The symphony of gene regulation

van Dijk, D.

Citation for published version (APA):
David van Dijk received the BSc and MSc degrees in Artificial Intelligence from the VU University in Amsterdam. He completed his PhD thesis at the University of Amsterdam in 2013.
The Symphony of Gene Regulation

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op woensdag 11 september 2013, te 12:00 uur

door

David van Dijk

geboren te Utrecht
Promotor: prof. dr. P.M.A. Sloot

Copromotor: dr. J.A. Kaandorp

Overige leden: dr. E. Segal

prof. dr. J. Heringa

prof. dr. R. van Driel

prof. dr. A.H.C. van Kampen

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The work described in this dissertation has been carried out in the Computational Science research group of the University of Amsterdam. I would like to acknowledge the financial support of the European Union through the projects DynaNets (EU project no: FET-233847), and ViroLab (EU project no: IST-027446). Furthermore I would like to acknowledge the financial support from an EMBO Short-Term Fellowship and a travel grant from the BioRange program of the Netherlands Bioinformatics Centre (NBIC) and Netherlands Genomics Initiative (NGI).

Author contact: davidvandijk@gmail.com
Table of Contents

CHAPTER 1 INTRODUCTION 5
1.1 A NEW AGE IN BIOLOGY 5
1.2 BIOLOGICAL BACKGROUND 6
1.3 RESEARCH QUESTIONS 11
1.4 THESIS OVERVIEW 11
1.5 REFERENCES 13

CHAPTER 2 IDENTIFYING POTENTIAL SURVIVAL STRATEGIES OF HIV-1 THROUGH VIRUS-HOST PROTEIN INTERACTION NETWORKS 17
2.1 INTRODUCTION 18
2.2 BACKGROUND 18
2.3 RESULTS 22
2.4 DISCUSSION 33
2.5 CONCLUSIONS 39
2.6 METHODS 39
2.7 REFERENCES 42
2.8 SUPPLEMENTARY FIGURES 48

CHAPTER 3 INFERENCE OF SURFACE MEMBRANE FACTORS OF HIV-1 INFECTION THROUGH FUNCTIONAL INTERACTION NETWORKS 55
3.1 SUMMARY 56
3.2 INTRODUCTION 56
3.3 MATERIALS AND METHODS 58
3.4 RESULTS 64
3.5 DISCUSSION 71
3.6 CONCLUSIONS 74
3.7 REFERENCES 76
3.8 SUPPLEMENTARY MATERIAL 84
3.9 ADDITIONAL MATERIAL REFERENCES 92

CHAPTER 4 PROMOTER SEQUENCE DETERMINES THE RELATIONSHIP BETWEEN EXPRESSION LEVEL AND NOISE. 93
4.1 SUMMARY 94
4.2 INTRODUCTION 94
4.3 RESULTS 96
4.4 DISCUSSION 111
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5 MATERIALS AND METHODS</td>
<td>113</td>
</tr>
<tr>
<td>4.6 FINANCIAL DISCLOSURE</td>
<td>120</td>
</tr>
<tr>
<td>4.7 ACKNOWLEDGMENTS</td>
<td>120</td>
</tr>
<tr>
<td>4.8 AUTHOR CONTRIBUTIONS</td>
<td>120</td>
</tr>
<tr>
<td>4.9 REFERENCES</td>
<td>121</td>
</tr>
<tr>
<td>4.10 SUPPLEMENTAL INFORMATION</td>
<td>124</td>
</tr>
<tr>
<td>4.11 SUPPLEMENTAL FIGURES</td>
<td>129</td>
</tr>
<tr>
<td>CHAPTER 5 TWO DNA-ENCODED STRATEGIES FOR INCREASING</td>
<td></td>
</tr>
<tr>
<td>EXPRESSION WITH OPPOSING EFFECTS ON PROMOTER DYNAMICS</td>
<td></td>
</tr>
<tr>
<td>AND TRANSCRIPTIONAL NOISE</td>
<td>138</td>
</tr>
<tr>
<td>5.1 SUMMARY</td>
<td>139</td>
</tr>
<tr>
<td>5.2 INTRODUCTION</td>
<td>139</td>
</tr>
<tr>
<td>5.3 RESULTS</td>
<td>141</td>
</tr>
<tr>
<td>5.4 CONCLUSIONS AND DISCUSSION</td>
<td>146</td>
</tr>
<tr>
<td>5.5 METHODS</td>
<td>147</td>
</tr>
<tr>
<td>5.6 FIGURES</td>
<td>149</td>
</tr>
<tr>
<td>5.7 REFERENCES</td>
<td>157</td>
</tr>
<tr>
<td>5.8 SUPPLEMENTARY INFORMATION</td>
<td>160</td>
</tr>
<tr>
<td>CHAPTER 6 CONCLUSIONS</td>
<td>196</td>
</tr>
<tr>
<td>SUMMARY</td>
<td>198</td>
</tr>
<tr>
<td>SAMENVATTING</td>
<td>200</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>202</td>
</tr>
<tr>
<td>LIST OF PUBLICATIONS</td>
<td>203</td>
</tr>
</tbody>
</table>
Chapter 1 Introduction

1.1 A new age in biology

The exponentially decreasing cost of DNA sequencing has resulted in an exponentially increasing number of sequenced genomes (Figure 1.1). In parallel, new high-throughput experimental techniques can generate equally large amounts of functional data, e.g. expression levels of all genes in a cell, or the interaction-network of all proteins with all other proteins. However the huge amount of data that we are able to generate has created a new bottleneck. One of the great challenges in the post genomic era is to infer biological functioning from DNA sequence and other high-throughput experimental data. Our ability to generate quantitative and accurate predictions regarding basic things such as mechanisms of gene regulation and protein-protein interactions remain severely limited. The rapidly expanding field of systems biology combines mathematical models and high-throughput experimental data in order to understand these fundamental processes.

![Figure 1.1: The increasing number of sequenced organisms. The exponentially decreasing cost of DNA sequencing has resulted in an exponentially increasing number of sequenced genomes. Shown here are the number of fully (blue) and partially (red) sequenced genomes.](image)

1.1.1 Biology and information processing

This systems biology approach in recent years has put forward a picture of great biological complexity yet one that appears to have emerged with a simple set of rules much like the laws of physics. It is the goal of systems biologists to uncover these rules or design principles and understand how they evolved. In particular, the quantitative sciences have contributed to our ability to measure, predict and understand biology. Mathematics, physics and computer science have provided the tools to model biological functioning and enable us to answer questions on how...
organisms respond to their environments and regulate reproduction given the constraints of biochemistry.

Diverse types of such regulatory systems lie at the heart of biology, most importantly the regulation of gene expression. In this thesis we therefore ask: “How is gene regulation encoded in the “wiring” of biological interactions?” Using two different biological systems and two different regulatory mechanisms – protein-protein interaction between the HIV virus and it’s host cell, and transcriptional regulation in Saccharomyces cerevisiae (budding yeast), we show that despite the complexity and diversity of biology there are clear design principles, employed by evolution, of how gene regulation and therefore information processing occurs in nature.

1.2 Biological Background

1.2.1 Virus-host protein-protein interaction networks with HIV-1

The DNA encoded cellular circuitry that determines biological behavior is for the most part implemented in protein-protein interactions. This means that the behavior of a protein is largely determined by its interaction partners. Therefore, large-scale maps of protein interaction can be used to understand biological functioning. Usually protein-protein interaction networks of single organisms are studied, however the interaction between two different organisms or biological entities result in similar interaction networks. This is the case for virus infection.

One of the most studied viruses is HIV-1, which causes around 2 million deaths per year. Currently 33.3 million people are estimated to be living with it and each year around 2.7 million people are newly infected. The HIV-1 positive strand RNA virus that, like all viruses, relies on the bio-molecular machinery of the host cell to reproduce. It is the physical interaction between the viral proteins and the proteins produced by the host cell that permits the virus to infect the host cell, replicate, and lyse the host cell in order to go on and infect more cells. Understanding this viral ‘life-cycle’ is essential for designing a cure for any viral disease. Thus, understanding the ways in which HIV interacts with it’s host cells, macrophages and CD4+ T-cells, has medical value, and has become a well-studied model system for the viral life-cycle.

During each stage of the virus life-cycle complex interaction occurs between virus and host. First, when virus particles enter the human body an immune response is triggered when immune cells recognize the virus exterior as a foreign entity. Next, infection occurs when the virus envelope recognizes receptors that are expressed on the surface of the host’s cells. Then, after the virus releases its contents into the host cell, reverse transcribes its genome to DNA and integrates it into the host, viral proteins regulate host transcription to prefer expression of the viral genome. In turn, intra-cellular immune response of the host tries to down-regulate virus expression and interfere with virion production. These stages of infection
combined give rise to a complex gene and protein interaction network. In this thesis
we seek to quantify this network using a set of computational techniques with the
goal to confirm existing and uncover new biological mechanisms that occur
between virus and host (see Chapter 2). In addition, we seek to uncover new
interactions by predicting cellular surface proteins that interact with HIV (see
Chapter 3).

1.2.2 Transcriptional regulation in budding yeast
Proper control of mRNA levels is critical in nearly all biological processes. Since
much of this control is encrypted within non-protein-coding regulatory regions,
deciphering the details of this mapping between DNA sequence and mRNA
expression levels is key for understanding transcriptional control. Such an
understanding could allow us to predict gene expression from DNA sequence, with
far-reaching implications. Most notably, genetic studies in a broad range of human
diseases found a substantial contribution of genetic variation in non-coding regions
to phenotypic diversity, and many expression changes have in turn been linked to
disease states. However, without a ‘regulatory code’ (comparable to the ‘genetic
code’), we cannot tell which sequence changes cause the observed expression
changes, and by what mechanism. Despite many studies of transcriptional control,
it is surprising how little we know about the quantitative effect on expression of
even the most basic organizational features of promoters.

Precise control of the average expression across a population of cells is not enough.
Gene expression is noisy, and this limits the precision with which cells can regulate
protein levels. Changes in transcription factor (TF) activity results in changes in the
expression of target promoters, but the way in which cell-to-cell variability in
expression (noise) changes as a function of TF activity is less well understood. This
is in spite of observations that noise in gene expression has a substantial fitness
cost and that the level of noise for each gene is under selective pressure.
Furthermore, stochastic variation in gene expression has been implicated in diverse
processes such as resistance to chemotherapy and antibiotics, stem cell
reprogramming and penetrance of otherwise Mendelian traits. In order to fully
understand how organisms and genetic networks function in vivo it is absolutely
essential that we understand the regulation of noise. Furthermore, because
different mechanisms of regulation have different effects on noise, measurements
of noise in gene expression provide insights into the molecular mechanisms of
transcriptional regulation.

