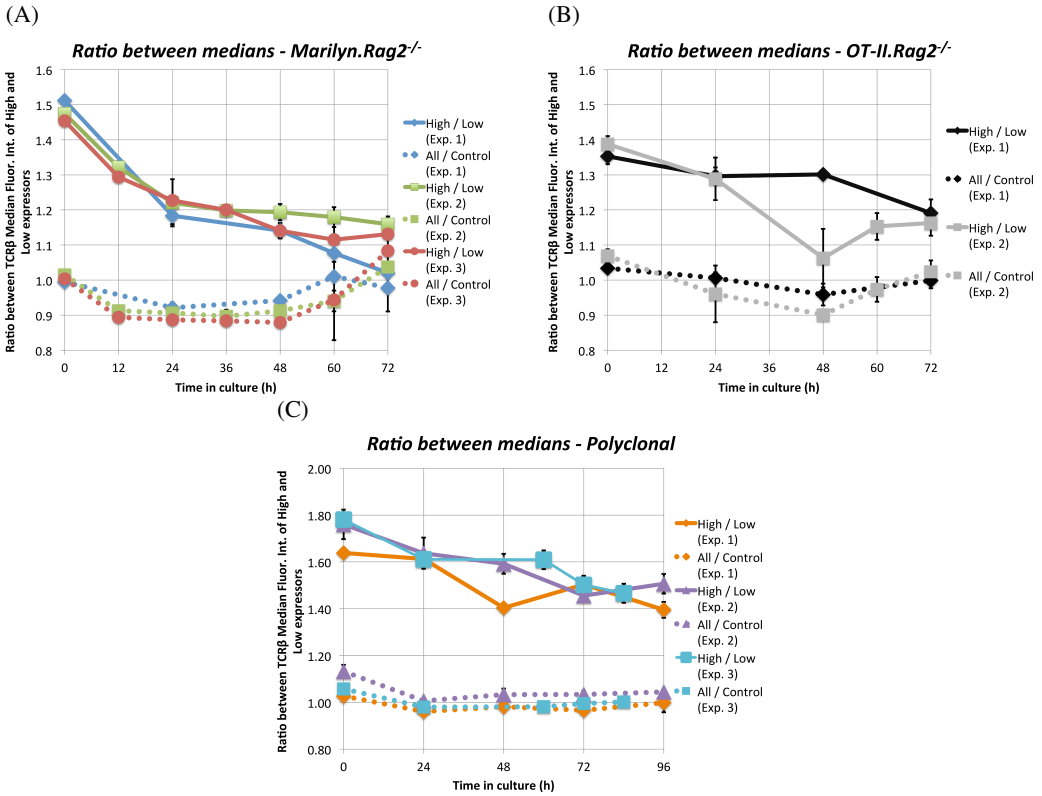

Figure 3.16: Details of the experimental data on TCR levels in unstimulated cells *in vitro*, in terms of the fold-ratio between the medians of TCR intensities (untransformed values). Shown are the data for Marilyn (A), OT-II (B) and the polyclonal population (C). In each figure, each pair of lines represents one experiment, with the ratios between high and low expressors as full lines, and between all expressors and the control population (see methods) as dotted lines. Error bars represent the standard error of the mean, calculated using uncertainty propagation.

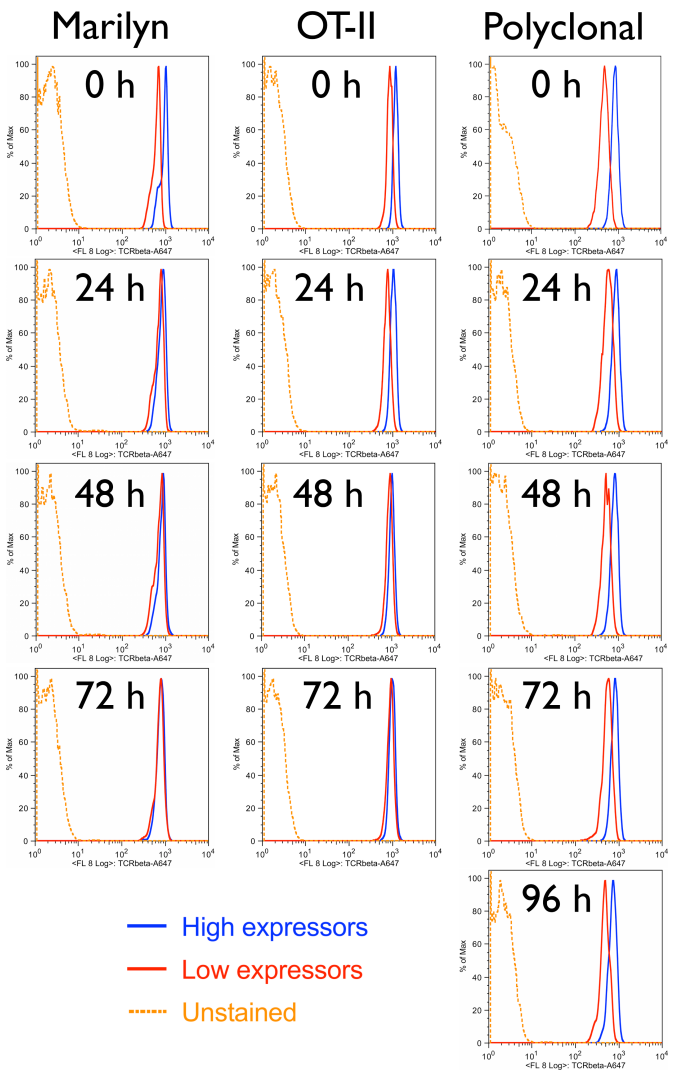

Figure 3.17: Dynamics of high and low expressors sorted from the Marilyn, OT-II and polyclonal populations in the different time-points. One of the three experiments done with each biological population is shown. In each time-point, ranging from 0 hours (after sorting, upon re-staining the cells for analysis) to 96 hours, only one replicate for each sorted population is shown, to facilitate visualization. Each figure includes an unstained population, to provide an estimate of background cellular fluorescence. In the case of the TCR-transgenic populations, cells were tracked only up to 72 hours, due to minimal cell viability afterwards. Data are shown spanning the 4 decades of fluorescence, as the typical approach in the visualization of flow cytometry data

Appendix 3.C Data on the first *in vivo* experiment

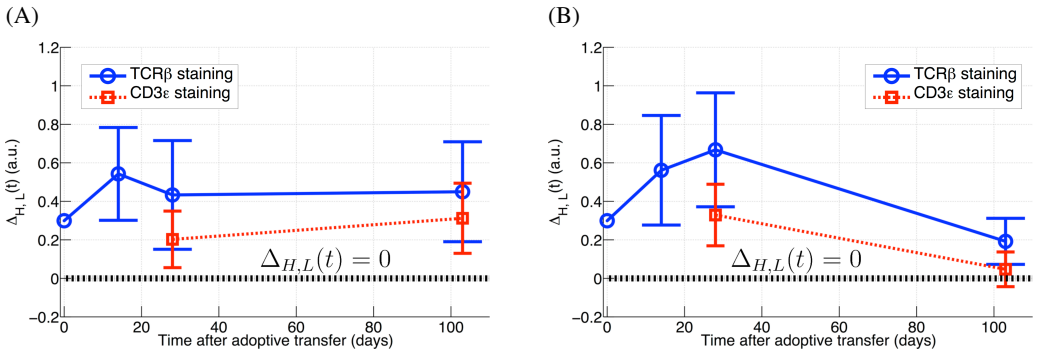

Figure 3.18: The stable component in a polyclonal population is robust to the highly stimulatory conditions provided by the lymphopenic mice upon adoptive transfer (first experiment). The function $\Delta_{H,L}(t)$ was estimated by staining cells for TCRβ or, alternatively, for CD3ϵ. Shown is data for lymph nodes (A) and spleen (B). Error bars denote 95% confidence intervals for each time-point, while the dotted black line highlights the threshold corresponding to $\Delta_{H,L}(t)$. No error bars are shown for the initial time-point ($t = 0$), as that estimate is based on a single replicate of high and low expressors, just before performing the adoptive transfer. Data correspond to the first of 2 independent experiments, with TCR levels quantified in different time-points, each having 3–4 animals per group.

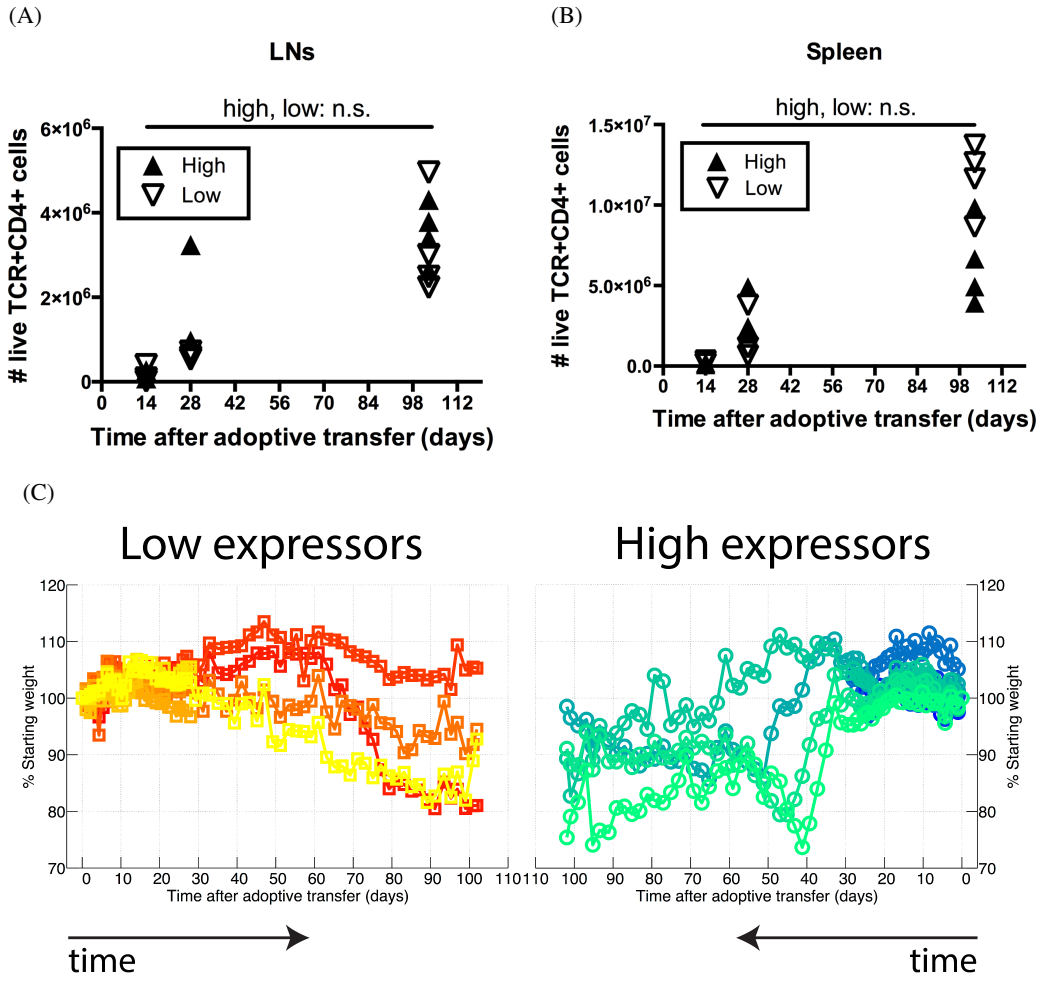

Figure 3.19: High and low expressors have indistinguishable abilities to reconstitute the peripheral pool (lymph nodes (A) and spleen (B); filled triangles pointing up denote animals that received high expressors, while open triangles pointing down denote those that received low expressors), and to induce weight loss (C) upon adoptive transfer to lymphopenic (*Rag2*$^{-/-}$) recipients (first experiment). Data correspond to the first of 2 independent experiments, with 10 animals per group at the start; n.s.: differences between the number of cells in animals receiving high or low expressors are not significant, based on a two-way ANOVA.

Appendix 3.D Analysis of the data on the hybridoma populations from (Bonnet et al., 2009)

This section provides additional information on the data on the two hybridoma populations, extracted from Bonnet et al. (2009). The two populations are clones grown out of hybridomas derived from WT CD4$^+$ T cells, referred to as WT hybridoma, and from CD4$^+$ T cells expressing a reduced version of the Eβ enhancer (E$\beta^{169/169}$ enhancer, or simply Eβ^{169}), referred to as mutant hybridoma. These data correspond, at the beginning of the experiment, to a histogram of the starting population and the thresholds used for isolating high and low expressors. The time-points for analysis of high and low expressors are available after 2 and 5 weeks in culture, in which as expected for these hybrid cell lines, there was extensive cell division taking place (M. Bonnet, personal communication).

The values of $\Delta_{H,L}(t)$ are shown in figure 3.20. It is interesting to note that the initial values $\Delta_{H,L}(0)$ for the two hybridomas are considerably larger than those for the *ex vivo* cells (figure 3.6), even though a greater percentile of cells were sorted as high and low expressors from the two hybridoma populations (estimated as 30–40% of the starting population). This is expected, since the total variation in the hybridomas (estimated as $\sigma_T^2 \approx 0.7$) is much larger than in the *ex vivo* cells (in which σ_T^2 ranges from 0.03 to 0.07). That hybridomas have greater total variation (σ_T^2) is consistent with data on two TCR-transgenic populations (on a Rag-deficient background) and hybridomas derived from them (Mandl et al., 2013).

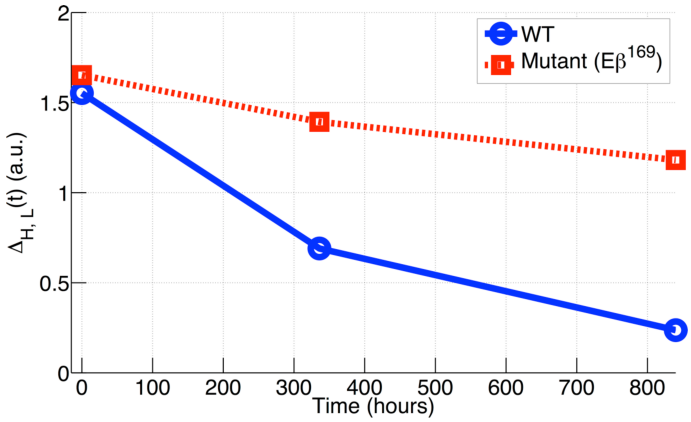

Figure 3.20: Summary of the data extracted from Bonnet et al. (2009), regarding TCR expression levels in clonal hybridomas, shown in terms of function $\Delta_{H,L}(t)$ for the two hybridomas.

Chapter 3

Bibliography

Alt, F. W. and Baltimore, D. (1982). Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-JH fusions. *PNAS*, 79(13):4118–22.

Azzam, H. S., Grinberg, A., Lui, K., Shen, H., Shores, E. W., and Love, P. E. (1998). CD5 expression is developmentally regulated by T cell receptor (TCR) signals and TCR avidity. *The Journal of Experimental Medicine*, 188(12):2301–11.

Babinet, C. (2000). Transgenic mice: an irreplaceable tool for the study of mammalian development and biology. *Journal of the American Society of Nephrology*, 11 Suppl 1:S88–94.

Baniyash, M. (2004). TCR zeta-chain downregulation: curtailing an excessive inflammatory immune response. *Nature Reviews Immunology*, 4(9):675–87.

Barnden, M. J., Allison, J., Heath, W. R., and Carbone, F. R. (1998). Defective TCR expression in transgenic mice constructed using cDNA-based alpha- and beta-chain genes under the control of heterologous regulatory elements. *Immunology and Cell Biology*, 76(1):34–40.

Bauer, A., McConkey, D. J., Howard, F. D., Clayton, L. K., Novick, D., Koyasu, S., and Reinherz, E. L. (1991). Differential signal transduction via T-cell receptor CD3 zeta 2, CD3 zeta-eta, and CD3 eta 2 isoforms. *PNAS*, 88(9):3842–6.

Bendelac, A., Savage, P. B., and Teyton, L. (2007). The biology of NKT cells. *Annual Review of Immunology*, 25:297–336.

Bonifacino, J. S., Chen, C., Lippincott-Schwartz, J., Ashwell, J. D., and Klausner, R. D. (1988). Subunit interactions within the T-cell antigen receptor: clues from the study of partial complexes. *PNAS*, 85(18):6929–33.

Bonnet, M., Huang, F., Benoukraf, T., Cabaud, O., Verthuy, C., Boucher, A., Jaeger, S., Ferrier, P., and Spicuglia, S. (2009). Duality of enhancer functioning mode revealed in a reduced TCR beta gene enhancer knockin mouse model. *The Journal of Immunology*, 183(12):7939–48.

Bosc, N. and Lefranc, M.-P. (2003). The mouse (Mus musculus) T cell receptor alpha (TRA) and delta (TRD) variable genes. *Developmental & Comparative Immunology*, 27(6-7):465–497.

Bosc, N., Lefranc, M.-P., and Ginestoux, C. (2011). Locus representation: Mouse (Mus musculus) TRB, IMGT Repertoire (IG and TR). *IMGT(R), the international ImMunoGenetics information system(R) http://www.imgt.org*, page Created: 15/01/2001. Version: 29/06/2011.

Brady, B. L., Steinel, N. C., and Bassing, C. H. (2010). Antigen receptor allelic exclusion: an update and reappraisal. *The Journal of Immunology*, 185(7):3801–8.

Bronstein-Sitton, N., Wang, L., Cohen, L., and Baniyash, M. (1999). Expression of the T cell antigen receptor zeta chain following activation is controlled at distinct checkpoints. Implications for cell surface receptor down-modulation and re-expression. *The Journal of Biological Chemistry*, 274(33):23659–65.

Burnham, K. P. and Anderson, D. R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York, NY.

Calado, D. P., Paixão, T., Holmberg, D., and Haury, M. (2006). Stochastic monoallelic expression of IL-10 in T cells. *The Journal of Immunology*, 177(8):5358–64.

Call, M. E., Pyrdol, J., Wiedmann, M., and Wucherpfennig, K. W. (2002). The organizing principle in the formation of the T cell receptor-CD3 complex. *Cell*, 111(7):967–79.

Call, M. E., Pyrdol, J., and Wucherpfennig, K. W. (2004). Stoichiometry of the T-cell receptor-CD3 complex and key intermediates assembled in the endoplasmic reticulum. *The EMBO Journal*, 23(12):2348–57.

Call, M. E. and Wucherpfennig, K. W. (2005). The T cell receptor: critical role of the membrane environment in receptor assembly and function. *Annual Review of Immunology*, 23:101–25.

Caramalho, I., Lopes-Carvalho, T., Ostler, D., Zelenay, S., Haury, M., and Demengeot, J. (2003). Regulatory T Cells Selectively Express Toll-like Receptors and Are Activated by Lipopolysaccharide. *The Journal of Experimental Medicine*, 197(4):403–411.

Carneiro, J., Duarte, L., and Padovan, E. (2009). Limiting Dilution Analysis of Antigen-Specific T Cells. In Libero, G., editor, *T Cell Protocols*, volume 514 of *Methods in Molecular Biology*, chapter 7, pages 95–105. Humana Press, Totowa, NJ, 2 edition.

Chen, F., Rowen, L., Hood, L., and Rothenberg, E. V. (2001). Differential transcriptional regulation of individual TCR V beta segments before gene rearrangement. *The Journal of Immunology*, 166(3):1771–80.

Clayton, L. K., D'Adamio, L., Howard, F. D., Sieh, M., Hussey, R. E., Koyasu, S., and Reinherz, E. L. (1991). CD3 eta and CD3 zeta are alternatively spliced products of a common genetic locus and are transcriptionally and/or post-transcriptionally regulated during T-cell development. *PNAS*, 88(12):5202–6.

Cobb, R. M., Oestreich, K. J., Osipovich, O. A., and Oltz, E. M. (2006). Accessibility control of V(D)J recombination. *Advances in Immunology*, 91:45–109.

Dave, V. P. (2009). Hierarchical role of CD3 chains in thymocyte development. *Immunological Reviews*, 232(1):22–33.

Davis, M. M. and Bjorkman, P. J. (1988). T-cell antigen receptor genes and T-cell recognition. *Nature*, 334(6181):395–402.

Deenick, E. K., Gett, A. V., and Hodgkin, P. D. (2003). Stochastic model of T cell proliferation: a calculus revealing IL-2 regulation of precursor frequencies, cell cycle time, and survival. *The Journal of Immunology*, 170(10):4963–72.

Durant, L., Watford, W. T., Ramos, H. L., Laurence, A., Vahedi, G., Wei, L., Takahashi, H., Sun, H.-W., Kanno, Y., Powrie, F., and O'Shea, J. J. (2010). Diverse targets of the transcription factor STAT3 contribute to T cell pathogenicity and homeostasis. *Immunity*, 32(5):605–15.

Farber, D. L., Luqman, M., Acuto, O., and Bottomly, K. (1995). Control of memory CD4 T cell activation: MHC class II molecules on APCs and CD4 ligation inhibit memory but not naive CD4 T cells. *Immunity*, 2(3):249–59.

Feinerman, O., Veiga, J., Dorfman, J. R., Germain, R. N., and Altan-Bonnet, G. (2008). Variability and robustness in T cell activation from regulated heterogeneity in protein levels. *Science*, 321(5892):1081–4.

Feng, T., Wang, L., Schoeb, T. R., Elson, C. O., and Cong, Y. (2010). Microbiota innate stimulation is a prerequisite for T cell spontaneous proliferation and induction of experimental colitis. *The Journal of Experimental Medicine*, 207(6):1321–32.

Fernández-Miguel, G., Alarcón, B., Iglesias, A., Bluethmann, H., Alvarez-Mon, M., Sanz, E., and de la Hera, A. (1999). Multivalent structure of an alpha beta T cell receptor. *PNAS*, 96(4):1547–52.

Gallucci, S., Lolkema, M., and Matzinger, P. (1999). Natural adjuvants: endogenous activators of dendritic cells. *Nature Medicine*, 5(11):1249–55.

Gascoigne, N. R., Chien, Y., Becker, D. M., Kavaler, J., and Davis, M. M. (1984). Genomic organization and sequence of T-cell receptor beta-chain constant- and joining-region genes. *Nature*, 310(5976):387–91.

Geisler, C. (1992). Failure to synthesize the CD3-gamma chain. Consequences for T cell antigen receptor assembly, processing, and expression. *The Journal of Immunology*, 148(8):2437–45.

Gellert, M. (2002). V(D)J recombination: RAG proteins, repair factors, and regulation. *Annual Review of Biochemistry*, 71(D):101–32.

Germain, R. N. (2002). T-cell development and the CD4-CD8 lineage decision. *Nature Reviews Immunology*, 2(5):309–22.

Ghendler, Y., Smolyar, A., Chang, H. C., and Reinherz, E. L. (1998). One of the CD3epsilon subunits within a T cell receptor complex lies in close proximity to the Cbeta FG loop. *The Journal of Experimental Medicine*, 187(9):1529–36.

Glusman, G., Rowen, L., Lee, I., Boysen, C., Roach, J. C., Smit, A. F., Wang, K., Koop, B. F., and Hood, L. (2001). Comparative genomics of the human and mouse T cell receptor loci. *Immunity*, 15(3):337–49.

Grandjean, I., Duban, L., Bonney, E. A., Corcuff, E., Di Santo, J. P., Matzinger, P., and Lantz, O. (2003). Are major histocompatibility complex molecules involved in the survival of naive CD4+ T cells? *The Journal of Experimental Medicine*, 198(7):1089–102.

Greenberg, N. M., Anderson, J. W., Hsueh, A. J., Nishimori, K., Reeves, J. J., DeAvila, D. M., Ward, D. N., and Rosen, J. M. (1991). Expression of biologically active heterodimeric bovine follicle-stimulating hormone in milk of transgenic mice. *PNAS*, 88(19):8327–31.

Guimond, M., Veenstra, R. G., Grindler, D. J., Zhang, H., Cui, Y., Murphy, R. D., Kim, S. Y., Na, R., Hennighausen, L., Kurtulus, S., Erman, B., Matzinger, P., Merchant, M. S.,

and Mackall, C. L. (2009). Interleukin 7 signaling in dendritic cells regulates the homeostatic proliferation and niche size of CD4+ T cells. *Nature Immunology*, 10(2):149–57.

Hayday, A. C. (2000). [gamma][delta] cells: a right time and a right place for a conserved third way of protection. *Annual Review of Immunology*, 18(c):975–1026.

Hayes, S. M., Shores, E. W., and Love, P. E. (2003). An architectural perspective on signaling by the pre-, alphabeta and gammadelta T cell receptors. *Immunological Reviews*, 191:28–37.

Hynes, R. O. (2002). Integrins: bidirectional, allosteric signaling machines. *Cell*, 110(6):673–87.

Jelley-Gibbs, D. M., Dibble, J. P., Filipson, S., Haynes, L., Kemp, R. A., and Swain, S. L. (2005). Repeated stimulation of CD4 effector T cells can limit their protective function. *The Journal of Experimental Medicine*, 201(7):1101–12.

Jouvin-Marche, E., Fuschiotti, P., and Marche, P. N. (2009). Dynamic aspects of TCRalpha gene recombination: qualitative and quantitative assessments of the TCRalpha chain repertoire in man and mouse. *Advances in Experimental Medicine and Biology*, 650(D):82–92.

Kaplan, M. H. (2013). Th9 cells: differentiation and disease. *Immunological Reviews*, 252(1):104–15.

Kassiotis, G., Zamoyska, R., and Stockinger, B. (2003). Involvement of avidity for major histocompatibility complex in homeostasis of naive and memory T cells. *The Journal of Experimental Medicine*, 197(8):1007–16.

Kearse, K. P., Roberts, J. L., and Singer, A. (1995). TCR alpha-CD3 delta epsilon association is the initial step in alpha beta dimer formation in murine T cells and is limiting in immature CD4+ CD8+ thymocytes. *Immunity*, 2(4):391–9.

Kieper, W. C., Troy, A., Burghardt, J. T., Ramsey, C., Lee, J. Y., Jiang, H.-q., Dummer, W., Shen, H., Cebra, J. J., and Surh, C. D. (2005). Recent immune status determines the source of antigens that drive homeostatic T cell expansion. *The Journal of Immunology*, 174(6):3158–63.

Klausner, R. D., Lippincott-Schwartz, J., and Bonifacino, J. S. (1990). The T cell antigen receptor: insights into organelle biology. *Annual Review of Cell Biology*, 6:403–31.

Koch, U. and Radtke, F. (2011). Mechanisms of T cell development and transformation. *Annual Review of Cell and Developmental Biology*, 27:539–62.

Kuhns, M. S. and Badgandi, H. B. (2012). Piecing together the family portrait of TCR-CD3 complexes. *Immunological Reviews*, 250(1):120–43.

Kuhns, M. S. and Davis, M. M. (2012). TCR Signaling Emerges from the Sum of Many Parts. *Frontiers in Immunology*, 3(June):159.

Kuhns, M. S., Davis, M. M., and Garcia, K. C. (2006). Deconstructing the form and function of the TCR/CD3 complex. *Immunity*, 24(2):133–9.

Kuhns, M. S., Girvin, A. T., Klein, L. O., Chen, R., Jensen, K. D. C., Newell, E. W., Huppa, J. B., Lillemeier, B. F., Huse, M., Chien, Y.-H., Garcia, K. C., and Davis, M. M. (2010). Evidence for a functional sidedness to the alphabetaTCR. *PNAS*, 107(11):5094–9.