The model organism *Saccharomyces cerevisiae* (budding yeast) is uniquely suited for
this work because its gene regulation is well characterized and because of the range
of regulatory manipulation possible. Budding yeast has long been a model organism
for genetics due to the ease with which it can be grown, mutated and mated in
order to conduct experiments in classical genetics. Budding yeast was at the
forefront of the revolution in molecular genetics due to the ease with which its
genoem can be manipulated. Thus it is well suited to continue to lead the
functional-genomics revolution due to the ease with which large-scale libraries of
altered organisms can be both generated and measured. For these reasons we use baker’s yeast as a model organism for studying many different types of molecular mechanisms, such as gene regulation.

1.2.3 Modeling Background
The complex and quantitative nature of gene regulation and interaction makes mathematical and computational modeling ideal for studying and predicting regulatory patterns. Different modeling techniques and mathematical representations have been proposed for studying biological regulatory systems (reviewed in 6,7). In this thesis we describe the topology of virus-host regulation using a network representation and the dynamics of individual regulatory interactions in yeast using kinetic and stochastic modeling.

In Chapter 2 and 3 we model gene regulation and gene interaction using a network approach. We focus on the wiring of the network as a whole and use a set of mathematical tools to infer biological meaning from the network structure and topology. In Chapter 4 and 5 we present a kinetic model of gene regulation. Here we focus on individual genes and model transcriptional regulation as a function of promoter DNA sequence. We use both stochastic simulations and an analytical solution to predict gene expression level and cell-to-cell variability.

1.2.3.1 Properties of complex biological networks
Interaction networks are central to biological regulation. Even when we do not know the quantitative properties of individual interactions between components in a network, the wiring itself, especially in larger networks, can be used to understand system-wide behaviors.

Networks that are made up of many interacting components and have a non-trivial topology are often called complex networks8,9. In such networks relatively simple interactions can give rise to complex behavior and network wiring which cannot be explained from the simple interactions alone. A wide range of mathematical concepts and computational tools has been developed to find structure in these networks. Complex networks are often found in the real world such as social, biochemical6,10 and computer networks and share common features such as heavy tail degree distributions, clustering and a hierarchical or community structure11,12. By studying properties of the topology we can learn about the behavior of the system. For example, the heavy tail degree distribution is often a sign of a process called preferential attachment13 in which new connections are more likely to form in already well-connected nodes. Many biological networks, such as transcriptional regulatory, protein-protein interaction and metabolic networks are “complex” and show the typical topological features. In this thesis we will use complex network analysis to study virus-host interaction (see Figure 1.2). We find that the HIV-human protein and gene interaction networks have a heavy-tail degree distribution and we use this topological property to find which HIV and which human proteins are important for viral infection.
A network motif is the simplest topological building block that a network is made of. By quantifying if certain three (see Figure 1.3) or four node patterns are present more than would be expected from chance, we can learn about how a network grew (evolved) and how it behaves. Network motifs have been found in a wide range of networks. For example the transcriptional regulatory network of yeast and E. coli showed network motifs that represent regulatory logic or logical gates. Interestingly, the same logical gates where found in the network motifs of electronic circuits. In this thesis we use network motifs to quantify virus-host interaction in terms of immune response and immune evasion.
Figure 1.3 Network Motifs. All possible three-node network motifs of a directed graph. (Figure is from 13).

1.2.3.2 Kinetic models of transcriptional regulation

If the connections in the network describe reactions, such as transcriptional regulation, we can describe not just the topology, or static behavior, but also the dynamics of the network.

A major challenge in biology is to predict the expression level for a set of interacting genes, given the cellular condition. This prediction requires knowledge of the regulatory circuit in which gene products, such as transcription factors, regulate other genes through promoter binding. While often the qualitative network, or “wiring”, is known, a quantitative prediction of gene expression level is much harder to achieve since it relies on knowledge of the specific stoichiometry of many (unknown) reactions.

In order to model these biochemical reactions we can setup a kinetic model that describes the different reactions of the system. In transcriptional regulation we can model gene expression as a function of transcription factor (TF) binding to the promoter region where TF binding alters the promoter configuration and as a consequence change the transcriptional efficiency of the gene 15. Such a system consists of a set of reactions describing the switching between different promoter states as a result of binding and unbinding of TFs. In addition there are transcription, translation and mRNA and protein degradation reactions. Mathematical representation of these reactions exist in the form of differential equations which, in order to get the steady state solutions, can be solved analytically or numerically.
Figure 1.4: Kinetic scheme. Kinetic scheme of gene expression in which a promoter switches between transcriptionally inactive (OFF) and active (ON) state. Such switching can be the result of transcription factor binding and unbinding or various chromatin-remodeling mechanisms such as nucleosome binding or repositioning.

Gene expression is a stochastic process. In order to model the distribution of mRNA or protein abundance, i.e. cell-to-cell variability, several techniques exist. Numerical simulation using the Gillespie algorithm is a useful method in particular for systems that do not have analytical solutions or are difficult to solve analytically. Preferably, the system is solved analytically, for example using the Master equation.

1.3 Research questions
Gene regulation is one of the main mechanisms with which biological systems process information. In this thesis our goal is to understand, using both a computational and experimental approach, how biology is designed to process information. Our goal is to find the principles of biological regulation and therefore ask: "How is information processing encoded in biological functioning?"

Using different mathematical representations and modeling techniques we set out to uncover the different layers of biological regulation. We start out at the network level where we quantify the topological wiring of regulation. For this we use the case study of HIV virus-host interaction. Here we ask: "How is gene regulation encoded in the topology of interaction networks?"

We then zoom in on the bio-molecular mechanisms of gene expression and investigate the kinetics of individual gene regulatory interactions. For this we use baker's yeast as a model organism. We ask: "How is transcriptional regulation encoded in the DNA sequence?" In specific we study the kinetic properties of transcription and translation that result in stochastic (noisy) gene expression, and ask to what extent the noise of gene expression reflects the underlying mechanisms of gene regulation.

1.4 Thesis overview
In chapter 2 we investigate how the HIV-1 virus infects and survives given the human immune system. We model HIV-1 infection using a virus-host interaction network and quantify infection dynamics and host immune response using complex network theory. We find that the central connectivity of the HIV-1 virus and it's
involvement in specific interaction patterns enable it to evade the human immune response.

In chapter 3 we predict new surface membrane receptors that interact with HIV-1 using a network analysis approach. We present a bioinformatics algorithm that predicts existing and new receptors. We confirm existing and find a set of putative surface receptors.

In chapter 4 we investigate the different regulatory mechanisms of transcription factors using the stochastic nature of gene expression. We measure the protein abundance of native and mutant promoters that are all regulated by the same TF. We perform these measurements in a dose response, where we incrementally change the concentration of the input TF and measure gene expression output. We model the behavior with a stochastic model that describes mean gene expression and cell-to-cell variability (noise) as a function of TF concentration. We find that the relationship between mean expression level and noise is unique to each promoter and reflects the underlying mechanisms of transcriptional regulation. We conclude that a single TF can regulate gene expression in multiple different ways, each resulting in a unique noise profile.

In chapter 5 we used single cell time-lapse microscopy to compare the effect on transcriptional dynamics of two distinct types of sequence changes in the promoter that can each increase the mean expression of a cell population by similar amounts but through different mechanisms. We show that increasing expression by strengthening a transcription factor binding site results in slower promoter dynamics and higher noise as compared to increasing expression by adding nucleosome-disfavoring sequences. The latter strategy likely reduces the expression variability of the cell population by increasing the frequency of transcription events. Our findings are explained well by a model in which the promoter stochastically switches between a transcriptionally active and inactive state, and where expression is modulated by changing either the frequency or period of activity. Taken together, our study demonstrates the effect of cis-regulatory elements on expression variability and points to concrete types of sequence changes that may allow partial decoupling of expression level and noise.

Finally, in chapter 6, we summarize our findings and state the contributions of this dissertation to the scientific community, and end with perspectives and future directions of this work.
1.5 References


57. Sanchez, A. & Kondev, J. Transcriptional control of noise in gene expression.


Chapter 2 Identifying potential survival strategies of HIV-1 through virus-host protein interaction networks

David van Dijk¹, Gokhan Ertaylan¹, Charles AB Boucher² and Peter MA Sloot¹

¹Computational Science, University of Amsterdam, Sciencepark 107, 1098 XG Amsterdam, The Netherlands

²Department of Virology, Erasmus Medical Center, 's-Gravendijkwal 230, 3015 CE Rotterdam, The Netherlands

2.1 Introduction
The National Institute of Allergy and Infectious Diseases has launched the HIV-1 Human Protein Interaction Database in an effort to catalogue all published interactions between HIV-1 and human proteins. In order to systematically investigate these interactions functionally and dynamically, we have constructed an HIV-1 human protein interaction network. This network was analyzed for important proteins and processes that are specific for the HIV life cycle. In order to expose viral strategies, network motif analysis was carried out showing reoccurring patterns in virus-host dynamics.

Our analyses show that human proteins interacting with HIV form a densely connected and central subnetwork within the total human protein interaction network. The evaluation of this sub-network for connectivity and centrality resulted in a set of proteins essential for the HIV life cycle. Remarkably, we were able to associate proteins involved in RNA polymerase II transcription with hubs and proteasome formation with bottlenecks. Inferred network motifs show significant over-representation of positive and negative feedback patterns between virus and host. Strikingly, such patterns have never been reported in combined virus-host systems.

We conclude that HIV infection results in a reprioritization of cellular processes reflected by an increase in the relative importance of transcriptional machinery and proteasome formation. We conclude that during the evolution of HIV, some patterns of interaction have been selected for resulting in a system where virus proteins preferably interact with central human proteins for direct control and with proteasomal proteins for indirect control over the cellular processes. Finally, the patterns described by network motifs illustrate how virus and host interact with one another.