Lafaille, J. J. (2004). T-cell receptor transgenic mice in the study of autoimmune diseases. *Journal of Autoimmunity*, 22(2):95–106.

Lantz, O., Grandjean, I., Matzinger, P., and Di Santo, J. P. (2000). Gamma chain required for naïve CD4+ T cell survival but not for antigen proliferation. *Nature Immunology*, 1(1):54–8.

Lee, W. T., Yin, X. M., and Vitetta, E. S. (1990). Functional and ontogenetic analysis of murine CD45Rhi and CD45Rlo CD4+ T cells. *The Journal of Immunology*, 144(9):3288–95.

Levine, B. L., Bernstein, W. B., Connors, M., Craighead, N., Lindsten, T., Thompson, C. B., and June, C. H. (1997). Effects of CD28 costimulation on long-term proliferation of CD4+ T cells in the absence of exogenous feeder cells. *The Journal of Immunology*, 159(12):5921–30.

Liew, F. Y. (2002). T(H)1 and T(H)2 cells: a historical perspective. *Nature Reviews Immunology*, 2(1):55–60.

Liu, H., Rhodes, M., Wiest, D. L., and Vignali, D. A. (2000). On the dynamics of TCR:CD3 complex cell surface expression and downmodulation. *Immunity*, 13(5):665–75.

Love, P. E., Shores, E. W., Johnson, M. D., Tremblay, M. L., Lee, E. J., Grinberg, A., Huang, S. P., Singer, A., and Westphal, H. (1993). T cell development in mice that lack the zeta chain of the T cell antigen receptor complex. *Science*, 261(5123):918–21.

Mandl, J. N., Monteiro, J. P., Vrisekoop, N., and Germain, R. N. (2013). T cell-positive selection uses self-ligand binding strength to optimize repertoire recognition of foreign antigens. *Immunity*, 38(2):263–74.

Manolios, N., Letourneur, F., Bonifacino, J. S., and Klausner, R. D. (1991). Pairwise, cooperative and inhibitory interactions describe the assembly and probable structure of the T-cell antigen receptor. *The EMBO Journal*, 10(7):1643–51.

Mariani, L., Schulz, E. G., Lexberg, M. H., Helmstetter, C., Radbruch, A., Löhning, M., and Höfer, T. (2010). Short-term memory in gene induction reveals the regulatory principle behind stochastic IL-4 expression. *Molecular Systems Biology*, 6(359):359.

Market, E. and Papavasiliou, F. N. (2003). V(D)J recombination and the evolution of the adaptive immune system. *PLoS Biology*, 1(1):E16.

Min, B., Yamane, H., Hu-Li, J., and Paul, W. E. (2005). Spontaneous and homeostatic proliferation of CD4 T cells are regulated by different mechanisms. *The Journal of Immunology*, 174(10):6039–44.

Minami, Y., Weissman, a. M., Samelson, L. E., and Klausner, R. D. (1987). Building a multichain receptor: synthesis, degradation, and assembly of the T-cell antigen receptor. *PNAS*, 84(9):2688–92.

Moon, J. J., Chu, H. H., Pepper, M., McSorley, S. J., Jameson, S. C., Kedl, R. M., and Jenkins, M. K. (2007). Naive CD4(+) T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude. *Immunity*, 27(2):203–13.

Mosmann, T. R., Cherwinski, H., Bond, M. W., Giedlin, M. A., and Coffman, R. L. (1986). Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins. *The Journal of Immunology*, 136(7):2348–57.

Mostoslavsky, R., Alt, F. W., and Rajewsky, K. (2004). The lingering enigma of the allelic exclusion mechanism. *Cell*, 118(5):539–44.

Nikolich-Zugich, J., Slifka, M. K., and Messaoudi, I. (2004). The many important facets of T-cell repertoire diversity. *Nature Reviews Immunology*, 4(2):123–32.

Oettinger, M. A., Schatz, D. G., Gorka, C., and Baltimore, D. (1990). RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. *Science*, 248(4962):1517–23.

Paixão, T. (2007). *The Stochastic Basis of Somatic Variation*. PhD thesis, University of Porto.

Paixão, T., Carvalho, T. P., Calado, D. P., and Carneiro, J. (2007). Quantitative insights into stochastic monoallelic expression of cytokine genes. *Immunology and Cell Biology*, 85(4):315–22.

Palmer, M. J., Mahajan, V. S., Chen, J., Irvine, D. J., and Lauffenburger, D. A. (2011). Signaling thresholds govern heterogeneity in IL-7-receptor-mediated responses of naïve CD8(+) T cells. *Immunology and Cell Biology*, 89(5):581–94.

Paoletti, P., Bellone, C., and Zhou, Q. (2013). NMDA receptor subunit diversity: impact on receptor properties, synaptic plasticity and disease. *Nature Reviews Neuroscience*, 14(6):383–400.

Paul, W. E. (2003). *Fundamental Immunology*. Lippincott Williams & Wilkins, 5 edition.

Punt, J. A., Roberts, J. L., Kearse, K. P., and Singer, A. (1994). Stoichiometry of the T cell antigen receptor (TCR) complex: each TCR/CD3 complex contains one TCR alpha, one TCR beta, and two CD3 epsilon chains. *The Journal of Experimental Medicine*, 180(2):587–93.

Raj, A. and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–26.

Saito, T., Sussman, J. L., Ashwell, J. D., and Germain, R. N. (1989). Marked differences in the efficiency of expression of distinct alpha beta T cell receptor heterodimers. *The Journal of Immunology*, 143(10):3379–84.

Sakaguchi, S., Yamaguchi, T., Nomura, T., and Ono, M. (2008). Regulatory T cells and immune tolerance. *Cell*, 133(5):775–87.

Schatz, D. G., Oettinger, M. A., and Baltimore, D. (1989). The V(D)J recombination activating gene, RAG-1. *Cell*, 59(6):1035–48.

Schrum, A. G., Gil, D., Turka, L. A., and Palmer, E. (2011). Physical and functional bivalency observed among TCR/CD3 complexes isolated from primary T cells. *The Journal of Immunology*, 187(2):870–8.

Schrum, A. G., Palmer, E., and Turka, L. A. (2005). Distinct temporal programming of naive CD4+ T cells for cell division versus TCR-dependent death susceptibility by antigen-presenting macrophages. *European Journal of Immunology*, 35(2):449–59.

Schrum, A. G., Turka, L. A., and Palmer, E. (2003). Surface T-cell antigen receptor expression and availability for long-term antigenic signaling. *Immunological Reviews*, 196:7–24.

Seber, G. A. F. and Wild, C. J. (2003). *Nonlinear Regression*. John Wiley & Sons.

Shores, E. W., Huang, K., Tran, T., Lee, E., Grinberg, A., and Love, P. E. (1994). Role of TCR zeta chain in T cell development and selection. *Science*, 266(5187):1047–50.

Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Liron, Y., Rosenfeld, N., Danon, T., Perzov, N., and Alon, U. (2006). Variability and memory of protein levels in human cells. *Nature*, 444(7119):643–6.

Singer, A., Adoro, S., and Park, J.-H. (2008). Lineage fate and intense debate: myths, models and mechanisms of CD4- versus CD8-lineage choice. *Nature Reviews Immunology*, 8(10):788–801.

Smith, K., Seddon, B., Purbhoo, M. A., Zamoyska, R., Fisher, A. G., and Merkenschlager, M. (2001). Sensory adaptation in naive peripheral CD4 T cells. *The Journal of Experimental Medicine*, 194(9):1253–61.

Smith-Garvin, J. E., Koretzky, G. A., and Jordan, M. S. (2009). T cell activation. *Annual Review of Immunology*, 27:591–619.

Sokal, R. R. and Rohlf, F. J. (1981). *Biometry*. Freeman.

Sousa, J. and Carneiro, J. (2000). A mathematical analysis of TCR serial triggering and down-regulation. *European Journal of Immunology*, 30(11):3219–27.

Starr, T. K., Jameson, S. C., and Hogquist, K. A. (2003). Positive and negative selection of T cells. *Annual Review of Immunology*, 21:139–76.

Sun, Z.-y. J., Kim, S. T., Kim, I. C., Fahmy, A., Reinherz, E. L., and Wagner, G. (2004). Solution structure of the CD3epsilondelta ectodomain and comparison with CD3epsilongamma as a basis for modeling T cell receptor topology and signaling. *PNAS*, 101(48):16867–72.

Surh, C. D. and Sprent, J. (2008). Homeostasis of naive and memory T cells. *Immunity*, 29(6):848–62.

Swanson, P. C. (2004). The bounty of RAGs: recombination signal complexes and reaction outcomes. *Immunological Reviews*, 200:90–114.

Taylor, J. R. (1997). *An introduction to error analysis: the study of uncertainties in physical measurements*. University Science Books, 2 edition.

Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature*, 302(5909):575–81.

Tubo, N. J., Pagán, A. J., Taylor, J. J., Nelson, R. W., Linehan, J. L., Ertelt, J. M., Huseby, E. S., Way, S. S., and Jenkins, M. K. (2013). Single Naive CD4+ T Cells from a Diverse Repertoire Produce Different Effector Cell Types during Infection. *Cell*, 153(4):785–796.

Wang, Y., Rickman, B. H., Poutahidis, T., Schlieper, K., Jackson, E. A., Erdman, S. E., Fox, J. G., and Horwitz, B. H. (2008). c-Rel is essential for the development of innate and T cell-induced colitis. *The Journal of Immunology*, 180(12):8118–25.

Weaver, C. T., Elson, C. O., Fouser, L. A., and Kolls, J. K. (2013). The Th17 pathway and inflammatory diseases of the intestines, lungs, and skin. *Annual Review of Pathology*, 8:477–512.

Wilson, C. B., Rowell, E., and Sekimata, M. (2009). Epigenetic control of T-helper-cell differentiation. *Nature Reviews Immunology*, 9(2):91–105.

Wucherpfennig, K. W., Gagnon, E., Call, M. J., Huseby, E. S., and Call, M. E. (2010). Structural biology of the T-cell receptor: insights into receptor assembly, ligand recognition, and initiation of signaling. *Cold Spring Harbor Perspectives in Biology*, 2(4):a005140.

Zhu, J., Yamane, H., and Paul, W. E. (2010). Differentiation of effector CD4 T cell populations. *Annual Review of Immunology*, 28:445–89.

Chapter 4

GENERAL DISCUSSION

Populations of cells are inherently heterogeneous, especially in terms of the amount of a particular protein or mRNA that is expressed, as often observed in a snapshot of the expression levels in the population. One aspect that is expected to contribute to this variation is the presence of stochastic fluctuations in the level expressed by each cell throughout time. With these fluctuations being asynchronous in different cells of the population, they would give rise to the variation that is observed in a snapshot of the population. Several studies (Elowitz et al., 2002; Raser and O'Shea, 2004; Rosenfeld et al., 2005; Sigal et al., 2006; Raj and van Oudenaarden, 2008) have pointed to the inherent stochasticity in the process of gene expression, simply referred to as noise in gene expression, as one process giving rise to such fluctuations.

Indeed, many works (Swain et al., 2002; Munsky et al., 2009; Rinott et al., 2011; Komorowski et al., 2013), focusing on isogenic cell populations, have developed approaches for quantifying the properties of cell populations. In doing so, it is implicitly assumed that noise, and hence the resulting fluctuations in single cells as a function of time, are the only factor underlying the variation that is observed in a snapshot of a cell population. Under this assumption, one would expect that the differences in the expression levels of two subsets of cells in a population would be transient. However, it remains unclear to which degree the potential presence of *stable variants* in the cell population would contribute to the variation that is observed. By stable variants (Chang et al., 2008), we refer to subsets of cells that are biased, by whichever mechanism, to have a limited range of expression levels, compared with those observed in the population. Such stable variants could come about if the population is genetically diverse (Market and Papavasiliou, 2003; Yates and Campbell, 2012), and/or due to epigenetic mechanisms, the latter notion being used in a broad sense, to refer to mechanisms that establish and maintain variant cell states (Jablonka and Raz, 2009). Such mechanisms underlie epigenetic variation, leading to phenotypic differences that are not explained differences by changes in DNA sequence, and persist throughout time.

Therefore, in the general case the basis for the variation in a cell population can be thought of as being due to a particular combination of the fluctuations in the levels of single cells throughout time and the presence of stable variants. On this subject, several works have relied on isolating cells from a given population, such as via cell sorting (Chang et al., 2008; Kalmar et al., 2009; Luo et al., 2012; Sisan et al., 2012), and analyzing statistics of their expression levels as a function of time. However, an important limitation is the lack of general approaches for quantifying properties and comparing the populations in this setup.

This thesis addressed the basis for variation in expression levels in a cell population. In order to provide formal definitions, and to study how to perform inference, chapter 2 developed a quantitative theoretical framework. This framework is based on classifying the mechanisms that shape expression levels in a cell population into two components, one stable and the other unstable. In an informal way, the stable component represents permanent differences between the expression levels of two subsets of cells, while the unstable component, on the other hand, represents transient differences, which will eventually vanish over time. This led to the concept of a sub-population, a set of cells in which all variation observed is due to the unstable component, and therefore that differences between the expression levels of two subsets of cells that belong to the same sub-population are transient. We used these definitions to describe protein expression levels in a heterogeneous cell population, based on a simplified model of constitutive protein expression, and to study how to infer parameters that quantify properties of the expression levels in such a cell population. Based on this formulation, the estimation of two parameters, termed R^2_α, quantifying the relative contribution of the stable component to the variation, and τ_T, referred to as the characteristic time of the variation, was considered, via isolating subsets of cells and quantifying their expression levels throughout time. The framework developed is general, and can be used to study genetically diverse or isogenic cell populations.