2.2 Background
Recent advances in high throughput genome-wide screening techniques have increased not only the amount of generated data, but also its quality. In combination with the completion of the human genome project, this has led to early expectations of revolutionizing medicine. However, as often is the case in life science, the devil is in the details. We have learned that before we can efficiently use genome-wide data for developing the next generation of drugs and treatments we have to revolutionize the way we use our data [1]. Since we have recognized that we are not yet equipped with the right tools to interpret this unprecedented amount of data we have been building large databases where data is waiting to be processed into information. Today interpreting this data stands as the grand challenge for bioinformatics in the post-genomic era.
Meanwhile, hoping to solve this problem, we have been broadening our view and have been looking elsewhere for answers. One of these is the field of network science. This relatively new field has emerged from graph theory and physics and has proved to be a powerful method for the mathematical representation, visualization and analysis of complex data that involves many interacting components. In this area powerful concepts have been developed, such as network centrality, scalability and network motifs, that have enabled us to understand a system through its network topology [2-9]. Subsequently many fields have benefited from these advances. For example in epidemiology the mapping of human interactions into social networks gave insight into how sexually transmitted diseases spread in a population [10-12]. In developmental biology the representation of interactions among different genes as gene regulatory networks has been widely accepted [13-17] and in social sciences the analysis of human mobility patterns using a human interaction network helped to shed light on the dynamics of our society [18].

However, the field of virology has not yet received the full attention it deserves from network research, despite the availability of data and ready to use scientific methodology. Only recently Dyer and colleagues have described a network between human proteins interacting with viruses and other pathogens based on manually curated data from literature as well as publicly available databases [19]. In their work they give an overview of the common interacting proteins of viruses such as HIV, Incense and Measles to pathogen groups like Toxoplasma and Plasmodium. Their findings emphasize that pathogens preferentially interact with two kinds of proteins: hubs (ones that interact with many other proteins) and bottlenecks (ones that lie on many shortest paths). They also provide evidence from Gene Ontology (GO) annotation that different sets of pathogens target the same processes even though they interact with different proteins. One remarkable feature of their data is that it is highly biased towards HIV interactions. Approximately eighty percent of all interactions are specific to Human Immunodeficiency Virus (HIV).

2.2.1 Human Immunodeficiency Virus
Human immunodeficiency virus (HIV) is recognized to be responsible for one of the most destructive pandemics in recorded history. It causes thousands of deaths and substantially decreases the life quality of millions of individuals each year, most of which live in Sub-Saharan Africa.

Since the first isolation of HIV in 1981, scientists are investigating every aspect of the virus hoping to find a vaccine. Genomic research has revealed that HIV has a compact genome, which consists of nine open reading frames (leading to nine primary translation products) that code for fifteen different translational products, represented by nineteen proteins. Most of the coding regions of HIV overlap, except for the genes rev and tat that are split by introns.
Despite the compactness of its genome, HIV has a very high nucleotide substitution rate, several million times faster than one of the average eukaryotic genome. Such a high substitution rate enables a virus population to exist in a cloud of genotypes called quasispecies and to rapidly adapt to environmental changes by means of this diversity. Varying conditions such as different humoral and innate immune system responses within and between hosts or varying treatment regiments result in selection pressures therefore shifting the dominant virus genotype [20]. This led to the understanding that the persistence of the virus in host relies on the complex web of interactions it has, rather than the fitness of its structural components. In other words, HIV’s strategy for dealing with environmental stress lies in its ability to change its structural components while maintaining their function. This is also the main reason why it is unlikely that a universal vaccine will be developed using conventional methods like targeting anchor proteins. Therefore, before we can expect to start developing a cure, we need to invest more in the understanding of the interplay between the virus and the host.

Describing an interplay between two systems requires the choice of an appropriate level at which the interactions will be studied. Since many HIV-human interactions have been studied on proteins, the protein interaction level appears to be the most suitable candidate. Recently many of these interactions have been collected in the HIV-1, Human Protein Interaction Database of the National Institute of Allergy and Infectious Diseases [21]. In this database HIV proteins, interacting human proteins as well as their interaction type are collected and organized (See Table 2.1). A general statistical analysis of this database has been performed recently [21,22].

Table 2.1: Fourteen most frequent types of interactions between HIV and human proteins.

<table>
<thead>
<tr>
<th>interaction</th>
<th>frequency</th>
<th>interaction</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>interacts with</td>
<td>575</td>
<td>processed by</td>
<td>99</td>
</tr>
<tr>
<td>upregulates</td>
<td>486</td>
<td>regulated by</td>
<td>99</td>
</tr>
<tr>
<td>binds</td>
<td>411</td>
<td>phosphorylated by</td>
<td>65</td>
</tr>
<tr>
<td>activates</td>
<td>365</td>
<td>enhances</td>
<td>62</td>
</tr>
<tr>
<td>inhibits</td>
<td>270</td>
<td>cleaves</td>
<td>61</td>
</tr>
<tr>
<td>downregulates</td>
<td>262</td>
<td>induces phosphorylation of</td>
<td>53</td>
</tr>
<tr>
<td>inhibited by</td>
<td>188</td>
<td>stimulates</td>
<td>53</td>
</tr>
</tbody>
</table>

In addition to the NCBI database there are three other independent data sets available as a result of small interfering RNA (siRNA) screens [23-25]. However, there is surprisingly little overlap between these four resources. A very recent review by Bushman et al. addresses this issue by comparing the results of these
three siRNA screens [26]. There were 34 genes called in at least two siRNA screens where as little as three genes were common in all three screens. Furthermore, of the 34 genes on two or three lists, only 11 were reported in the NCBI database. They have explained several reasons that could contribute to this variation. In addition they have included the interactions from NCBI database and other related work to assemble a "host-pathogen" interaction network. The analysis of this all-combined host-pathogen network revealed ten clusters that are identified with a distinct biochemical or cellular function. The clusters that were identified not only confirm understanding of some known processes such as immune response and tat activation/transcriptional elongation but also suggest the existence of new processes previously overlooked such as proteasome and mediator complex activity.

Nevertheless there are two important shortcomings associated with siRNA screening. First, the siRNA method cannot be used to identify genes if their knockdown is toxic (i.e. resulting in cell death). Hence the method can be argued to be biased towards the identification of genes that have a phenotype, yet on the periphery of a pathway within the total HIV-1 Human interactome. Second, it does not explain the type of interaction that the suggested gene might have with HIV proteins. Therefore we argue that if one aims to identify "core proteins" involved in important processes for viral survival and also wants to analyze resulting dynamics, one has to rely on relatively less-biased and well annotated data such as the NCBI database. However the quality of the published manuscripts differ among those present in the database. In this report, all individual calls reporting interactions are treated equally for computational analyses.

2.2.2 HIV-1, Human Protein-Protein Interaction Network and Analysis
In the remaining of this paper we introduce the HIV-1 Human Protein-Protein Interaction Network based on the database by the National Institute of Allergy and Infectious Diseases (NIAID) called HIV-1, Human Protein Interaction Database. In the results section we present our findings from network centrality and network motif analysis. In the discussion section we discuss the analysis of network topology and patterns that has led to the Identification of HIV specific proteins and processes associated with viral survival. In the methods section we explain how our network was inferred and annotated with publicly available human protein interaction data and gene ontology (GO) terms. Subsequently, newly developed algorithms are described in the methods section.
2.3 Results

2.3.1 Connectivity Analysis
The National Institute of Allergy and Infectious Diseases’ (NIAID) HIV-1, Human Protein Interaction Database offers comprehensive data on nineteen HIV proteins (fifteen structural and four intermediate proteins) interacting with 1452 human proteins via 3959 interactions. The most frequent types of these interactions are summarized in Table 2.1 with their frequency. We can see that regulatory (up-regulates, down-regulates, regulated by) and activation/inhibition (activates, inhibits, inhibited by) are among the most common interactions. We have inferred an HIV-1 Human Protein interaction network from this data. A visualization of the network can be seen in Figure 2.1. It is apparent from this figure that some HIV proteins have many more interactions than others and some of the human proteins are interacting with more than one HIV protein. Furthermore we can state that these interactions are also different in nature. In order to distinguish between two different functional levels of interaction we have divided the total network into two distinct directed sub-networks by placing all regulatory interactions (upregulation/downregulation) in one sub-network (HIV-host regulatory network) and activation-inhibition related interactions in another (HIV-host activation/inhibition network). In order to study the influence of the pathogen on the host and vice versa only directed interactions were considered non-directed interactions, like “binds” and “interacts with” were left out. The annotations in the database are somewhat ambiguous, i.e. the regulatory interactions not only point to transcriptional regulatory processes and activation/inhibition interactions not exclusively are signaling interactions. Therefore, the concepts that we use for both sub-networks (regulation and activation/inhibition) have a broader meaning and should not be directly interpreted as transcriptional regulation and signaling networks. Nonetheless, semantically a distinction between the two can be made. Also, regulation and activation/inhibition between proteins usually act at different time-scales and on different molecular levels, even though they are not decoupled processes but are co-occurring in many pathways in the cell. For this reason distinguishing between these two functional sub-networks also gives us the opportunity to study the different levels of involvement of the HIV proteins in these sub-networks. We have therefore conducted a connectivity analysis for each HIV protein in both networks to address this issue.
Figure 2.1: HIV-Human protein interaction network. Nineteen HIV proteins that interact with 1452 human proteins through 3959 interactions. Blue nodes are human proteins and red nodes are HIV proteins. Visualization is done with the Cytoscape [54] software using the spring layout algorithm.

Figure 2.2 shows a bar-plot of all nineteen HIV proteins and their connectivity in the total, regulatory and activation/inhibition network. From this we observe a non-uniform distribution of human interactions with HIV proteins, suggesting distinct functionalities (see Figure 2.2-A). It is not surprising that Tat has many connections given its central role as transactivator in promoting viral transcription and its effect on disease progression by interacting with neighboring cells after being released to the intercellular medium [27]. Gp120 also has many interactions due to its essential function in facilitating cell entry in different cell types [28] and Gp120 shedding of the virus [29]. Gp120 shows a similar distribution as Tat and is found in infected cells as well as in the intercellular space. The structural proteins P1, P6 and Nucleocapsid as well as unspliced Pol only have a small number of interactions. This is most likely due to their specific involvement in cellular processes and their presence which is confined to the intracellular space [30]. Proteins like Tat and Gp120 have been studied extensively, possibly because of their central role in HIV infection and their potential as drug targets.
on many levels. Yet these interactions become meaningful only when we can put them into context. Therefore we have enriched our HIV-1 human protein interaction network with interactions

HIV has many interactions with human proteins, and on many levels. Yet these interactions become meaningful only when we can put them into context. Therefore we have enriched our HIV-1 human protein interaction network with interactions

This explains their overrepresentation in Figure 2.2-A. To correct for this bias we have calculated a relative connectivity distribution of the activation/inhibition and regulatory sub-networks using normalization (see section Methods for details). This allows for direct comparison of connectivities between HIV proteins and between the two sub-networks (see Figure 2.2-B).