Afterwards, based on this theoretical framework, chapter 3 addressed experimentally the contributions to variation in expression levels of the T Cell Receptor (TCR) in $CD4^+$ T cells. We quantified the relative contribution of the stable and unstable components (R^2_α) and the characteristic time of the variation (τ_T) in expression levels of the TCR in a polyclonal (genetically diverse) and in two isogenic (TCR-transgenic on a *Rag2*$^{-/-}$ background) populations. In the polyclonal population, both genetic and epigenetic variation have the potential to mold the stable component, while in the TCR-transgenic populations, the stable component is due to epigenetic variation. By evaluating the relative contribution of the two components in these various populations, we assessed the impact of genetic and epigenetic variation on the stable component in the polyclonal population. The description of the data from an *in vitro* setup indicates the unstable component as the main contribution in the two TCR-transgenic populations analyzed, but also provides preliminary evidence for the stable component in these populations. On the other hand, for the polyclonal population, the description of the data in this *in vitro* setup implies that the stable component is the main contribution, and, along with the results of adoptive transfers to *Rag2*$^{-/-}$ recipients, provides strong evidence for the stable component in the polyclonal population. Based on the comparison between the values of R^2_α estimated *in vitro*, it was concluded that in this pop-

ulation genetic variation would be the main explanatory factor for the stable component, and that epigenetic variation would have a relatively small impact. These results establish the TCR in a polyclonal population of $CD4^+$ T cells as a model system to study how the stable and unstable components contribute to variation in expression levels.

Therefore, this thesis contributes to understanding how two different processes, termed the stable and unstable components, shape the expression levels in a cell population by developing a theoretical quantitative framework to formalize and quantify certain properties of the expression levels, and by putting forward an experimental model system to study the interplay of such processes. The theoretical framework defined two parameters, and developed an approach for inferring these parameters by isolating cells and quantifying the statistics of the expression levels of such cells over a sufficiently long period of time. An important contribution of this approach is that it allows one to readily interpret the results even if analysis is done over a limited window of time, in which case the results constrain parameter values that are compatible with the data. Crucially, this theoretical framework grounds an experimental setup that has been used by several authors recently (Chang et al., 2008; Huang, 2009; Kalmar et al., 2009; Pina et al., 2012; Sisan et al., 2012). We then turned to analyze the factors that influence the variation in expression levels of the TCR in $CD4^+$ T cells, putting forward this as an experimental model system to study the interplay between the stable and unstable components. Finally, the present chapter provides a general discussion of the results obtained in this thesis, in a broader context and in relation to other works in the literature.

4.1 Attractors in gene regulatory networks

Recent works rely on the notion of an attractor of a gene regulatory network to conceptualize a particular cell fate (Huang et al., 2005; MacArthur et al., 2009). An attractor denotes a state, in terms of genome-wide expression levels, that is essentially stable and self-sustainable. This stability would arise as a consequence of interactions between genes in a gene regulatory network, making this state relatively robust to noise and external perturbations. These interactions are in the form of feedback loops influencing the expression of various molecules, including transcription factors, with the potential to give rise to a set of so-called molecular switches (Ferrell, 2002; Tyson et al., 2003), displaying a relatively stable, long-lasting pattern of expression in a single cell throughout time. This may also extend to chromatin modifications, as some theoretical studies have predicted these marks to have the potential to be relatively stable once fully established, at least under some con-

ditions (Dodd et al., 2007; Paixão et al., 2007).

Therefore, in this view, the particular genome-wide pattern of expression in a single cell would be maintained for a relatively long time, even though it may be variable among different cells, in the sense of which of these switches are on or off. Focusing on the factors that influence, directly or indirectly, production of the protein of interest, cells from a given population could be clustered in terms of a particular pattern of on/off switches. In this way, as defined in chapter 2, each cluster of cells would be equivalent to a sub-population. In the sub-populations where the balance of these factors is tilted towards those inducing expression, the average rate of protein production, and consequently the expression levels of the protein of interest, would be higher than average, while in sub-populations characterized by lower rates of production, this balance would be shifted in the direction of those leading to decreased expression. However, this mechanism may be highly constrained, in terms of the cell populations and molecules affected, given that transcription factors are often highly pleiotropic, such that the effects of a particular pleiotropic transcription factor being off would directly affect expression of not only one protein, but many. One possibility is that some of these switches are not binary, but attain multiple levels.

With this picture in mind, one consequence of entry into the cell cycle may be the perturbation of the global pattern that is established in a single cell, given the disruption of chromatin that takes place during DNA replication (Alabert and Groth, 2012) and given that, during mitosis, most transcription factors are dissociated from the genome and nuclear organization is disrupted (Egli et al., 2008). Therefore, one can envisage that the daughter cells could, with a certain probability, adopt a particular pattern of expression that is different from that of the mother cell, equivalent to switching among states. Such switching may also come about due to relatively large fluctuations in the expression of one or more molecular switches, so as to revert their state. In other words, by allowing cells to switch among states, the impact of cell division would be akin to inducing "turnover" of the particular pattern of expression. If this is the case, the expected effect would be that of reducing the value of R^2_α, in comparison with that obtained without cell division. In the limit that all the transitions between states have non-zero probabilities, the overall effect would be that of making the unstable component as the only contribution to the variation that is observed, with relatively slow dynamics, giving rise to a memory of expression levels spanning several cellular generations. Moreover, since switching leads to an abrupt change in the time-averaged rate of protein production, it is expected to also modulate the dynamics of changes in the instantaneous rate of protein production, which is related to parameter τ, which in turn affects τ_T. Therefore, switching may also impinge on τ_T, depending on the

expected change in the time-averaged rate of protein production per cell division.

This highlights an important aspect. In the general case in which analysis of the expression levels is constrained to a limited time interval, the processes modulating expression levels can be operationally described based on three categories. First, processes that shape the unstable component, with dynamics that are relatively fast when compared with the duration of analysis. Second, processes that shape the unstable component, but with dynamics that are relatively slow, such that they will appear as effectively stable in the window of observation. And, finally, processes that shape the "true" stable component.

Some experimental studies analyzing the expression levels of a molecule in a cell population have deduced the existence of multiple attractor states based on the observation of the histogram of expression levels being multimodal. By isolating cells and tracking the statistics of expression levels as a function of time, such studies (Chang et al., 2008; Kalmar et al., 2009; Luo et al., 2012; Sisan et al., 2012) have considered, whether explicitly or not, the switching between these states as the explanation for the tendency of the isolated populations to become more similar to each other as a function of time. In this context, some works (Chang et al., 2008; Luo et al., 2012) have directly fitted the distribution to mixture models, and then assigned each component of the mixture to an attractor state. Alternatively, such states have been obtained based on the specific model describing expression levels, whether inferred based on the stationary distribution of expression levels of the population (Sisan et al., 2012) or based on the particular gene network model being considered (Kalmar et al., 2009). In all these studies (Chang et al., 2008; Kalmar et al., 2009; Luo et al., 2012; Sisan et al., 2012), the number of states was determined to be small, equal to 2 or 3.

In chapter 2, the description of a population based on a mixture model formulation centers around the concept of a sub-population. In this way, each sub-population corresponds to a single attractor state, that is maintained indefinitely once established, given that it is assumed that cells cannot switch from one sub-population to the other. Moreover, in chapter 2, it was assumed that the number of sub-populations, and hence the number of attractors, is relatively large. This assumption allowed for the simplification of the analysis, such that the parameters (mean and variance) describing expression levels of each sub-population, which determine the contribution of the stable and unstable components to the variation that is observed, could be replaced by the statistics across all sub-populations.

Given a population that is described in this way, such that each state is equivalent to a sub-population, the case in which there is switching among states can be readily dealt with from a conceptual point of view. For this, recall that a sub-population is defined such

that all variation observed is due to the unstable component. It follows that the impact of switching among states is simply that of defining a sub-population that is described by expression levels that are a combination of those from all the states that can be reached from one another. Formally, such combination is a mixture model of the states. This would likely imply describing each sub-population by a different model compared with the simplified version, with a lognormal distribution of expression levels, considered in section 2.3.1.

The assumption in chapter 2 of the number of states being relatively large is, at least in principle, a mathematical simplification. On the other hand, the consideration of a small number of attractor states in the previously mentioned studies (Chang et al., 2008; Kalmar et al., 2009; Luo et al., 2012; Sisan et al., 2012) stems from arguments of parsimony, given the stationary distribution of expression levels of the population, or from relatively simple, low-dimensional models of gene networks. In particular, the latter lump interactions between the molecule(s) of interest with other components of the genome-wide gene regulatory network into relatively few, constant, parameters, taken to be identical in all cells of the population. This has the effect of averaging out any variation that would be present in other components of the network, so that the small number of states is related to the specific parameterization used, and this may not necessarily be the case when accounting for the full network. For a molecule of interest in a particular cell population, it may be that the "true" number of states of expression levels, dictated by the properties of the gene regulatory network, falls into any of the two cases, being relatively large or small. We note, however, that one possible limitation in experimentally distinguishing between two states is that each may be associated with a relatively large variance in expression levels, and hence a considerable overlap between the ranges of expression levels measured for cells in each state, such that cells would appear to be in a single state. A similar argument has been recently made by Antebi et al. (2013), in the context of comparing models for cell differentiation having 2 or 3 stable states. In such a case, the assumption of chapter 2 of a large number of states would constitute merely a mathematical simplification, amenable to refinement as a consequence of further understanding of the mechanisms regulating the expression levels.

4.2 Variation in expression levels in T cells

A remarkable property of the adaptive immune system in jawed vertebrates is the phenotypic variation of B and T lymphocytes, which are genetically diverse. Consequently, variation in the expression levels of a particular molecule in such lymphocyte populations

may come about due to genetic diversity, epigenetic variation and to fluctuations in the expression levels of each cell throughout time.

The analysis of the (naive) polyclonal population in chapter 3 provided evidence of the stable component in this population. This conclusion was obtained by combining results of the analysis of this population done under different conditions. Under one condition, cells were maintained *in vitro* for up to 4 days, in which neither stimulation of the cells nor cell division is expected, and cells do not interact with APCs. Another condition used was an *in vivo* setup, of adoptive transfer to *Rag2*$^{-/-}$ recipients, in which the cells were analyzed up to around 10 weeks after transfer, upon activation and extensive proliferation. Moreover, together with the results of the analysis of two isogenic T cell populations (TCR-transgenic on a *Rag2*$^{-/-}$ background) in the *in vitro* setup, we concluded that genetic variation would be the main factor underlying the stable component in the polyclonal population. However, the precise mechanism being unclear, in the following we speculate on its basis.

While considerable changes in TCR expression levels take place throughout T cell development, one of the features of the TCR is that, in mature non-activated T cells, its levels seem to be independent of the steady-state contact with antigens (Smith et al., 2001), with further modulation of the levels tending to be associated strictly with activation (Schrum et al., 2003). It is important to highlight an important property of the *in vivo* environment, whether of lymphopenic or lymphoreplete animals, which is the expected vast diversity of ligands that can be presented by APCs (Hunt et al., 1992). Hence, stable variants may be described as being cell-extrinsic, if external signals are required for the variation in the average rate of protein expression in different cells to be maintained, or cell-intrinsic, if the differences are maintained even in the absence of such signals. A third possibility is that the stable variants are due to both cell-intrinsic and cell-extrinsic mechanisms. This is an important aspect to keep in mind because, in a polyclonal population, one would expect some level of variation in the TCR-derived signals that are received by the different clones in the periphery, since they have different TCR specificities. Importantly, note that the distinction between cell-intrinsic and cell-extrinsic regards mechanisms that explain the maintenance of the differences. Therefore, external signals may originally establish the expression levels in different cells, but the underlying mechanism would be classified as cell-intrinsic if the differences are afterwards maintained, for example when such external signals are eliminated.

In order to explain the cell-intrinsic stable component in the mature polyclonal population, two broad, non-exclusive models can be considered. A simple model would rest on differences in the expression levels of the different Vα and Vβ promoters (suggested

mainly for the germline Vβ promoters; Chen et al., 2001) and/or variation in the efficiency of pairing of the chains (Saito et al., 1989). This mechanism would imply that expression of the α and/or β chains is rate-limiting for assembly of the TCR in at least some clones of the polyclonal population. Another model postulates that the average TCR level for each clone is defined based on the strength of signals that are received during T cell development in the thymus, upon interaction with peptide-MHC presented by APCs, and afterwards maintained once development is completed. Such mechanism could operate independently or coupled with the steps involving selection (namely, positive and negative selection) during T cell development. As it is thought that the ζ chain is the rate-limiting sub-unit for assembly of the TCR (Baniyash, 2004), if this indeed holds for all clones, such a model would act by defining the average level of this chain in each clone.

In contrast to the TCR, there is evidence that the levels of some molecules are continuously modulated even in naive cells, depending on signals that are received, such as the strength of interaction between the TCR and self peptide-MHC, and other signals such as IL-7. One of the molecules for which there is evidence of continuous modulation of expression levels is CD5, which is expressed on the surface of T cells (and also B cells), and has been described as a negative regulator of TCR-derived signals (Tarakhovsky et al., 1995). Interestingly, there is evidence that differences between CD5 levels of two subsets of cells may be maintained, in studies analyzing $CD4^+$ or $CD8^+$ T cells (Azzam et al., 1998; Smith et al., 2001; Palmer et al., 2011; Mandl et al., 2013). An early work related the expression levels of CD5 in both thymocytes and mature T cells and the strength of signals derived from self peptide-MHC presented by APCs, by comparing TCR-transgenic strains in various MHC genetic backgrounds (Azzam et al., 1998). That work suggested a model in which CD5 levels would be defined during positive selection based on TCR specificity (Azzam et al., 1998). Furthermore, a subsequent work showed that there is also ongoing modulation of the CD5 levels depending on the strength of signals continuously received from self peptide-MHC, on top of the basal values that would be defined during T cell development (Smith et al., 2001). On this issue, upon separate transfer of CD5 high and low expressors from a naive polyclonal $CD4^+$ population to recipients lacking both MHC-I and MHC-II, it was shown that the differences between high and low expressors persisted, albeit they were compared only up to 9–12 days after transfer (Smith et al., 2001). Also, Palmer et al. (2011) recently showed that the differences between the mean CD5 levels of two $CD8^+$ TCR-transgenic populations (*Rag2*$^{-/-}$) are maintained *in vitro*, without any stimulation, for up to 5 days, and after 1 week of stimulation via the TCR and with cytokines. However, in this case no detailed quantitative analysis was done.

Altogether, the data from Smith et al. (2001) and Palmer et al. (2011) are consistent with a cell-intrinsic stable component of variation in expression levels of CD5, resulting in persistent differences between CD5 expression levels of different TCR-transgenic strains (based on Palmer et al., 2011), and among the clones of a polyclonal population (based on Smith et al., 2001). While genetic diversity could explain the stable component in the case of the TCR, the exact same reasoning does not apply to CD5, since it is not known to undergo a process of somatic rearrangement. In order to explain the cell-intrinsic stable component of variation in CD5, two non-exclusive models can be considered, having an intrinsic relationship to those previously discussed in the context of the TCR. First, that there is a physical association between CD5 and TCR, which has indeed been reported by early studies based on co-immunoprecipitation experiments in human and rat T cells (Burgess et al., 1992; Osman et al., 1992; Beyers et al., 1992). This model would imply a positive correlation between the levels of CD5 and TCR in a cell population, since the cells having higher TCR levels would be expected to have higher CD5 levels. However, data from studies reporting TCR and CD5 levels of different T cell populations in various cases are discordant, with some studies showing a negative correlation (Kieper et al., 2004), no clear relationship (Kassiotis et al., 2003), or a positive correlation between the levels of the TCR and CD5 (Palmer et al., 2011; Mandl et al., 2013). This may be related to the partial dependence of CD5 levels on the interaction with self peptide-MHC, since these studies have analyzed this relationship in fresh *ex vivo* cells. The second model is that the cell-intrinsic CD5 levels describing each clone would be defined during T cell development, as proposed by Azzam et al. (1998). This substantiates the basis for the previously discussed model postulating the definition of average TCR levels based on the strength of signals at this same time, which would be consistent with the view of the ζ chain as the rate-limiting sub-unit for assembly of the TCR (Baniyash, 2004).

Nevertheless, independently of the model explaining the stable component of TCR levels in the polyclonal population, the aforementioned data on CD5 (Smith et al., 2001; Palmer et al., 2011), together with the data in chapter 3, suggest that stable variants, in terms of the expression levels of various molecules, may be a widespread phenomenon in T cell populations. However, it remains unclear whether this reflects a common dependence on the stable component of variation in TCR levels, such as physical association between a particular molecule and the TCR, or independent pathways giving rise to sub-populations in terms of the expression levels of each molecule. In associating each sub-population to a stable genome-wide pattern of expression of transcription factors and various chromatin modifications, the ongoing advancement of genome-wide analysis, starting to allow the

analysis of the transcriptome at the single-cell level (Moignard et al., 2013; Shalek et al., 2013), may shed some light on the possible underlying molecular mechanism in the future.

This discussion highlights the difference between the two conditions in which the polyclonal population was analyzed, which were *in vitro* (section 3.4), and *in vivo* (section 3.5). While the former centers on cell-intrinsic stable variants, the latter includes both cell-intrinsic and cell-extrinsic variants. In the *in vivo* setting based on the adoptive transfer to *Rag2*$^{-/-}$ recipients, we showed that the differences in expression levels between high and low expressors in the polyclonal population persisted on a time frame of several weeks, consistent with the notion of the stable component of variation in TCR levels, and that the differences were also heritable, in light of extensive cell division. Indeed, it is possible that, as a consequence of lymphopenic *in vivo* environment of the recipient mice, the activation of the transferred cells adds up a cell-extrinsic effect to the stable component, given the trend of the values of function $\Delta_{H,L}(t)$ being greater than the initial value $\Delta_{H,L}(0)$. One way of addressing whether there is still a cell-intrinsic stable component would be to re-isolate the cells that were initially transferred to the recipients, and compare whether the different TCR expression levels of high and low expressors are maintained when cells are maintained *in vitro*.

The preliminary evidence for the stable component in TCR-transgenic populations, albeit not being the main contribution to the variation, may be a general property of these cells in the absence of cell division, and given the possible perturbation due to the cell cycle (see section 4.1), may not be heritable. When it comes to the TCR-transgenic populations, the mechanism based on the specification of expression levels can explain the stable component in such a population, provided that there is variation in the signals received by such an isogenic population during development in the thymus. However, this putative mechanism must be considered in light of the particular features of these populations. First, the dependence of expression of the α and β chains on transgenes, rather than on the endogenous TCRα and TCRβ loci. Second, the altered T cell development that takes place in many TCR-transgenic mice (see, for example, Baldwin et al., 2005), since the α chain tends to be expressed very early on depending on the transgene used, disturbing the order of events that take place during thymic development. Although not assessed in this work, it is presumed that this takes place in the two TCR-transgenic *Rag2*$^{-/-}$ populations used (Marilyn and OT-II), based on the particular vectors used to drive expression of the α chain (Barnden et al., 1998; Lantz et al., 2000). An additional concern is that the population is overloaded with cells all having the same TCR specificity, and there is evidence that the very high competition for access to ligands leads to extremely inefficient thymic development (Wong

et al., 2000; Canelles et al., 2003).

Given the relative contribution of the stable component (R_α^2) and the characteristic time of the variation (τ_T) of the expression levels of a certain molecule in a cell population, one question that arises is how stimulation modulates these two properties. In this context, the TCR may be a good model system, since much is known about its regulation and modulation upon stimulation to induce T cell activation (Schrum et al., 2003). Such stimulation is often achieved via the peptide to which the given isogenic T cell population is specific to, or high doses of immobilized anti-TCR antibodies which stimulate all T cells (Kruisbeek et al., 2004). Intriguingly, the events involved in T cell activation operate on quite distinct time frames. On the first few hours, there is internalization and degradation of the TCRs that have been engaged (triggered) by the stimulus, in a process referred to as TCR down-regulation (Valitutti et al., 1997). This continues to take place for as long as the stimulus remains, such that TCRs that were assembled and exported after the commencement of stimulation can be engaged, down-regulated and contribute to the signaling response (Schrum et al., 2000). It is thought that TCR down-regulation protects T cells from excessive stimulation, which could lead to activation-induced cell death (Schrum et al., 2003). By decreasing the effective value of parameter β, related to the mean lifetime of the TCR, TCR down-regulation is expected to decrease the characteristic time τ_T. In the subsequent hours to days, there is an up-regulation of TCR expression levels, which contributes to replenish the TCRs that were down-regulated and to further modulate the ongoing signaling response (Schrum et al., 2000), followed by cell division, typically after at least 48 hours since stimulation started (Gett and Hodgkin, 2000). By modulating the average rate of protein production, the up-regulation of TCR levels may impinge on R_α^2. Cell division may also influence R_α^2, by inducing "turnover" of the particular pattern of expression, as previously discussed (section 4.1). If strong stimulation progresses for much longer, induction of activation-induced cell death (AICD) becomes pronounced (Jelley-Gibbs et al., 2005; Schrum et al., 2005). Moreover, T cell differentiation, depending on additional factors (typically cytokines) provided, takes place after several days (Iezzi et al., 1999). This also has the potential to modulate R_α^2, by giving rise to differentiated T cell types, such as Th1 and Th2 (Liew, 2002), that could be associated with distinct TCR levels. However, the experimental analysis under such conditions of explicit stimulation should be addressed with proper caution in the context of stimulation, given the need of staining for the TCR in order to isolate high and low expressors, as done in chapter 3. Although not having a pronounced impact on TCR levels, at least in the *in vitro* setup used in section 3.4, the antibody used for sorting may interfere with the first instants of TCR-derived stimulation.

If so, analysis would need to be limited to later time-points, in which TCR down-regulation and cell division would be expected to essentially "dilute-out" the labeled receptors.

Interestingly, the process of T cell activation via the TCR is modulated by the intensity and/or the duration of the stimulus, as reflected by changes in the different processes that are induced, ranging from the number of TCRs that are down-regulated, the timing and number of cell divisions (Itoh and Germain, 1997; Iezzi et al., 1998; Schrum et al., 2003; Jelley-Gibbs et al., 2005; Schrum et al., 2005). In particular, stronger stimulation (greater doses) induce greater down-regulation (Valitutti et al., 1997; Schrum et al., 2000), such that, during the process of stimulation, surface TCR levels reflect a balance between production and degradation, the latter in the form of down-regulation and basal degradation (Schrum et al., 2003).

In the context of the response of a heterogenous population to stimulation, the theoretical framework developed in this thesis may provide an interesting perspective. Consider a molecule whose expression level influences a particular cellular response, such as activation, survival or proliferation, and that it takes a certain amount of time for this response to be induced once that molecule starts acting. Furthermore, neglect variation in any other property of the cells. If the expression level of the molecule fluctuates with relatively fast dynamics, compared with the time for the response to be induced, then it is unlikely that cells initially expressing different amounts of the molecule will vary in their responses. This is so because the pathway downstream of the molecule would effectively average out the fluctuations in the expression level of the molecule. Therefore, one requirement for two subsets of cells having different expression levels to respond distinctly to the same stimulus would be that the differences in the expression levels of the two subsets are permanent or relatively long-lasting.

At present the functional consequences or correlates of the variation in expression levels of the TCR remain unclear. Given the evidence for the stable component in the polyclonal population, the permanent differences in TCR expression levels may lead to variation in the responses of different clones under some conditions. There are, however, two main points that argue against such possibility. First, it may be that the differences in TCR levels, quantified in his work as being less than 2-fold when comparing high and low expressors in the various populations considered, are too small to impact on the responses. In such case, one could speculate that, in the polyclonal population, in which the total variation (σ^2_T) is larger, such variation could reflect mainly a neutral by-product of the need to assemble the TCR out of sub-units with diverse sequences. The second is based on the inference of genetic variation being an important factor explaining the variation in expression levels

of the TCR in the polyclonal population. It implies that each particular TCR clonotype, defined by its particular α and β chains, would be inherently associated with a particular expression level. This association may limit how the variation in expression levels impacts on the response of the cells. This limit would hold in particular under conditions in which the particular clonotype is the main determining factor for activation to take place, being relatively insensitive to the associated expression level. For example, it may be that the levels expressed are not limiting. One exception for this second point may be conditions in which a polyclonal population is stimulated in a manner that is independent of the actual TCR specificity, such as via immobilized anti-TCR antibodies, a commonly used *in vitro* technique (Kruisbeek et al., 2004). Under these conditions, it may be that high and low expressors for the TCR respond differently, especially in the case of strong stimulation, in which TCR levels would be limiting due to the extensive down-regulation.

As discussed in chapter 3, the equivalent ability of high and low expressors to reconstitute the peripheral pool of *Rag2*$^{-/-}$ recipients perhaps reflects particular underlying features of this *in vivo* setup. One possible explanation is that, under these conditions, TCR levels are not limiting, such that all cells receive essentially the same TCR-derived signals, and the ability to reconstitute the peripheral pool in this setup is mainly determined by other factors. However, even if the variation in TCR levels results in different signals being received, the particular pattern of variation in expression of downstream molecules may result in these differences being eliminated. For example, this would happen if high expressors for the TCR, even though receiving stronger TCR-derived signals when compared with low expressors, express lower levels of a kinase in comparison with the low expressors. If the levels of this kinase are important in determining the activation of a downstream pathway that results in a particular response, the negative correlation between the levels of the TCR and the kinase would have the overall effect of canceling differences in signaling. Indeed, Feinerman et al. (2008) argued for such a mechanism, in a study of the impact of variation on the expression levels of two molecules on the response of an important pathway in T cell activation, using a CD8$^+$ TCR-transgenic (*Rag2*$^{-/-}$) population.

While these points argue against the possibility of functional consequences of the variation in TCR expression levels, some studies have suggested or pointed to mechanisms that suggest consequences. There is evidence that additional signals provided to T cells during stimulation can modulate TCR levels on a fine scale, and also that the strength of the stimulus determines the fraction of the total TCRs that are engaged (Schrum et al., 2000). Based on these two aspects, a comparison between two T cell populations, having different average TCR levels due to initial stimulation in the presence of different signals, revealed

that the one having more TCRs would down-regulate a greater number, with this number being correlated with the magnitude of the signal transduced (Schrum et al., 2000). A similar point had been made in an early work using human T cell clones, which argued for the number of TCR triggered upon stimulation as a threshold for the acquisition of certain effector functions (Viola and Lanzavecchia, 1996). Some of the clones in that study (Viola and Lanzavecchia, 1996) had multiple β chains, and multiple TCRs clonotypes, each expressed in different amounts when compared among clones. Indeed, cells that have two functional α and/or β chains may have multiple $\alpha\beta$ pairings, and hence multiple TCRs clonotypes. They may be obtained, for instance, by crossing two TCR-transgenic mouse strains. The levels of each TCR clonotype in the cells expressing many of them may be lower than the levels in cells expressing only one clonotype, for example in the case that the total number of TCRs does not vary considerably in these two groups of cells. Indeed, such reduced levels of each of multiple clonotypes have been reported by some studies, which also found evidence for a correlation between the expression level of a particular clonotype and the response elicited upon stimulation of that clonotype (Dave et al., 1999; Legrand and Freitas, 2001a,b; Schrum and Turka, 2002). There are, however, two potentially complicating factors in these data. First, some particular $\alpha\beta$ pairs tend to be expressed at higher levels than other pairs in the same cells. This results in levels of a particular TCR clonotype that are much lower (for example, 10-fold; Dave et al., 1999) than the levels in cells having that particular clonotype alone, such that the differences that are necessary in order for functional consequences to be observed may be relatively large. Second, the fact that, upon stimulation with a given peptide, expected to engage only a given TCR clonotype, that peptide may interact with some of the additional clonotypes present in cells and provide inhibitory signals (similar to what is known as TCR antagonism; see, for example, Yang and Grey, 2003).