One interesting aspect to note is that the HIV proteins that are exposed to the host environment (in the case of Gp120 and Gp41 by expression on the virus’s outer membrane or secretion to the extracellular environment in the case of Tat and Vpr) have almost exactly the same number of interactions expected from their overall connectivity. In other words, they show no sign of specification for the activation/inhibition or regulatory networks. The unspliced Gp160 on the other hand, is under-represented in the regulatory network. Furthermore, HIV Integrase has very little involvement in activation/inhibition and virtually no involvement in the regulatory processes. No significant correlation was observed with the amino-acid length of each viral protein and its involvement in any of the networks (Data not shown). We hypothesize that HIV-1 interacting central human proteins may play a significantly more important role than non-central ones in the life cycle of HIV-1. Therefore, we have conducted a similar connectivity analysis for the human proteins from the total network. From the point of view of the human proteins we once again observe a non-uniform distribution of interactions. (Data not shown) Table 2 shows the ten most connected HIV Dependency Factors (HDFs) with varying degrees. Not surprisingly three kinases Atmpk1, Prkca and Mapk3 (which take part in a wide variety of cellular processes) and the immune system cytokine Ifng are identified as the most connected proteins.

Figure 2.2: A: Number of HIV-human interactions for each HIV protein, B: Normalized relative number of HIV-human interactions for each HIV protein. The y-axis enumerates the n-fold representation of interactions per HIV protein divided by the relative number of interactions of the respective protein in the total network. An n-fold representation of n > 1 shows an over-representation, whereas n < 1 signifies an under-representation.

Table 2 shows the ten most connected HIV Dependency Factors (HDFs) with varying degrees. Not surprisingly three kinases Atmpk1, Prkca and Mapk3 (which take part in a wide variety of cellular processes) and the immune system cytokine Ifng are identified as the most connected proteins.
from human protein interaction databases BIND, BioGRID and HPRD (see Methods). First we have included interactions between the HDFs (the local network) and interactions with non-HDF human proteins (the global network). The resulting network is a human protein interaction network where HIV interacting human proteins or HDFs are connected to each other and also to non-HDF human proteins. Figure 2.3 shows an abstract representation of the structure of this network. Two degree distributions of the networks are shown in Figure 2.4. In Figure 2.4-A, we can see the degree distribution of HDFs considering only interactions with HIV proteins. In Figure 2.4-B, we only consider the HDF-HDF interactions. On both graphs the power-law distribution indicates the scale-free nature of the networks, caused by a topology where most proteins have few connections, but a small number of proteins are highly connected, thus acting as hubs. Networks with scale-free properties are thought to be resilient to random perturbations and are therefore robust [5].

Table 2.2: Top ten highest connected HDFs, considering only HIV-HDF connections.

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATMKP1 [GenBank:NP_036197.1]</td>
<td>mitogen-activated protein kinase 1</td>
<td>10</td>
</tr>
<tr>
<td>IFNG [GenBank:NP_009375.1]</td>
<td>interferon, gamma</td>
<td>9</td>
</tr>
<tr>
<td>PRKCA [GenBank:NP_000522.2]</td>
<td>protein kinase C, alpha</td>
<td>9</td>
</tr>
<tr>
<td>MAPK3 [GenBank:NP_0046810.2]</td>
<td>mitogen-activated protein kinase 3 isoform 1</td>
<td>9</td>
</tr>
<tr>
<td>ACTB [GenBank:NP_852564.1]</td>
<td>beta actin</td>
<td>8</td>
</tr>
<tr>
<td>ACTG1 [GenBank:NP_004518.2]</td>
<td>actin, gamma 1 propeptide</td>
<td>8</td>
</tr>
<tr>
<td>HLA-A [GenBank:NP_004317.1]</td>
<td>major histocompatibility complex class I A precursor</td>
<td>8</td>
</tr>
<tr>
<td>CD4  [GenBank:NP_062077.1]</td>
<td>CD4 antigen precursor</td>
<td>8</td>
</tr>
<tr>
<td>IL10 [GenBank:NP_061320.2]</td>
<td>interleukin 10 precursor</td>
<td>7</td>
</tr>
<tr>
<td>IFNA1 [GenBank:NP_062115.1]</td>
<td>interferon, alpha 1</td>
<td>7</td>
</tr>
</tbody>
</table>
Table 3: Mean values of centrality measures on HDFs andidentify proteins of global importance by re-ranking them
is also important globally. To counteract this effect we fil-
signifying strong correlations.

We plot the local against the global measures of all our
centrality in the HDF network is a predictor for high cen-
tralization in the HDF network with the total human protein interaction
and the HDF sub-network (local). Comparison of the
removal of these proteins 
removed from the network [31]. See the methods section
removal of these proteins 
removed from the network [31]. See the methods section

2.3.1.2 Metrics: Centrality
We hypothesize that central genes or proteins in the human protein interaction
network are more likely to be important players in the life cycle of the virus than
non-central ones. Therefore, after constructing the HIV-1 human protein
interaction network we have measured three types of network centrality: degree,
betweenness and eigenvector centrality on both local and global networks.

2.3.1.2 The HDF sub-network is Central
To determine the importance of individual HDFs regarding connectivity in the total
human protein interaction network we define two scores: a hub score and a
bottleneck score. The degree and the eigenvector centrality of a protein describe
how well it is connected to other proteins (see Methods for a detailed description of both measures). For this reason we have associated the term "hub" with these measures. Network centrality encompasses several different notions in connectivity analysis, degree and eigenvector centrality being two of them. Another concept that is used to describe the position in a network is by looking at paths rather than connections. Betweenness centrality is used to measure the centrality of a node in the network by counting the number of shortest paths that go through that node. In other words, how many shortest paths would increase in length if the node is removed from the network [31]. See the methods section for a definition of and earlier work on network ranking [19,32]. Table 2.3 shows these centrality metrics measured for the total human protein interaction network (global) and the HDF sub-network (local). Comparison of the HDF network with the total human protein interaction network using a Kolmogorov-Smirnov test shows that the measured degrees, eigenvector centralities and betweenness scores in the local and global network are not from the same distribution (see Table 2.3). Because the Kolmogorov-Smirnov test was performed one-sided, we can conclude that the local network is significantly more central than the global network with respect to the three metrics. This indicates a densely connected HDF network that takes on a central position in the whole human protein interaction network. Subsequently, this shows that the human proteins interacting with HIV tend to be involved in other important processes as well.

### Table 2.3: Mean values of centrality measures on HDFs and on proteins of the whole human protein interaction network, with standard deviations between brackets.

<table>
<thead>
<tr>
<th></th>
<th>HDF network</th>
<th>total human network</th>
<th>p(ks) (HDF &gt; total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>degree</td>
<td>16 (25)</td>
<td>6 (11)</td>
<td>2.42·10^{-87}</td>
</tr>
<tr>
<td>betweenness</td>
<td>63·10^{-3} (19·10^{-3})</td>
<td>17·10^{-3} (79·10^{-3})</td>
<td>9.22·10^{-44}</td>
</tr>
<tr>
<td>eigenv. centr.</td>
<td>0.049 (0.10)</td>
<td>0.013 (0.04)</td>
<td>3.88·10^{-97}</td>
</tr>
</tbody>
</table>

### 2.3.1.3 Hubs

We define a hub as a protein with high degree and eigenvector centrality (see Methods section). Table 2.4 shows the proteins that were commonly identified as central nodes by both of these metrics. Table 2.4 summarizes the top one percent of the highest ranked HDFs in the total network. We notice from this table that both centrality metrics result in very similar sets of top ranked proteins. We can see that P53, Brca-1 and Retinoblastoma-1 have been identified as being highly central by both metrics. This result is not surprising since all three are well-established oncogenes and have been extensively studied. Therefore their connections with other proteins are expected to be better documented.
Table 2.4: Set of proteins that is found to be hubs by both the degree and eigenvector centrality metrics.

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
<th>Degree</th>
<th>Eigenv. centr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP53 [GenBank:NP_000537.3]</td>
<td>tumor protein p53</td>
<td>93</td>
<td>1.00</td>
</tr>
<tr>
<td>BICA1 [GenBank:NP_002225.1]</td>
<td>breast cancer 1, early onset isoform 1</td>
<td>74</td>
<td>0.98</td>
</tr>
<tr>
<td>ESR1 [GenBank:NP_000116.2-P]</td>
<td>estrogen receptor 1</td>
<td>59</td>
<td>0.83</td>
</tr>
<tr>
<td>CREB1 [GenBank:NP_004371.2]</td>
<td>CREB binding protein isoform a</td>
<td>58</td>
<td>0.81</td>
</tr>
<tr>
<td>RB1 [GenBank:NP_003132.2]</td>
<td>retinoblastoma 1</td>
<td>58</td>
<td>0.72</td>
</tr>
<tr>
<td>RELA [GenBank:NP_068810.2]</td>
<td>v-rel reticuloendotheliosis viral oncogene homolog A</td>
<td>57</td>
<td>0.75</td>
</tr>
<tr>
<td>SRC [GenBank:NP_005408.1]</td>
<td>proto-oncogene tyrosine-protein kinase SRC</td>
<td>57</td>
<td>0.53</td>
</tr>
<tr>
<td>TBP [GenBank:NP_006355.1]</td>
<td>TATA box binding protein</td>
<td>56</td>
<td>0.61</td>
</tr>
<tr>
<td>MYC [GenBank:NP_002458.2]</td>
<td>myc proto-oncogene protein</td>
<td>55</td>
<td>0.67</td>
</tr>
<tr>
<td>JUN [GenBank:NP_002272.1]</td>
<td>jun oncogene</td>
<td>51</td>
<td>0.72</td>
</tr>
<tr>
<td>EP300 [GenBank:NP_002451.2]</td>
<td>E1A binding protein p300</td>
<td>51</td>
<td>0.69</td>
</tr>
</tbody>
</table>

The top one percent of highest ranked proteins are shown here.