For other molecules in T cells, some studies have reported functional correlates of the variation in expression levels. Differential response of cells correlated with differences in CD5 levels have been reported under various conditions, such as *in vitro* responses of $CD8^+$ T cells to cytokines (Cho et al., 2010), and proliferation and/or other functional responses upon adoptive transfer of $CD4^+$ or $CD8^+$ T cells to lymphoreplete or lymphopenic recipients (Smith et al., 2001; Palmer et al., 2011; Mandl et al., 2013). Focusing on the expression levels of CD127 (the α chain of the IL-7 receptor) and using single-positive $CD4^+$ thymocytes, Sinclair et al. (2011) reported that, upon transfer of CD127 high and low expressors separately to unmanipulated wild-type recipients, high expressors were recovered in higher numbers after 2 weeks. In this case, it was argued that at this point both high

and low expressors were broadly similar to the recipient-derived cells in terms of CD127 levels, which was interpreted as indicating that the low expressors recovered were either those with initially the highest levels or were those that up-regulated CD127 sufficiently quickly and survived.

Therefore, while in the case of the TCR the functional consequences or correlates remain unclear, the results of the previously mentioned studies (Smith et al., 2001; Cho et al., 2010; Palmer et al., 2011; Sinclair et al., 2011; Mandl et al., 2013) indicate that the variation in expression levels of a molecule may have an impact. This may depend on the particular molecule, and also on the conditions in which the response of the cells is studied. Altogether, the results of this thesis allow for framing the question of functional impact in a much more general way, which will be the subject of the next section.

4.3 An integrated view of the expression level as a time-varying quantitative trait

An important theme in systems biology is the quantitative analysis of gene expression, and the works on noise in gene expression (Raj and van Oudenaarden, 2008) have studied extensively the stochastic effects on the fluctuations in expression levels in single cells and on the variation observed in a snapshot of a cell population. In relating the fluctuations in expression levels in single cells and the variation that is observed in a snapshot of the population, an implicit assumption is that every cell may attain any of the levels observed in the population. This has been related (Brock et al., 2009; Huang, 2009; Garcia-Ojalvo and Martinez Arias, 2012; MacArthur and Lemischka, 2013) to the so-called ergodic hypothesis in physics (see, for example, Patrascioiu, 1987; von Plato, 1991), such that time averages of an observable property of a cell population would be equivalent to an ensemble average of this property in a given time instant. However, if the cell population is composed of a set of stable variants (Chang et al., 2008), groups of cells that are biased, by whichever mechanism, to have a limited range of expression levels, compared with those observed in the population, this assumption is violated. In an extreme scenario, each cell would be described as a unique, stable variant, with expression level constant throughout time. Rather than constant, it may be that the expression levels in single cells fluctuate with very slow dynamics, compared with the duration of the time interval of analysis. In this case, each cell will not assume all the levels observed in the population, thereby also complicating the relationship between the dynamics of single cells and a snapshot of the population, which is often discussed in terms of invalidating the assumption of ergodicity, in reference to the

ergodic hypothesis (Brock et al., 2009; Huang, 2009; Rocco et al., 2013).

As an important technical aspect, it should be emphasized that ergodicity refers to a particular property (an "observable") of a stochastic process or a dynamical system. For example, a stochastic process that is ergodic with respect to the mean is one in which the time average of values in a sufficiently long time window is equal to the ensemble average of values (Papoulis and Pillai, 2002). This may also be referred to as a mean-ergodic stochastic process. Therefore, if a cell population is described using a general stochastic process, with each cell being a realization of the process, a population is said to be mean-ergodic if the time average, over a sufficiently long window of time, of every cell is equal to the average expression level of the population. A stochastic process may also be variance-ergodic, in that the variance of the values throughout time is equal to the variance of a snapshot, and also in terms of the probability density function, or simply density for short (Papoulis and Pillai, 2002). In the latter case, the density of the values throughout time is identical to the density of the values in a snapshot (density-ergodic). In the following, we will adopt this distinction when referring to the ergodicity of a cell population.

The theoretical framework of chapter 2 was initially framed in terms of relating the mean and variance of expression levels of a cell population, referred to as full population, to the properties of sub-populations, which are groups of cells that compose that population. A sub-population is a particular parameterization of the general stochastic process describing the full population, and is therefore by definition ergodic with respect to the density (and, consequently, also with the mean and variance). In the scope of the particular model of protein expression adopted, the full population is also density-ergodic if $R_\alpha^2 = 0$, since in this case it is equivalent to a single sub-population. For $R_\alpha^2 = 100\%$, the discrepancy between the variance of (log-transformed) expression levels of each cell throughout time ($\sigma_W^2 = 0$), and that of the full population (σ_T^2), is maximal, and σ_W^2 approaches σ_T^2 as R_α^2 decreases. In this sense, R_α^2 can be thought of as quantifying the degree to which the full population is ergodic with respect to the variance, with the characteristic time of the variation τ_T defining a time interval $t^* > 3\,\tau_T$ that is related to the minimum time necessary for each single cell to attain essentially all expression levels it can. In the setup of isolating high and low expressors to estimate R_α^2, the degree to which the starting population is ergodic with respect to the variance is hence related to the degree to which the means of log-transformed values of high and low expressors become similar after an amount of time t^*. Furthermore, if the duration of analysis $t^\dagger$ is shorter than $3\,\tau_T$, the relationship $\Omega_{H,L}(t^\dagger)$ (section 2.5.2) quantifies the degree to which the full population is variance-ergodic within this duration of analysis. Similarly, in a setup of time-lapse imaging, Sigal et al. (2006)

had previously devised a measure of the degree to which a cell attains all expression levels observed in the population (see also Cohen et al., 2009), and discussed that it could be considered as a measure of ergodicity. By focusing on the properties of populations that have been isolated, such as high and low expressors, the quantification of ergodicity with respect to the variance based on R_α^2 would be expected to average the data on the lineages in the isolated populations. However it provides for a more general approach for analysis, as it does not require a setup of time-lapse imaging in which individual cells and their lineages are tracked throughout time.

The relationship between the duration of the observation and R_α^2 emerged in the analysis of the *in vitro* experimental data in chapter 3, comparing the polyclonal and the two TCR-transgenic T cell populations (section 3.4). While unable to distinguish whether the T cell populations can be better described by different values of R_α^2 or τ_T, due to the limited window of observation, it was argued that a description in terms of different values of R_α^2, but equal τ_T for all populations was the most appropriate, by constraining the value estimated for τ_T to the duration of the observation. In this experimental setup, with the limited time for observation, we concluded that all three T cell populations could be described as being composed of a set of stable variants, and, in other words, are populations that are non-ergodic with respect to the variance, at least up to the duration of the analysis.

Originally framed in terms of the quantification of properties of the expression levels in a cell population, the theoretical framework developed in chapter 2 also has parallels with quantitative genetics, and the following presentation is based on Falconer and Mackay (1996); Lynch and Walsh (1998). Quantitative genetics extends the principles of Mendelian inheritance to continuous-valued traits, that are often influenced by the segregation of genes at many loci, such that the inheritance of quantitative differences is a consequence of the transmission of genetic information along generations. The phenotypic value of a particular trait in an individual is described as the sum of a genotypic value, which represents the impact of genes, and an environmental deviation, quantifying the influence of non-genetic causes, which are overall lumped into the concept of "environment". The genotypic value and environmental deviation are often assumed to follow normal distributions. Based on the pioneering work of Fisher in his 1918 paper (Fisher, 1918), the variance of phenotypic values of the individuals, termed V_P, is decomposed into a genetic variance V_G and an environmental variance V_E. Since it is the case most often studied in quantitative genetics, the presentation hereafter assumes an infinitely large, diploid and sexually-reproducing population with random mating. Afterwards, the view of a cell population in such context will be discussed.

The genetic variance is often further decomposed into an additive (V_A) and a non-additive genetic variance (V_N). Although not a part of genetic theory, strictly speaking, the environmental variance can also be further partitioned into two general categories, designated as special environmental variance (V_{Es}) and general environmental variance (V_{Eg}), whose interpretations depend on the character of interest (Falconer and Mackay, 1996; Lynch and Walsh, 1998). Therefore, the phenotypic variance V_P can be written as:

$$V_P = V_G + V_E = \underbrace{\underbrace{\underbrace{V_A}_{h^2\,V_P} + V_N}_{H^2\,V_P} + V_{Eg}}_{r\,V_P} + V_{Es} \tag{4.1}$$

which neglects the joint genetic-environmental contribution (often referred to as gene-environment interactions).

The genetic variance describes the population, and depends on the properties of the alleles that influence the trait. In an individual, the presence of a particular allele of a gene has the effect of altering the genotypic value of that individual, relative to the mean genotypic value of the population. The two terms decomposing the genetic variance, namely V_A and V_N, are related, respectively, to the so-called additive and non-additive effects of the alleles. The additive effect of an allele is the expected effect of that allele averaged over the effect observed in the combination with all other alleles. In turn, a non-additive effect reflects a deviation between the effect of the particular combination of alleles present in an individual and the sum of the additive effects of these alleles. This deviation occurs due to statistical interactions between alleles of the same gene (dominance) and among alleles of different genes (epistasis). The fundamental distinction between additive (V_A) and non-additive genetic variance (V_N) arises due to the segregation of alleles during the formation of gametes, since in a diploid population that reproduces sexually each parent contributes with only one allele for each autosomal locus of its offspring, and consequently non-additive effects are not transmitted to the next generation.

When a character is measured multiple times in each individual, one may partition V_P so as to reflect the variance within and among individuals. One simplified scenario is the assessment of milk-yield in dairy cattle, as a character measured several times during the life of an individual. In this case, the special environmental variance V_{Es} is defined as the variance of the trait values within individuals throughout time. The variance among individuals is then $V_A + V_N + V_{Eg}$, such that the general environmental variance V_{Eg} is the

variance among individuals that is due to non-genetic origins. Therefore, for a trait whose value fluctuates in an individual throughout time, quantitative genetics relies on the partitioning of the phenotypic variance that results in three parameters, relating the proportion of the phenotypic variance that is due to variation among individuals (repeatability of the trait, r), due to the total genetic variance (broad-sense heritability, H^2), and that is due to the additive genetic variance (narrow-sense heritability h^2). These three parameters can be organized as shown in equation 4.1, reflecting the property that $h^2 \leq H^2 \leq r$.

Focusing on the analysis of the expression levels in a population of cells via flow cytometry, this thesis bridges the concept of ergodicity in physics, the quantification of expression levels in cell populations (including noise in gene expression), and the study of continuous-valued traits in the context of quantitative genetics. In the following, we present an integrated discussion of the present work in the context of these subjects.

Interestingly, in both quantitative genetics (Falconer and Mackay, 1996; Lynch and Walsh, 1998) and systems biology, in particular works focused on noise in gene expression (Raj and van Oudenaarden, 2008), approaches for decomposing the variance, or another measure of variation, have been instrumental in studying the properties of different biological systems. In the study of noise in gene expression, the notion of intrinsic and extrinsic noise put forward by Elowitz and colleagues (Elowitz et al., 2002; Swain et al., 2002), is based on decomposing the coefficient of variation of expression levels into these two noise sources. Other works have focused on either generalizing this distinction or developing new decompositions (Rausenberger and Kollmann, 2008; Hilfinger and Paulsson, 2011; Rinott et al., 2011; Komorowski et al., 2013). It may be possible to combine these approaches with the framework developed in this thesis, to further partition the unstable component, for example, into intrinsic and extrinsic noise.

Considering expression levels, typically assessed via mRNA levels, as quantitative traits, advances in genomics have provided the possibility of addressing the genetic contributions to variation in gene expression, in terms of so-called expression quantitative trait loci (eQTL) (Rockman and Kruglyak, 2006). In this case, the objective is to determine, or map, and estimate the effects of the specific loci underlying variation in expression levels among different individuals of a population. In this case, the specific quantitative trait of interest is the average expression level in each biological sample being analyzed, given that cells and/or multicellular organisms are typically pooled in each biological sample. The possibility of assessing thousands of quantitative traits (transcripts) provides with the potential to identify several loci contributing to differences in average transcript abundance. Recent works have sought to extend this to study properties of expression levels that re-

quire analysis of single cells (Wills et al., 2013), and how genetic variation influences the response of pathways to stimulation (Gat-Viks et al., 2013). It is tempting to consider mapping the molecular basis for the stable component in terms of genome-wide pattern of expression in single cells (see section 4.1), much like a QTL-based approach proposed to quantify the impact of alternative chromatin states to phenotypic variation (Johannes et al., 2008).

One of the key results of this thesis, parameter R^2_α, which quantifies the fraction of the total variance of expression levels that is due to the stable component, representing differences that persist throughout time, is related, at least conceptually, to the concepts of heritability and repeatability. In the analogy with quantitative genetics, each cell is taken as an individual in a population, such that theoretical framework deals with a population of individuals that reproduces asexually. In this sense, for a cell population analyzed in the absence of cell division, R^2_α would be analogous to the repeatability r of the expression levels in single cells, while, if there is cell division, R^2_α would be akin to a notion of heritability, but now extended to a sufficiently long amount of time. Implicit in this comparison is that the notion of time in quantitative genetics is present in terms of the generation time, given the transmission of genetic information to the progeny, and throughout the lifetime of the organism.