2.3.1.4 Bottlenecks

We define a protein with high betweenness score as a bottleneck [19].

Table 2.5 shows the top one percent of proteins that are called bottlenecks. Once again, highly documented proteins such as tumor protein Tp53, Ubiquitin C (UBC), Grb2 and Brca-1 are identified as the highest ranked proteins.

Table 2.5: Top one percent of proteins that have the highest score from the betweenness centrality metric.

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
<th>Betweenness Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP53 [GenBank:NP_000537.3]</td>
<td>tumor protein p53</td>
<td>44550</td>
</tr>
<tr>
<td>UBC [GenBank:NP_066289.2]</td>
<td>ubiquitin C</td>
<td>36458</td>
</tr>
<tr>
<td>GRB2 [GenBank:NP_002077.1]</td>
<td>growth factor receptor-bound protein 2 isoform 1</td>
<td>22792</td>
</tr>
<tr>
<td>BRCA1 [GenBank:NP_000272.1]</td>
<td>breast cancer 1, early onset isoform 1</td>
<td>21622</td>
</tr>
<tr>
<td>SRC [GenBank:NP_005408.1]</td>
<td>proto-oncogene tyrosine-protein kinase SRC</td>
<td>21568</td>
</tr>
<tr>
<td>EGF [GenBank:NP_005128.2]</td>
<td>epidermal growth factor receptor isoform a</td>
<td>20472</td>
</tr>
<tr>
<td>STAT3 [GenBank:NP_644605.1]</td>
<td>signal transducer and activator of transcription 3 isoform 1</td>
<td>18503</td>
</tr>
<tr>
<td>ESR1 [GenBank:NP_000116.2]</td>
<td>estrogen receptor 1</td>
<td>18424</td>
</tr>
<tr>
<td>RB1 [GenBank:NP_003132.2]</td>
<td>retinoblastoma 1</td>
<td>16777</td>
</tr>
<tr>
<td>PKRS1 [GenBank:NP_852644.4]</td>
<td>phosphoinositol-3-kinase, regulatory subunit, polypeptide 1 isoform 1</td>
<td>16048</td>
</tr>
<tr>
<td>MYC [GenBank:NP_002458.2]</td>
<td>myc proto-oncogene protein</td>
<td>15686</td>
</tr>
<tr>
<td>SPI1 [GenBank:NP_672462.1]</td>
<td>Sp1 transcription factor</td>
<td>14620</td>
</tr>
<tr>
<td>RELA [GenBank:NP_068810.2]</td>
<td>v-rel reticuloendotheliosis viral oncogene homolog A</td>
<td>14114</td>
</tr>
<tr>
<td>SHC1 [GenBank:NP_001020.2]</td>
<td>Src homology 2 domain containing transforming protein 1 isoform p52Shc</td>
<td>14011</td>
</tr>
</tbody>
</table>
2.3.1.5 Identification of host factors that are specific to HIV infection

It is not surprising that from our centrality analysis the proteins that are important for the functioning of a cell are also crucial for the viral survival. The question that remains is "Are there HIV specific processes that are crucial for viral existence but not as important for the cell?"

In order to understand the relation between local (related to other HDFs) and global (related to all human proteins) properties of HDFs, we check whether high centrality in the HDF network is a predictor for high centrality in the total human protein interaction network. We plot the local against the global measures of all our metrics. In Figure 2.5 these three plots are shown, clearly signifying strong correlations. Because of this strong correlation between local and global properties almost any protein that is identified as highly essential using a ranking based on local properties is also important globally. To correct for this bias we filter out proteins of global importance by re-ranking them using an adjusted metric (see methods for details).

In Table 2.6 the top one percent of proteins that are identified by both corrected degree and corrected eigenvector centrality metrics are shown. We observe from this table that highly studied oncogene products are replaced by the transcription machinery related proteins TBP-associated factor 1 isoform 1 (Taf1), Activating transcription factor 2 (Taf2), General transcription factor H1B (Gtf2b). This finding is important because it indicates that transcription is a vital process for HIV to synthesize proteins necessary for forming progeny.

Table 2.7 is the result of normalizing the local betweenness measure by the global value. Proteasome subunits have gained significant importance and they constitute the three highest ranked proteins in terms of betweenness. Interestingly, in a recent review on the meta-analysis of genome-wide studies [26], proteasome has been reported to play an important role in HIV functioning. However, there is conflicting evidence regarding its action. Proteasome is predominantly reported in degradation of viral products in earlier literature [33,34] whereas recent genome-wide siRNA studies indicate a role in the facilitation of HIV infection [24-26]. Our result confirms the importance of proteasome and identifies it as a bottleneck.

To understand why proteasome appears as a crucial process specifically for HIV we have isolated all proteasomal proteins from the total network and included their first and second level interacting neighbors. The resulting network consists of three distinct clusters, where the first cluster clearly only involves proteasomal proteins. For functional annotation of the other clusters we have used the DAVID bioinformatics resources online service [35] and performed clustering with GO Biological Process (BP). After annotation, the second cluster is associated with regulation of metabolic process (80% of all proteins), and regulation of progression through cell cycle (49% of all proteins). The members of this cluster are proteins from highly connected oncogenes Tp53, Tp73, Brca-1 and Rb1. The third cluster is associated with signal transduction and cell communication (both 78%).
These findings suggest that the proteasomal proteins are identified as bottlenecks because they are connected to important cellular processes mentioned above, as well as to the rest of the network. The network visualization of the proteasomal proteins with their first and second neighbors, the lists of proteins associated with each cluster and heat maps of associated GO terms can be found in Figure 2.12.

However, some of the virus-host interaction studies have been done on individual subunits of a complex, but at other times a complex is implicated in a virus-host interaction and all subunits of that complex are linked to a virus protein even though only a few subunits might be involved in the interaction. This might lead to spurious over-represented motifs. On the other hand, if those data describing interaction of complexes rather than individual subunits is discarded this might lead to an under-representation of complexes which would in reality be present in the motif analysis. We have chosen to include these in favor of over-representation of motifs since the HIV-1 human protein interaction data is already sparse.

Table 2.6: Set of proteins that are identified as central using both adjusted centrality metrics (degree and eigenvector centrality).

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAF1 [GenBank:NP006592.2]</td>
<td>TBP-associated factor 1 isoform 1</td>
</tr>
<tr>
<td>ATF2 [GenBank:NP001871.2]</td>
<td>activating transcription factor 2</td>
</tr>
<tr>
<td>GTF2B [GenBank:NP001505.1]</td>
<td>general transcription factor II B</td>
</tr>
<tr>
<td>CCN1 [GenBank:NP444294.1]</td>
<td>cyclin D1</td>
</tr>
<tr>
<td>STAT1 [GenBank:NP009520.1]</td>
<td>signal transducer and activator of transcription 1 isoform alpha</td>
</tr>
<tr>
<td>TBP [GenBank:NP003185.1]</td>
<td>TATA box binding protein</td>
</tr>
<tr>
<td>CDKN1A [GenBank:NP003380.1]</td>
<td>cyclin-dependent kinase inhibitor 1 A</td>
</tr>
<tr>
<td>CEBPB [GenBank:NP005185.2]</td>
<td>CCAAT/enhancer binding protein beta</td>
</tr>
</tbody>
</table>

The top one percent of highest ranked proteins are shown here.
Complex networks in general and biological networks since the HIV-1 human protein interaction data is already under-representation of complexes which would in real-life be linked to a virus protein even though only a few subunits of that complex are HIV associated.

Both adjusted centrality metrics (degree and eigenvector centrality, with an R2 of 0.942). Next we describe the types of motifs found.

### Table 2.7: Top ten bottlenecks after normalization.

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
<th>Bottleneck Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSMD6 [GenBank:NP001543.1]</td>
<td>proteasome (prosome, macropain) 26S subunit, non-ATPase, 6</td>
<td>0.44</td>
</tr>
<tr>
<td>PSMA2 [GenBank:NP001779.2]</td>
<td>proteasome alpha 2 subunit</td>
<td>0.25</td>
</tr>
<tr>
<td>PSMD10 [GenBank:NP002085.1]</td>
<td>proteasome 26S non-ATPase subunit 10 isoform 1</td>
<td>0.15</td>
</tr>
<tr>
<td>DOX9 [GenBank:NP001543.5]</td>
<td>DEAH (Asp-Glu-Ala-His) box polypeptide 9</td>
<td>0.08</td>
</tr>
<tr>
<td>C04 [GenBank:AF006097.1]</td>
<td>CD4 antigen precursor</td>
<td>0.07</td>
</tr>
<tr>
<td>CD82 [GenBank:NP002222.1]</td>
<td>CD82 antigen isoform 1</td>
<td>0.07</td>
</tr>
<tr>
<td>IKBKE [GenBank:NP047211.1]</td>
<td>IKB-related kinase epsilon</td>
<td>0.06</td>
</tr>
<tr>
<td>PTEN [GenBank:NP002292.9]</td>
<td>protein tyrosine phosphatase, receptor type, C isoform 1 precursor</td>
<td>0.06</td>
</tr>
<tr>
<td>AKT1 [GenBank:NP000005.1]</td>
<td>alpha-2-macroglobulin precursor</td>
<td>0.06</td>
</tr>
<tr>
<td>CCR5 [GenBank:NP000570.1]</td>
<td>chemokine (C-C motif) receptor</td>
<td>0.05</td>
</tr>
</tbody>
</table>

#### 2.3.2 Network Motifs

Complex networks in general and biological networks specifically have been found to consist of small recurring patterns, so-called network motifs [2,4,36]. These building blocks have been used to study the structure and dynamic behavior of networks.

We carried out a network motif analysis on the regulatory and activation/inhibition sub-networks (inferred from the HIV-1 Human protein interaction network) by comparing the sub-networks with one thousand randomized networks, which were created by randomly rewiring the original networks (see Figure 2.6 and section Methods for details on the rewiring algorithm). This resulted in a number of statistically significant motifs. In Figure 2.7 a selection (see Figure 2.13 for the complete table) of motifs are listed that were found to be significant (Zscore > 2, P value < 0.02). Next we describe the types of motifs found.
32

Figure 2.6: Diagram representing the rewiring method used by the randomization algorithm. Two random edges are chosen and either the sources or the targets are switched with equal probability.