Moreover, in quantitative genetics, the variance components (equation 4.1) and the proportions r, H^2 and h^2 are often calculated based on the results of an ANOVA (analysis of variance) or by regression (Sokal and Rohlf, 1981; Lessells and Boag, 1987; Lynch and Walsh, 1998), or alternatively by more sophisticated approaches (for example, Alvarez-Castro, 2012). For example, the narrow-sense heritability is commonly estimated via a regression of phenotypic values of parent and offspring (Lynch and Walsh, 1998). Such approaches have in common the requirement of information on the association between the offspring and their parents. In the context of the quantification of expression levels in cell populations, this is equivalent to a setup of time-lapse imaging, of single cells being directly tracked throughout time, and then relating the properties of the daughter cells to those of the mother cell. In contrast, the present work focused on the analysis of snapshots of a population, such as via flow cytometry, where such information is not available. The approach of the present work more closely resembles the application of quantitative genetics in the context of animal breeding, in which a few individuals of the population having traits with values that are considered most adequate for the application of interest are selected to reproduce. In particular, the narrow-sense heritability, quantifying the proportion of the variance that is passed on to the offspring, is a key predictor of the short-term

response of a population to strong selection acting on a trait, as represented in the so-called breeder's equation:

$$R = h^2 S \tag{4.2}$$

where S is the selection differential, given by the change in the mean trait value by selecting the parents, in comparison with the original population, and R is the response, quantifying the change in the mean trait value of the progeny of the selected parents, relative to the original population.

In this view, a distinguishing property of expression levels, as emphasized by recent works on mammalian cell populations (Chang et al., 2008; Kalmar et al., 2009; Luo et al., 2012; Sisan et al., 2012), is that the characteristic time of the variation (τ_T), may be much longer than the generation time of the cell populations considered. Therefore, two subsets of genetically identical cells may differ in their expression levels for a prolonged duration of time, much like they would do if the two cells were genetically different. This highlights the expression level of the respective molecules as a potentially good model system in the context of transgenerational inheritance, given the potential non-ergodicity in terms of the mean and variance. In analogy with the breeder's equation, it would be expected that function $\Omega_{H,L}(t^\star)$, where $t^\star$ is the generation time of the population of interest, provides an estimate of the short-term ability of a cell population to respond to selection acting on the expression levels, for example in the context of a molecule whose expression levels affect the rate of cell division.

4.4 Perspectives on the study of variation in expression levels

This thesis sought to address the contributions to variation in expression levels in cell populations. By considering the factors that relate the fluctuations in expression levels in a single cell throughout time, and a snapshot of the population, we aimed at formalizing this relationship, deriving an approach for quantification, and addressing experimentally how the expression levels in a cell population are shaped.

The theoretical framework of chapter 2 started from a general formulation, but ultimately a relatively simple scenario to describe expression levels in a cell population was considered and studied in detail. For example, for simplicity, it was considered that all subpopulations have the same variance of log-transformed protein levels. This allowed one to formally refer to the unstable component as a single entity in the resultant full population, since all sub-populations and cells have the same variance. In this way, all the differences

among sub-populations are lumped into the term of variance of log-transformed average rates of protein expression. In this setting, inference of R^2_α emerged as a relatively simple procedure. One may investigate inference in the more general case in which this assumption is relaxed, to account for the possibility that each sub-population has a different variance, and to assess the possibility of quantifying additional properties of the population.

In the context of cell division and the processes preceding it (such as DNA replication), the likely resulting perturbations to the global pattern of genome-wide gene expression (Egli et al., 2008; Alabert and Groth, 2012) established in a single cell may shed new light on the dynamics of expression levels in a cell population. In particular, asymmetric cell division, whereby the two resultant daughter cells are markedly distinct in their properties, such as protein amounts (Neumüller and Knoblich, 2009), may lead to a considerable increase in the variation in expression levels, and, for the case of regulatory molecules, provide a large perturbation in the expression of the downstream genes. Indeed, recent reports suggest that under some conditions T cells may asymmetrically segregate some molecules (Chang et al., 2007, 2011; King et al., 2012). Therefore, given the potential for asymmetric cell division to further shape the expression levels in a cell population, a detailed study of cell division could provide more details into how it impacts on the dynamics of expression levels in a cell population.

The integrated view of the expression level as a quantitative trait provided by this work provides a framework that may be used to address the impact of variation in expression levels on the response of cell populations. Theoretical studies have shown that stochastic variation, corresponding to what we referred to as the unstable component, may contribute to the adaptive response of a cell population (Paixão, 2007; Tanase-Nicola and ten Wolde, 2008). Furthermore, in an experimental evolution experiment, it has been shown that protein expression levels can evolve (Dekel and Alon, 2005). In this context, we believe that the formulation of the expression level as a quantitative trait can be very fruitful, by reinforcing the connection with quantitative genetics, in which case the view of the impact of variation on expression levels is cast in terms of the evolution and selection of a quantitative trait. Furthermore, in keeping in mind the possibility that the characteristic time of the variation (τ_T) may be comparable or even larger than the generation time (in between cell divisions), such that there is a transgenerational memory of the expression levels, this may provide a good model system in which to study the impact of such property on evolution.

Immunology is naturally concerned with the response of heterogeneous cell populations, with one of the cornerstones being the clonal selection theory, proposed by Burnet in the 1950's (Burnet, 1959). By casting the immune response in a Darwinian setting, based

on the selection of clones that are capable of responding to an antigen, one of the key points of the theory was that cells specific for the antigen would be present before encountering it. The theory therefore posited that there would be standing variation in a cell population, such that the antigen would to selection for a clone capable of recognizing it, rather than instructing the generation of such clone. In contrast, we can consider the duality between instruction and selection in processes leading to cell differentiation. Indeed, this duality has been discussed by Szabo et al. (2003), in the context of Th1/Th2 differentiation, in terms of whether cytokines act by guiding cells in a precursor-like state to a specific differentiated fate (instruction), or by favoring a particular set of "variants" that were present *a priori* (selection). In this sense, one possible avenue of research highlighted by the work developed in this thesis may be studying how stimulation of a cell population modifies the parameters that regulate protein expression in that population, not only in terms of the mean and the amount of variation in expression levels, but the underlying properties such as the relative contribution of the stable component (R^2_α) and the characteristic time of the variation (τ_T). This may provide a new perspective on the immune response, in terms of the ability to establish stable patterns of expression of a particular molecule in different cells upon stimulation (Beuneu et al., 2010; Antebi et al., 2013; Fang et al., 2013).

The analysis of the response of cell populations to stimulation raises an additional point. In order to increase the average expression level of a molecule, one can consider two hypothetical mechanisms. First, simply increasing the average rate of production. Provided that none of the additional parameters β, τ and σ change, an overall expected impact would be that of simply changing the mean expression level of the population. A second mechanism would be the increase of the mean lifetime of the protein (β), for example by reducing the rate of turnover, so as to produce the same overall increase in the mean expression level of the population. However, increasing parameter β is expected to increase the characteristic time τ_T, and hence the expected amount of time over which transient differences in expression levels of two subsets of cells last. This raises the possibility of using properties of the dynamics as a way to compare different models for the response of the population. There is evidence that up-regulation of MHC-II levels in some APCs in response to certain stimuli takes place via such a mechanism, reducing the rate of protein turnover (Cella et al., 1997). This may also be an attractive model system to study the impact of variation in expression levels in the context of cell-cell interactions, given indication of the impact of MHC-II expression levels in the ability to stimulate and activate T cells upon presentation of peptides (Kuwano et al., 2007).

Therefore, the quantitative understanding of cellular regulatory and response pathways

will require considering not only static properties, such as mean and the variation in expression levels, but also underlying properties of the population that are not easily analyzed from a single snapshot, such as the degree to which every cell can attain all expression levels observed in a snapshot of the population, and the dynamics of the fluctuations in expression in single cells. This will provide a more faithful view of how the expression levels of a molecule are regulated on a cell population, and of its impact on the survival, proliferation and response of this population to stimulation. On this undertaking, the theoretical framework and the analysis of experimental data conducted in this thesis may provide an interesting perspective.

Bibliography

Alabert, C. and Groth, A. (2012). Chromatin replication and epigenome maintenance. *Nature Reviews Molecular Cell Biology*, 13(3):153–67.

Alvarez-Castro, J. M. (2012). Current applications of models of genetic effects with interactions across the genome. *Current Genomics*, 13(2):163–75.

Antebi, Y. E., Reich-Zeliger, S., Hart, Y., Mayo, A., Eizenberg, I., Rimer, J., Putheti, P., Pe'er, D., and Friedman, N. (2013). Mapping Differentiation under Mixed Culture Conditions Reveals a Tunable Continuum of T Cell Fates. *PLoS Biology*, 11(7):e1001616.

Azzam, H. S., Grinberg, A., Lui, K., Shen, H., Shores, E. W., and Love, P. E. (1998). CD5 expression is developmentally regulated by T cell receptor (TCR) signals and TCR avidity. *The Journal of Experimental Medicine*, 188(12):2301–11.

Baldwin, T. A., Sandau, M. M., Jameson, S. C., and Hogquist, K. A. (2005). The timing of TCR alpha expression critically influences T cell development and selection. *The Journal of Experimental Medicine*, 202(1):111–21.

Baniyash, M. (2004). TCR zeta-chain downregulation: curtailing an excessive inflammatory immune response. *Nature Reviews Immunology*, 4(9):675–87.

Barnden, M. J., Allison, J., Heath, W. R., and Carbone, F. R. (1998). Defective TCR expression in transgenic mice constructed using cDNA-based alpha- and beta-chain genes under the control of heterologous regulatory elements. *Immunology and Cell Biology*, 76(1):34–40.

Beuneu, H., Lemaître, F., Deguine, J., Moreau, H. D., Bouvier, I., Garcia, Z., Albert, M. L., and Bousso, P. (2010). Visualizing the functional diversification of CD8+ T cell responses in lymph nodes. *Immunity*, 33(3):412–23.

Beyers, A. D., Spruyt, L. L., and Williams, A. F. (1992). Molecular associations between the T-lymphocyte antigen receptor complex and the surface antigens CD2, CD4, or CD8 and CD5. *PNAS*, 89(7):2945–9.

Brock, A., Chang, H., and Huang, S. (2009). Non-genetic heterogeneity–a mutation-independent driving force for the somatic evolution of tumours. *Nature Reviews Genetics*, 10(5):336–42.

Burgess, K. E., Yamamoto, M., Prasad, K. V., and Rudd, C. E. (1992). CD5 acts as a tyrosine kinase substrate within a receptor complex comprising T-cell receptor zeta chain/CD3 and protein-tyrosine kinases p56lck and p59fyn. *PNAS*, 89(19):9311–5.

Burnet, F. M. (1959). *The Clonal Selection Theory of Acquired Immunity*. Cambridge University Press.

Canelles, M., Park, M. L., Schwartz, O. M., and Fowlkes, B. J. (2003). The influence of the thymic environment on the CD4-versus-CD8 T lineage decision. *Nature Immunology*, 4(8):756–64.

Cella, M., Engering, A., Pinet, V., Pieters, J., and Lanzavecchia, A. (1997). Inflammatory stimuli induce accumulation of MHC class II complexes on dendritic cells. *Nature*, 388(6644):782–7.

Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E., and Huang, S. (2008). Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, 453(7194):544–7.

Chang, J. T., Ciocca, M. L., Kinjyo, I., Palanivel, V. R., McClurkin, C. E., Dejong, C. S., Mooney, E. C., Kim, J. S., Steinel, N. C., Oliaro, J., Yin, C. C., Florea, B. I., Overkleeft, H. S., Berg, L. J., Russell, S. M., Koretzky, G. A., Jordan, M. S., and Reiner, S. L. (2011). Asymmetric proteasome segregation as a mechanism for unequal partitioning of the transcription factor T-bet during T lymphocyte division. *Immunity*, 34(4):492–504.

Chang, J. T., Palanivel, V. R., Kinjyo, I., Schambach, F., Intlekofer, A. M., Banerjee, A., Longworth, S. A., Vinup, K. E., Mrass, P., Oliaro, J., Killeen, N., Orange, J. S., Russell,

S. M., Weninger, W., and Reiner, S. L. (2007). Asymmetric T lymphocyte division in the initiation of adaptive immune responses. *Science*, 315(5819):1687–91.

Chen, F., Rowen, L., Hood, L., and Rothenberg, E. V. (2001). Differential transcriptional regulation of individual TCR V beta segments before gene rearrangement. *The Journal of Immunology*, 166(3):1771–80.

Cho, J.-H., Kim, H.-O., Surh, C. D., and Sprent, J. (2010). T cell receptor-dependent regulation of lipid rafts controls naive CD8+ T cell homeostasis. *Immunity*, 32(2):214–26.

Cohen, A. A., Kalisky, T., Mayo, A., Geva-Zatorsky, N., Danon, T., Issaeva, I., Kopito, R. B., Perzov, N., Milo, R., Sigal, A., and Alon, U. (2009). Protein dynamics in individual human cells: experiment and theory. *PLoS ONE*, 4(4):e4901.

Dave, V. P., Allman, D., Wiest, D. L., and Kappes, D. J. (1999). Limiting TCR expression leads to quantitative but not qualitative changes in thymic selection. *The Journal of Immunology*, 162(10):5764–74.

Dekel, E. and Alon, U. (2005). Optimality and evolutionary tuning of the expression level of a protein. *Nature*, 436(7050):588–92.

Dodd, I. B., Micheelsen, M. A., Sneppen, K., and Thon, G. (2007). Theoretical analysis of epigenetic cell memory by nucleosome modification. *Cell*, 129(4):813–22.

Egli, D., Birkhoff, G., and Eggan, K. (2008). Mediators of reprogramming: transcription factors and transitions through mitosis. *Nature Reviews Molecular Cell Biology*, 9(7):505–16.

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584):1183–6.

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. Benjamin Cummings, 4 edition.

Fang, M., Xie, H., Dougan, S. K., Ploegh, H., and van Oudenaarden, A. (2013). Stochastic Cytokine Expression Induces Mixed T Helper Cell States. *PLoS Biology*, 11(7):e1001618.

Feinerman, O., Veiga, J., Dorfman, J. R., Germain, R. N., and Altan-Bonnet, G. (2008). Variability and robustness in T cell activation from regulated heterogeneity in protein levels. *Science*, 321(5892):1081–4.

Ferrell, J. E. (2002). Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Current Opinion in Cell Biology*, 14(2):140–8.