2.3.2.1 Self-regulation, feedback
A feedback pattern was found for both two and three node motifs, consisting of one human protein and one or two HIV proteins. The three node feedback loop motif (see Figure 2.8), identified as indirect self-regulation, is a pattern where an HIV protein regulates or signals a human protein that in turn regulates/signals another HIV protein. As the two HIV proteins are part of the same organic structure (the HIV pathogen) we observe a process of self-regulation or activation/inhibition (feedback) depending on the nature of the interactions. The two-node feedback loop [self-regulation motif, see Figure 2.8] consists of one HIV protein that regulates/signals a human protein that in turn regulates/signals the HIV protein. The specific type of interactions between the proteins is what determines the nature of the feedback, e.g., two up-regulations result in a positive feedback, as well as two down-regulations. On the other hand a negative feedback will be the result of one up- and one down-regulation. The three-node feedback pattern was observed in seven different regulatory motifs and in one activation/inhibition motif, additionally a two-node feedback motif was found in the regulatory network as well as in the activation/inhibition network.

2.3.2.2 Co-regulation
Co-regulation, or co-activation/inhibition is what we describe as two HIV proteins regulating/activation/inhibition one human protein (see Figure 2.8). The two interactions can be of the same type (e.g., both up-regulation, or inhibition), where they can show a potential redundancy in the system. Of the co-regulation motif we found six types of regulatory and two types of activation/inhibition motifs to be significantly over-represented.

2.3.2.3 Clique
Inclusion of interactions between HDFs (collected from human protein interaction databases, see methods section) gives the ability to study the relationship between HDFs that have a common interacting HIV protein. The network motif that is associated with this pattern is what we identify as a "clique" (see Figure 2.8). Traditionally the term clique has been used to denote a group of fully interconnected nodes [37], but has also been used to describe network motifs of the fully connected three-node subgraph [2]. In this work we study such a clique that consists of two human proteins and one HIV protein. As the interactions between
HIV and human nodes have directionality a number of different clique patterns arise, similar to the ones without HDF-HDF interactions.

A feed-forward type [2-4,36] (or self-regulatory) motif occurs when two connected HDFs are also (indirectly) interacting via an HIV protein. Co-regulation (or activation/inhibition) is also observed in the clique. Two interacting human proteins both also regulate/signal the same HIV protein. Again when the two interactions are of the same type this might indicate a redundancy (see Discussion).

Nine different clique patterns were observed in the regulatory network and five in the activation/inhibition network.

2.4 Discussion
In this study we have analyzed a pathogen-host protein interaction network in an effort to relate network topology to biological functioning. Topologically central proteins have shown to be crucial for HIV functioning and network motifs appear to be the result of the complex virus-host interplay. In this section we discuss these results from the network centrality metrics and the network motif analysis.

2.4.1 Network Centrality

2.4.1.1 HIV Human Protein Interaction Network Meta-Analysis
First we have conducted a meta-analysis of the HIVhuman protein interaction network to examine the distribution of interactions among HIV proteins as well as HDFs. Network analysis identified key components in the life cycle of HIV. The normalized relative connectivity analysis revealed involvement of viral proteins in distinct sub-functions (activation/inhibition and regulatory). Integrase is a viral enzyme that enables the viral genome to be integrated into the DNA of the host cell. In addition to this it is present at the time of the initial infection of a cell in only small amounts [38]. One can speculate that any dual function of activation/inhibition or regulatory nature would end up in reduced efficiency and probably early detection by the human immune machinery before completing the job. This might be the reason why it is involved in neither the activation/inhibition network, nor the regulatory network.

HIV proteins that are exposed to the extracellular environment (Gp120, Gp41, Tat and Vpr) have approximately an equal number of interactions inferred from their global connectivity in the total network. This is probably due to the large variety of function related to these proteins. It is indeed true for Tat and Vpr and possibly for Gp120, that they are hyperactive in terms of their role in different processes. All three proteins are also directly exposed to the extracellular factors such as antibodies. Gp41 on the other hand, is originally buried in the viral envelope and is exposed only after Gp120 binds to a CD4 receptor. In addition, Gp41 has been associated with a specific role in viral membrane fusion. So it is puzzling that Gp41 is sharing this generic connectivity profile. On the other side of the spectrum, viral enzymes RT, retropepsin and integrase all show interaction profiles that are highly
specific for activation and inhibition interactions. These enzymes are reaction specific and functional changes are likely to be too costly for the virus, therefore might be favorable to keep these proteins uni-functional.

Similar connectivity analysis for human proteins revealed Mitogen-activated protein kinase 1 (Mapk1), Interferon gamma (Ifng) and Protein kinase C alpha (Prkca) and Mitogen activated protein kinase 3 (Mapk3) as the most HIV connected nodes in HIV-human protein interaction network, having degrees 10, 9, 9 and 9 respectively. Mapk1 is identified as the integration point for multiple pathways and takes part in a wide variety of cellular processes [39]. Ifng is an important cytokine for innate and adaptive immunity. Prkca and Mapk3 are both known to be involved in various critical cellular processes. It is not unexpected that we find them to be overrepresented in the HIV-1 human protein interaction network.

2.4.1.2 Centrality Analysis
Meta-analysis of the HIV-human protein interaction network revealed that HDFs interacting with HIV constitute a non-random sub-network (HDF network) in the human interactome. We employed three centrality measures (degree, betweenness and eigenvector centrality) to analyze the HDF sub-network in detail. We calculated the average centrality measures for the HDF network as well as the total human protein interaction network. It is clear that the HDF network is located topologically central in the human-protein interaction network and is significantly densely connected.

Hub analysis of the HDF network resulted in fifteen proteins that are found to be central for at least one of the two centrality metrics (degree and eigenvector centrality) where six of them were oncogenes. Bottleneck analysis was conducted based on the betweenness centrality and resulted in a similar list to the hub analysis. Further inspection showed that both were also highly central in the total human protein interaction network.

We calculated the correlation between local and global centrality for each of the centrality metrics that resulted in high correlation for each measure. This means that the centrality assigned to each protein in the HDF network was a result of its high connectivity in the total network. To overcome this problem and identify HIV specific processes we have normalized each centrality measure from the HDF network by its global network counterpart. We observe from the normalized list that highly studied oncogenes are replaced by transcription factors, transcription factor sub-units (TBP) and transcription activators. This finding is important because although transcription is important for the cell, it is probably a vital processes for HIV to synthesize proteins necessary for forming progeny. It is important to note that in the normalized bottlenecks list, three proteasome subunits constitute the most important bottlenecks specific for the HDF network. Proteasome subunits were also identified as one of the important processes by Bushman et al. [26]. It is known that cellular proteasome can act negatively on HIV infection by destroying viral proteins but it is not clear what the overall effect is on
the infection. Our results show that the importance of protease stems from the close interaction between vital proteins in regulation of gene expression and cell communication with proteasomal proteins. Therefore proteasome seems to connect the processes governed by these proteins and the rest of the HDF network. All biochemical reactions in the cell are dynamic and their equilibrium depends on the concentration of the substrates available. Proteasomes have a unique role in this scenario by being the regulator of the concentration of particular proteins. A strong line of evidence for HIV's exploitation of proteasomal pathways comes from the innate restriction host factors that inhibit viral replication at the cellular level. Human CD317/Tetherin and APOBEC proteins (APOBEC3G and APOBEC3F) have been identified to inhibit HIV replication and render resistance to HIV infection. There is growing evidence that HIV proteins Vpu and Vif accelerate proteasomal degradation of CD317/Tetherin [40-43] and APOBEC3G/F [44,45] respectively, thus suppressing their expression and overcoming the innate resistance. Strikingly, the human restriction factor tetherin mentioned above is not curated into the NIAID database. Yet, the importance of proteasomal degradation for HIV infection has been identified independently in this work. Given the critical role of HIV’s Vif and Vpu in suppressing APOBEC3G/F and CD317 activity, we argue that pharmacologic compounds designed for restoring the activity of these intrinsic anti-viral factors in infected cells in-vivo, could have strong therapeutic benefits, and therefore deserve serious attention.

As a result, we hypothesize that after infection, apart from degrading HIV proteins, re-prioritization of proteasomal pathways is an indirect control mechanism actively engaged by the virus to manage the concentrations of pivotal proteins in the cell. We have shown that regulation of gene expression and cell communication are major processes that are directly linked to proteasome functioning.

2.4.2 Network motifs
Traditionally networks of single systems have been studied using network motifs (e.g. gene regulatory network of yeast, see [3]). Discovered patterns, in terms of over-represented network motifs, hold information on network structure and dynamics of that system. HIV infection and its life cycle are based on the interplay between two systems, namely the virus itself and the human host. Consequently, network analysis using motifs results in understanding of dynamics and structure of interplay as opposed to the functioning of the two systems independently.

2.4.2.1 General patterns
By interpreting the inferred network motifs (see Figure 2.7 and Figure 2.8) we achieve insight into this interplay. Self-regulation or feedback is a pattern that is commonly found in gene regulatory systems (see [2-4,36]). Generally these patterns indicate a response mechanism, where a signal such as a gene regulation (up-regulation, down-regulation) or a phosphorylation (activation, inhibition) of a protein A triggers a similar signal to protein B. In the two-node case the source of the signal to A is B, thus potentially resulting in a positive or negative feedback loop.
In the three-node case (two different HIV proteins) interpretation is less trivial. When we consider all HIV proteins that make up the virus as a unity, we may consider the motif as a feedback or self-regulation. Since current available data is lacking information on interactions between HIV proteins, we are not able to interpret it as a feedback loop. Yet interactions between HIV proteins, especially with the regulatory protein Tat, are known to be prevalent [27]. Therefore it is plausible to assume the existence of three node feedback loops.