Fisher, R. A. (1918). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433.

Garcia-Ojalvo, J. and Martinez Arias, A. (2012). Towards a statistical mechanics of cell fate decisions. *Current Opinion in Genetics & Development*, 22(6):619–26.

Gat-Viks, I., Chevrier, N., Wilentzik, R., Eisenhaure, T., Raychowdhury, R., Steuerman, Y., Shalek, A. K., Hacohen, N., Amit, I., and Regev, A. (2013). Deciphering molecular circuits from genetic variation underlying transcriptional responsiveness to stimuli. *Nature Biotechnology*, 31(4):342–349.

Gett, A. V. and Hodgkin, P. D. (2000). A cellular calculus for signal integration by T cells. *Nature Immunology*, 1(3):239–44.

Hilfinger, A. and Paulsson, J. (2011). Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *PNAS*, 108(29):12167–72.

Huang, S. (2009). Non-genetic heterogeneity of cells in development: more than just noise. *Development*, 136(23):3853–62.

Huang, S., Eichler, G., Bar-Yam, Y., and Ingber, D. E. (2005). Cell Fates as High-Dimensional Attractor States of a Complex Gene Regulatory Network. *Physical Review Letters*, 94(12):128701.

Hunt, D. F., Michel, H., Dickinson, T. A., Shabanowitz, J., Cox, A. L., Sakaguchi, K., Appella, E., Grey, H. M., and Sette, A. (1992). Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad. *Science*, 256(5065):1817–20.

Iezzi, G., Karjalainen, K., and Lanzavecchia, A. (1998). The duration of antigenic stimulation determines the fate of naive and effector T cells. *Immunity*, 8(1):89–95.

Iezzi, G., Scotet, E., Scheidegger, D., and Lanzavecchia, A. (1999). The interplay between the duration of TCR and cytokine signaling determines T cell polarization. *European Journal of Immunology*, 29(12):4092–101.

Itoh, Y. and Germain, R. N. (1997). Single cell analysis reveals regulated hierarchical T cell antigen receptor signaling thresholds and intraclonal heterogeneity for individual cytokine responses of CD4+ T cells. *The Journal of Experimental Medicine*, 186(5):757–66.

Jablonka, E. and Raz, G. (2009). Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *The Quarterly Review of Biology*, 84(2):131–76.

Jelley-Gibbs, D. M., Dibble, J. P., Filipson, S., Haynes, L., Kemp, R. A., and Swain, S. L. (2005). Repeated stimulation of CD4 effector T cells can limit their protective function. *The Journal of Experimental Medicine*, 201(7):1101–12.

Johannes, F., Colot, V., and Jansen, R. C. (2008). Epigenome dynamics: a quantitative genetics perspective. *Nature Reviews Genetics*, 9(11):883–90.

Kalmar, T., Lim, C., Hayward, P., Muñoz Descalzo, S., Nichols, J., Garcia-Ojalvo, J., and Martinez Arias, A. (2009). Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biology*, 7(7):e1000149.

Kassiotis, G., Zamoyska, R., and Stockinger, B. (2003). Involvement of avidity for major histocompatibility complex in homeostasis of naive and memory T cells. *The Journal of Experimental Medicine*, 197(8):1007–16.

Kieper, W. C., Burghardt, J. T., and Surh, C. D. (2004). A role for TCR affinity in regulating naive T cell homeostasis. *The Journal of Immunology*, 172(1):40–4.

King, C. G., Koehli, S., Hausmann, B., Schmaler, M., Zehn, D., and Palmer, E. (2012). T cell affinity regulates asymmetric division, effector cell differentiation, and tissue pathology. *Immunity*, 37(4):709–20.

Komorowski, M., Miekisz, J., and Stumpf, M. P. (2013). Decomposing Noise in Biochemical Signaling Systems Highlights the Role of Protein Degradation. *Biophysical Journal*, 104(8):1783–1793.

Kruisbeek, A. M., Shevach, E., and Thornton, A. M. (2004). Proliferative assays for T cell function. *Current Protocols In Immunology*, pages 3.12.1–3.12.20.

Kuwano, Y., Prazma, C. M., Yazawa, N., Watanabe, R., Ishiura, N., Kumanogoh, A., Okochi, H., Tamaki, K., Fujimoto, M., and Tedder, T. F. (2007). CD83 influences cell-surface MHC class II expression on B cells and other antigen-presenting cells. *International Immunology*, 19(8):977–92.

Lantz, O., Grandjean, I., Matzinger, P., and Di Santo, J. P. (2000). Gamma chain required for naïve CD4+ T cell survival but not for antigen proliferation. *Nature Immunology*, 1(1):54–8.

Legrand, N. and Freitas, A. A. (2001a). CD8+ T lymphocytes in double alpha beta TCR transgenic mice. I. TCR expression and thymus selection in the absence or in the presence of self-antigen. *The Journal of Immunology*, 167(11):6150–7.

Legrand, N. and Freitas, A. A. (2001b). CD8+ T lymphocytes in double alpha beta TCR transgenic mice. II. Competitive fitness of dual alpha beta TCR CD8+ T lymphocytes in the peripheral pools. *The Journal of Immunology*, 167(11):6158–64.

Lessells, C. M. and Boag, P. T. (1987). Unrepeatable Repeatabilities : A Common Mistake. *The Auk*, 104(1):116–121.

Liew, F. Y. (2002). T(H)1 and T(H)2 cells: a historical perspective. *Nature Reviews Immunology*, 2(1):55–60.

Luo, Y., Lim, C. L., Nichols, J., Martinez-Arias, A., and Wernisch, L. (2012). Cell signalling regulates dynamics of Nanog distribution in embryonic stem cell populations. *Journal of the Royal Society, Interface*, 10(78):20120525.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, 1 edition.

MacArthur, B. D. and Lemischka, I. R. (2013). Statistical Mechanics of Pluripotency. *Cell*, 154(3):484–489.

MacArthur, B. D., Ma'ayan, A., and Lemischka, I. R. (2009). Systems biology of stem cell fate and cellular reprogramming. *Nature Reviews Molecular Cell Biology*, 10(10):672–81.

Mandl, J. N., Monteiro, J. P., Vrisekoop, N., and Germain, R. N. (2013). T cell-positive selection uses self-ligand binding strength to optimize repertoire recognition of foreign antigens. *Immunity*, 38(2):263–74.

Market, E. and Papavasiliou, F. N. (2003). V(D)J recombination and the evolution of the adaptive immune system. *PLoS Biology*, 1(1):E16.

Moignard, V., Macaulay, I. C., Swiers, G., Buettner, F., Schütte, J., Calero-Nieto, F. J., Kinston, S., Joshi, A., Hannah, R., Theis, F. J., Jacobsen, S. E., de Bruijn, M. F., and Göttgens, B. (2013). Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nature Cell Biology*, 15(4):1–11.

Munsky, B., Trinh, B., and Khammash, M. (2009). Listening to the noise: random fluctuations reveal gene network parameters. *Molecular Systems Biology*, 5(318):318.

Neumüller, R. A. and Knoblich, J. A. (2009). Dividing cellular asymmetry: asymmetric cell division and its implications for stem cells and cancer. *Genes & Development*, 23(23):2675–99.

Osman, N., Ley, S. C., and Crumpton, M. J. (1992). Evidence for an association between the T cell receptor/CD3 antigen complex and the CD5 antigen in human T lymphocytes. *European journal of immunology*, 22(11):2995–3000.

Paixão, T. (2007). *The Stochastic Basis of Somatic Variation*. PhD thesis, University of Porto.

Paixão, T., Carvalho, T. P., Calado, D. P., and Carneiro, J. (2007). Quantitative insights into stochastic monoallelic expression of cytokine genes. *Immunology and Cell Biology*, 85(4):315–22.

Palmer, M. J., Mahajan, V. S., Chen, J., Irvine, D. J., and Lauffenburger, D. A. (2011). Signaling thresholds govern heterogeneity in IL-7-receptor-mediated responses of naïve CD8(+) T cells. *Immunology and Cell Biology*, 89(5):581–94.

Papoulis, A. and Pillai, S. U. (2002). *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 4 edition.

Patrascioiu, A. (1987). The Ergodic-Hypothesis, A Complicated Problem in Mathematics and Physics. *Los Alamos Science*, 15:263–279.

Pina, C., Fugazza, C., Tipping, A. J., Brown, J., Soneji, S., Teles, J., Peterson, C., and Enver, T. (2012). Inferring rules of lineage commitment in haematopoiesis. *Nature Cell Biology*, 14(3):287–94.

Raj, A. and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–26.

Raser, J. M. and O'Shea, E. K. (2004). Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678):1811–4.

Rausenberger, J. and Kollmann, M. (2008). Quantifying origins of cell-to-cell variations in gene expression. *Biophysical Journal*, 95(10):4523–8.

Rinott, R., Jaimovich, A., and Friedman, N. (2011). Exploring transcription regulation through cell-to-cell variability. *PNAS*, 108(15):6329–34.

Rocco, A., Kierzek, A. M., and McFadden, J. (2013). Slow protein fluctuations explain the emergence of growth phenotypes and persistence in clonal bacterial populations. *PLoS ONE*, 8(1):e54272.

Rockman, M. V. and Kruglyak, L. (2006). Genetics of global gene expression. *Nature Reviews Genetics*, 7(11):862–72.

Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S., and Elowitz, M. B. (2005). Gene regulation at the single-cell level. *Science*, 307(5717):1962–5.

Saito, T., Sussman, J. L., Ashwell, J. D., and Germain, R. N. (1989). Marked differences in the efficiency of expression of distinct alpha beta T cell receptor heterodimers. *The Journal of Immunology*, 143(10):3379–84.

Schrum, A. G., Palmer, E., and Turka, L. A. (2005). Distinct temporal programming of naive CD4+ T cells for cell division versus TCR-dependent death susceptibility by antigen-presenting macrophages. *European Journal of Immunology*, 35(2):449–59.

Schrum, A. G. and Turka, L. A. (2002). The Proliferative Capacity of Individual Naive CD4+T Cells Is Amplified by Prolonged T Cell Antigen Receptor Triggering. *The Journal of Experimental Medicine*, 196(6):793–803.

Schrum, A. G., Turka, L. A., and Palmer, E. (2003). Surface T-cell antigen receptor expression and availability for long-term antigenic signaling. *Immunological Reviews*, 196:7–24.

Schrum, A. G., Wells, A. D., and Turka, L. A. (2000). Enhanced surface TCR replenishment mediated by CD28 leads to greater TCR engagement during primary stimulation. *International Immunology*, 12(6):833–42.

Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., Trombetta, J. T., Gennert, D., Gnirke, A., Goren, A., Hacohen, N., Levin, J. Z., Park, H., and Regev, A. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, pages 1–5.

Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Liron, Y., Rosenfeld, N., Danon, T., Perzov, N., and Alon, U. (2006). Variability and memory of protein levels in human cells. *Nature*, 444(7119):643–6.

Sinclair, C., Saini, M., van der Loeff, I. S., Sakaguchi, S., and Seddon, B. (2011). The Long-Term Survival Potential of Mature T Lymphocytes Is Programmed During Development in the Thymus. *Science Signaling*, 4(199):ra77.

Sisan, D. R., Halter, M., Hubbard, J. B., and Plant, A. L. (2012). Predicting rates of cell state change caused by stochastic fluctuations using a data-driven landscape model. *PNAS*, 109(47):19262–7.

Smith, K., Seddon, B., Purbhoo, M. A., Zamoyska, R., Fisher, A. G., and Merkenschlager, M. (2001). Sensory adaptation in naive peripheral CD4 T cells. *The Journal of Experimental Medicine*, 194(9):1253–61.

Sokal, R. R. and Rohlf, F. J. (1981). *Biometry*. Freeman.

Swain, P. S., Elowitz, M. B., and Siggia, E. D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. *PNAS*, 99(20):12795–800.

Szabo, S. J., Sullivan, B. M., Peng, S. L., and Glimcher, L. H. (2003). Molecular mechanisms regulating Th1 immune responses. *Annual Review of Immunology*, 21:713–58.

Tanase-Nicola, S. and ten Wolde, P. R. (2008). Regulatory control and the costs and benefits of biochemical noise. *PLoS Computational Biology*, 4(8):e1000125.

Tarakhovsky, A., Kanner, S., Hombach, J., Ledbetter, J., Muller, W., Killeen, N., and Rajewsky, K. (1995). A role for CD5 in TCR-mediated signal transduction and thymocyte selection. *Science*, 269(5223):535–537.

Tyson, J. J., Chen, K. C., and Novak, B. (2003). Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Current Opinion in Cell Biology*, 15(2):221–231.

Valitutti, S., Müller, S., Salio, M., and Lanzavecchia, A. (1997). Degradation of T cell receptor (TCR)-CD3-zeta complexes after antigenic stimulation. *The Journal of Experimental Medicine*, 185(10):1859–64.

Viola, A. and Lanzavecchia, A. (1996). T Cell Activation Determined by T Cell Receptor Number and Tunable Thresholds. *Science*, 273(5271):104–106.

von Plato, J. (1991). Boltzmann's ergodic hypothesis. *Archive for History of Exact Sciences*, 42(1):71–89.

Wills, Q. F., Livak, K. J., Tipping, A. J., Enver, T., Goldson, A. J., Sexton, D. W., and Holmes, C. (2013). Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nature Biotechnology*, 31(8):748–752.

Wong, P., Goldrath, A. W., and Rudensky, A. Y. (2000). Competition for specific intrathymic ligands limits positive selection in a TCR transgenic model of CD4+ T cell development. *The Journal of Immunology*, 164(12):6252–9.

Yang, W. and Grey, H. M. (2003). Study of the mechanism of TCR antagonism using dual-TCR-expressing T cells. *The Journal of Immunology*, 170(9):4532–8.

Yates, L. R. and Campbell, P. J. (2012). Evolution of the cancer genome. *Nature Reviews Genetics*, 13(11):795–806.