One limitation of the network motif analysis is the absence of time (or causality) and spatial information associated with each event in the database. Therefore, reconstruction of pathway dynamics by means of network motifs is not possible. One way to overcome this problem, at least for some motifs, is to include interactions among human proteins that indicate shared compartments and time. For instance, co-regulation, specifically in the case of two of the same interactions, points to a potential redundance. This only holds when we assume that the two similar interactions occur in a shared spatial and temporal frame, i.e. the interactions happen in the same cellular compartment and roughly at the same time. This assumption becomes more plausible when HDF-HDF interactions are incorporated, serving as proof for the co-occurrence in time and space, of the two proteins. Co-regulation that occurs within a clique thus more strongly points to redundance. It is these redundancies that are known to contribute to the robustness of regulatory networks in general [46-48] and give evidence for a potential cause of the robust nature of HIV infections.

\subsection*{2.4.2.2 Survival Strategy}

Studying HIV-human interaction in terms of network motifs gives us the opportunity to reconstruct dynamics on the protein level. It is known that under selection pressure by the immune system the HIV virus undertakes a number of actions to evade this defense. This interplay where the host tries to undermine virus reproduction and where the virus evades immune response is the key concept for understanding virus-host relations.

Network motifs that have been found to be significantly over-represented, i.e. when their existence can not only be accounted for by randomness, show patterns that apparently have been selected for. By investigating these motifs individually we observe these strategies on the protein interaction level.

One of such motifs is a two-node feedback loop, found in the HIV-host activation/inhibition network (see motif B2 in Figure 2.7). Significant over-representation of this network motif shows the inhibitory behavior of HIV proteins on human proteins that in turn inhibit the HIV protein. We therefore refer to these patterns as an "indirect positive feedback" and in this specific case "self-activation" as inhibition of an inhibitor results in (relative) activation. Closer inspection of all occurrences of this network motif shows that the HIV Tat and Gp120 protein and the human protein Interferon Gamma (Ifng) have the highest level of involvement. Gene Ontology analysis of the observed network motif indicates that the human
proteins involved in the network motif are involved in immune response (see Figure 2.10-Figure 2.12).

Ifng, or type II interferon, is a cytokine critical for innate and adaptive immunity against viral and intracellular bacterial infections and for tumor control. The importance of Ifng in the immune system stems in part from its ability to inhibit viral replication directly, but most importantly derives from its immunostimulatory and immunomodulatory effects [49,50].

We want to acknowledge that the results presented in this paper are based on annotated protein interaction data from the NIAID database. This data varies strongly in quality and it can be argued to contain a bias as a result from translating individual reports into a structured database. Therefore the results presented above should be interpreted qualitatively authentic rather than quantitatively accurate. Nonetheless, the presented work is the first in the field, according to our knowledge, to incorporate network centrality analysis and network motifs in a virus-host protein interaction network. We encourage experimental testing of the results in this paper to study their potential role in HIV infection.
38 | Identifying potential survival strategies of HIV-1 through virus-host protein interaction networks

Gp41 on the other hand, is originally buried in the viral membrane and not directly exposed to the extracellular factors such as antibodies. Therefore, it has a limited role in different processes. All three proteins are also likely for Gp120, that they are not present on the cell surface.

van Dijk et al. classified network motifs as significant when the deviation of the motif found in one thousand randomized networks. Real number of specific motifs found. 

Interactions can either be activations/up-regulations (+), inhibitions/down-regulations (-), activation/inhibition/regulation (±), or both (arrow without sign). Real is the number of specific motifs found. Network motifs were classified as significant when P-value < 0.02 and Z-score > 2.

<table>
<thead>
<tr>
<th>Motif</th>
<th>N_real</th>
<th>N_rand ± SD</th>
<th>P_value</th>
<th>Z_score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>15</td>
<td>7 ± 4</td>
<td>0.0134</td>
<td>2.21</td>
<td>positive feedback</td>
</tr>
<tr>
<td>A2</td>
<td>8</td>
<td>3 ± 2</td>
<td>0.0181</td>
<td>2.09</td>
<td>positive feedback</td>
</tr>
<tr>
<td>A3</td>
<td>200</td>
<td>163 ± 13</td>
<td>2.5 \cdot 10^{-3}</td>
<td>2.80</td>
<td>positive co-regulation</td>
</tr>
<tr>
<td>A4</td>
<td>115</td>
<td>51 ± 10</td>
<td>1.6 \cdot 10^{-10}</td>
<td>6.29</td>
<td>negative co-regulation</td>
</tr>
<tr>
<td>A5</td>
<td>4</td>
<td>1 ± 1</td>
<td>1.7 \cdot 10^{-05}</td>
<td>4.14</td>
<td>feedback loop</td>
</tr>
<tr>
<td>B1</td>
<td>167</td>
<td>78 ± 11</td>
<td>1.1 \cdot 10^{-16}</td>
<td>8.19</td>
<td>co-activation</td>
</tr>
<tr>
<td>B2</td>
<td>22</td>
<td>15 ± 3</td>
<td>0.0130</td>
<td>2.20</td>
<td>positive feedback loop (activation)</td>
</tr>
<tr>
<td>B3</td>
<td>72</td>
<td>36 ± 10</td>
<td>1.7 \cdot 10^{-4}</td>
<td>3.59</td>
<td>co-inhibition clique</td>
</tr>
<tr>
<td>B4</td>
<td>324</td>
<td>131 ± 27</td>
<td>4.5 \cdot 10^{-13}</td>
<td>7.14</td>
<td>activation clique</td>
</tr>
<tr>
<td>B5</td>
<td>175</td>
<td>74 ± 17</td>
<td>1.4 \cdot 10^{-09}</td>
<td>5.95</td>
<td>inhibition clique</td>
</tr>
</tbody>
</table>

Figure 2.7: Significantly over-represented network motifs in HIV-host protein interaction network. Black nodes are HIV proteins and white nodes are human proteins. Interactions can either be activations/up-regulations (+), inhibitions/down-regulations (-), activation/inhibition/regulation (±), or both (arrow without sign). N_real is the number of specific motifs found. N_rand ± SD is the average number and standard deviation of the motif found in one thousand randomized networks. P-value is the probability that N_real or more motifs are found in the randomized networks. Z_score is the number of standard deviations N_rand differs from N_real. Network motifs were classified as significant when P-value < 0.02 and Z_score > 2.
2.5 Conclusions

We have demonstrated that infection with HIV results in re-prioritization of cellular processes such as transcription and proteasome activity. The primary success of the HIV-1 virus depends on the synthesis of new virions in a reasonable amount of time. This has to be accomplished before the infected cells are detected by patrolling CD8+ T cells or a humoral response has emerged. Therefore it is highly plausible that hijacking of the transcriptional machinery is one of the key processes that has a pronounced role post-infection.

In addition, proteasomes not only gain significant importance for the survival of the cell by degrading HIV proteins early in the infection, but arguably also for HIV, since they regulate the concentration of the innate antiviral host factors such as APOBEC3G/F and CD317 and can be targeted by HIV proteins Vpu and Vif.

We have shown that using network motifs one can identify recurring patterns that have consequences in the virus-host dynamics. Specifically, we observed patterns that show strategies of the virus used to evade the host immune system. Finally, we conclude that the survival of HIV within the host requires direct control of the cellular machinery via the pivotal human proteins and indirect control via the proteasomes. Network motifs and complex network theory provide a promising framework to study these dynamics.

2.6 Methods

2.6.1 NCBI Database to network

The NCBI HIV-Human Protein Interaction database is used to construct a protein interaction network. The obtained network consists of nineteen HIV proteins that interact with 1452 human proteins through 3959 interactions (See Figure 1.) In this protein interaction network nodes represent either HIV or human proteins and edges interactions between them. Because interactions between HIV and human proteins are annotated (see Figure 2.8) for most common interaction types), edges in our network are directed and have an interaction type. As interactions are only between HIV and human protein, the resulting network is bipartite.
2.6.2 HIV protein connectivity

Figure 2.2 shows the connectivity of the nineteen HIV proteins in the HIV-Human protein interaction networks. Figure 2.2A shows the absolute number of interactions per HIV proteins for each of the two sub-networks and the total network. Figure 2.2B shows the normalized relative connectivity. This was achieved by first calculating the relative connectivity, by dividing the number of interactions for each protein and network by the total number of interactions in that network. Next the numbers were normalized by dividing the relative connectivity for each protein and each of the two sub-networks by the relative connectivity of that protein in the total network. This normalization permits the comparison of proteins and sub-networks.

2.6.3 Human Protein interactions

To incorporate interactions between HDFs and between HDFs and human non-HDF proteins, data on protein interaction was collected from several databases (BIND, BioGRID, HPRD) and added to the network [51-53]. As a result the network consists of nineteen HIV proteins, 1,452 HDFs and 12,557 non-HDF human proteins, and 3,959 HIV-HDF interactions, 4,540 HDF-HDF interactions and 13,189 interactions between HDFs and nonHDF human proteins.

2.6.4 Network visualization

Cytoscape [54] was used to visualize the HIV-1 human protein interaction network. The spring embedded layout algorithm was used for Figure 2.1.

2.6.5 Network statistics

Network analysis and filtering was performed using IGraph [55], an R [56] package for complex network analysis.

The metrics that are used to rank HDFs according to their importance in the network are based on a number of network centrality measures (measured per node):

- Degree: number of connected edges, i.e. number of protein interactions
- Eigenvector centrality: measure of connectedness to other well connected nodes [57,58]
- Betweenness: number of shortest paths that go through the node [59,60]

In contrast to the degree, which is a measure of direct connectedness (number of interacting proteins in our case), the eigenvector centrality measures direct and indirect connectedness. Because well connected nodes contribute more to the score of their neighbors than low connected nodes, a protein with a relative high eigenvector centrality not just indicates high activity in terms of different interactions, but also points to activity in important pathways. The betweenness centrality, on the contrary, only measures pathway activity. A protein with high betweenness is positioned at a central location in the network, as relatively many shortest paths cross it. This does not necessarily imply well connectedness in terms
of degree; a low connected protein might still have a high centrality. This way important "cross-roads" in the network can be identified, that would not have been noticed using standard degree analysis.

Using these three metrics we seek to measure the importance of human proteins that interact with HIV proteins (HDFs). In order to distinguish between HDFs that are important to whole human functioning and HDFs that are specifically important to the HIV life-cycle, we normalize our centrality ranking using a distinction between "local" and "global" metrics.

For instance, we define local degree of an HDF as the number of edges to other HDFs, and global degree of an HDF as the number of edges to any other human protein (including HDFs). So local degree measures connectivity within the HDF network, whereas global degree measures connectivity in the whole human protein interaction network. Similarly, we define local and global measures for eigenvector centrality and betweenness.

To use this for normalization, first, we filter for proteins in the top five percent of local degree, eigenvector centrality and betweenness. This results in 73 proteins for each metric. Second, to calculate the adjusted centrality metrics we divide the local by the global value. This results in three lists of proteins that are important specifically for HIV regarding these three metrics (see Table 2.6 and Table 2.7).

2.6.6 Network motif detection

The HIV-Host protein interaction network was analyzed for network motifs using a motif detection algorithm implemented in Prolog. The prolog programming language presents a useful alternative for network motif finding as the definition and detection of network patterns is highly intuitive (prolog is a declarative language used for logic programming). In contrast to the motif detection tools MAVisto [61] and Mfinder [36]), our implementation in Prolog and the FANMOD [62] motif-finding tool are able to find any annotated network pattern consisting of any number of nodes and edges. This means that we are able to specify the type of edges and nodes, thereby distinguishing between different functional motifs even though they have the same topology (i.e. distinguishing between regulatory and activation/inhibition motifs). Motif detection was carried out for all possible two and three node patterns. To determine the significance of the observed motifs, motif detection was repeated on one thousand randomized networks using a strict randomization algorithm. This ensures an unchanged connectivity distribution.

2.6.7 Randomization Algorithm

Fully randomized networks would make any found network motif to be significant. For this reason a randomized network should be as similar to the original network as possible, yet randomized. In [2,3,36] this is achieved by introducing a rewiring algorithm that iteratively switches the sources or targets of two random edges until the network is sufficiently randomized. This results in a network where the edges are randomized without changing the number of nodes or edges and without
changing the degree distribution of the network. In our approach we used a similar algorithm (see Figure 2.6) for randomizing the networks. Because edges can be of different type, we either switch the sources or targets of two randomly chosen edges with equal probability.

2.6.8 Significance
As described in [3,36] the significance of network motifs is determined using the P-value and Z-score which are calculated using the number of a specific motif found in the original network (Nreal) and the average number found in the randomized networks (Nrand) with standard deviation (SD). A network motif is found to be significant if the probability of finding the motif Nreal times in the randomized networks (P-value) is smaller than 0.02 and the number of standard deviations Nreal is removed from Nrand is at least 2. As a result the network motifs that are found to be significant cannot just be attributed to randomness.

2.7 References


Identifying potential survival strategies of HIV-1 through virus-host protein interaction networks


Identifying potential survival strategies of HIV-1 through virus-host protein interaction networks


2.8 Supplementary Figures

Figure 2.9: proteasomal network generated with Cytoscape, in spring embedded layout. The visualization of the proteasomal network with their first and second neighbors. Resulting network consists of three distinct clusters. 1) The first cluster consists of only proteasomal proteins. 2) Second cluster is associated with regulation of metabolic process (80 percent of all proteins), and regulation of progression through cell cycle (49 percent of all proteins). The members of this cluster are proteins from highly connected oncogenes Tp53, Tp73, Brca-1 and Rb1. 3) The third cluster is associated with signal transduction and cell communication (both 78 percent).
Figure 2.10: Heat-map image from proteasomal network to annotate for the common GeneOntology terms associated with cell cycle in cluster-2.
Figure 2.11: Heat-map image from proteasomal network to annotate for the common GeneOntology terms associated with regulation in duster-2.
Figure 2.12: Heat-map image from proteasomal network to annotate for the common GeneOntology terms in cluster-3.
Identifying potential survival strategies of HIV-1 through virus-host protein interaction networks

Figure 2.13: Significantly over-represented network motifs in HIV-host regulatory network. Black nodes are HIV proteins and white nodes are human proteins. Interactions can either be activations/up-regulations (+), inhibitions/down-regulations (−), signaling/regulation (±), or both (arrow without sign). Nreal is the number of specific motifs found. Nrand ± SD is the average number and standard deviation of the motif found in one thousand randomized networks. Pvalue is the probability that Nreal or more motifs are found in the randomized networks. Zscore is the number of standard deviations Nrand differs from Nreal. Network motifs were classified as significant when Pvalue < 0.02 and Zscore > 2.

<table>
<thead>
<tr>
<th>ID</th>
<th>motif</th>
<th>Nreal</th>
<th>Nrand ± SD</th>
<th>Pvalue</th>
<th>Zscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="image1" alt="Motif 1" /></td>
<td>1</td>
<td>0.01 ± 0.1</td>
<td>0</td>
<td>9.9</td>
</tr>
<tr>
<td>2</td>
<td><img src="image2" alt="Motif 2" /></td>
<td>15</td>
<td>7 ± 4</td>
<td>0.0134</td>
<td>2.21</td>
</tr>
<tr>
<td>3</td>
<td><img src="image3" alt="Motif 3" /></td>
<td>8</td>
<td>3 ± 2</td>
<td>0.0183</td>
<td>2.09</td>
</tr>
<tr>
<td>4</td>
<td><img src="image4" alt="Motif 4" /></td>
<td>2</td>
<td>0.3 ± 0.8</td>
<td>0.0167</td>
<td>2.13</td>
</tr>
<tr>
<td>5</td>
<td><img src="image5" alt="Motif 5" /></td>
<td>577</td>
<td>425 ± 20</td>
<td>3.1 · 10⁻¹⁴</td>
<td>7.50</td>
</tr>
<tr>
<td>6</td>
<td><img src="image6" alt="Motif 6" /></td>
<td>290</td>
<td>163 ± 13</td>
<td>2.5 · 10⁻³</td>
<td>2.80</td>
</tr>
<tr>
<td>7</td>
<td><img src="image7" alt="Motif 7" /></td>
<td>115</td>
<td>51 ± 10</td>
<td>1.6 · 10⁻¹⁰</td>
<td>6.29</td>
</tr>
<tr>
<td>8</td>
<td><img src="image8" alt="Motif 8" /></td>
<td>336</td>
<td>278 ± 25</td>
<td>9.4 · 10⁻³</td>
<td>2.35</td>
</tr>
<tr>
<td>9</td>
<td><img src="image9" alt="Motif 9" /></td>
<td>4</td>
<td>0.7 ± 0.8</td>
<td>1.7 · 10⁻⁵</td>
<td>4.14</td>
</tr>
<tr>
<td>10</td>
<td><img src="image10" alt="Motif 10" /></td>
<td>2341</td>
<td>1569 ± 156</td>
<td>3.9 · 10⁻⁷</td>
<td>4.94</td>
</tr>
<tr>
<td>11</td>
<td><img src="image11" alt="Motif 11" /></td>
<td>2155</td>
<td>1246 ± 128</td>
<td>6.2 · 10⁻¹³</td>
<td>7.10</td>
</tr>
<tr>
<td>12</td>
<td><img src="image12" alt="Motif 12" /></td>
<td>24466</td>
<td>21914 ± 927</td>
<td>2.9 · 10⁻³</td>
<td>2.8</td>
</tr>
</tbody>
</table>
Table 2.14: Significantly over-represented network motifs in HIV-host regulatory network. Black nodes are HIV proteins and white nodes are human proteins. Interactions can either be activations/up-regulations (+), inhibitions/down-regulations (−), signaling/regulation (±), or both (arrow without sign). \( N_{\text{real}} \) is the number of specific motifs found. \( N_{\text{rand}} \pm SD \) is the average number and standard deviation of the motif found in one thousand randomized networks. \( P_{\text{value}} \) is the probability that \( N_{\text{real}} \) or more motifs are found in the randomized networks. \( Z_{\text{score}} \) is the number of standard deviations \( N_{\text{rand}} \) differs from \( N_{\text{real}} \). Network motifs were classified as significant when \( P_{\text{value}} < 0.02 \) and \( Z_{\text{score}} > 2 \).

<table>
<thead>
<tr>
<th>ID</th>
<th>motif</th>
<th>( N_{\text{real}} )</th>
<th>( N_{\text{rand}} \pm SD )</th>
<th>( P_{\text{value}} )</th>
<th>( Z_{\text{score}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td><img src="image1" alt="motif" /></td>
<td>482</td>
<td>374 ± 48</td>
<td>0.0123</td>
<td>2.25</td>
</tr>
<tr>
<td>14</td>
<td><img src="image2" alt="motif" /></td>
<td>14718</td>
<td>12457 ± 572</td>
<td>3.9 ( \times ) 10^{-5}</td>
<td>3.95</td>
</tr>
<tr>
<td>15</td>
<td><img src="image3" alt="motif" /></td>
<td>8348</td>
<td>6709 ± 377</td>
<td>6.8 ( \times ) 10^{-6}</td>
<td>4.35</td>
</tr>
<tr>
<td>16</td>
<td><img src="image4" alt="motif" /></td>
<td>74</td>
<td>16 ± 6</td>
<td>0</td>
<td>9.97</td>
</tr>
<tr>
<td>17</td>
<td><img src="image5" alt="motif" /></td>
<td>71</td>
<td>12 ± 5</td>
<td>0</td>
<td>11.0</td>
</tr>
<tr>
<td>18</td>
<td><img src="image6" alt="motif" /></td>
<td>322</td>
<td>218 ± 35</td>
<td>1.6 ( \times ) 10^{-3}</td>
<td>2.94</td>
</tr>
<tr>
<td>19</td>
<td><img src="image7" alt="motif" /></td>
<td>22</td>
<td>4 ± 2</td>
<td>1.11 ( \times ) 10^{-16}</td>
<td>8.28</td>
</tr>
<tr>
<td>20</td>
<td><img src="image8" alt="motif" /></td>
<td>181</td>
<td>124 ± 23</td>
<td>7.7 ( \times ) 10^{-3}</td>
<td>2.43</td>
</tr>
<tr>
<td>21</td>
<td><img src="image9" alt="motif" /></td>
<td>107</td>
<td>67 ± 15</td>
<td>2.8 ( \times ) 10^{-3}</td>
<td>2.77</td>
</tr>
<tr>
<td>22</td>
<td><img src="image10" alt="motif" /></td>
<td>130</td>
<td>90 ± 19</td>
<td>0.0175</td>
<td>2.11</td>
</tr>
<tr>
<td>23</td>
<td><img src="image11" alt="motif" /></td>
<td>30</td>
<td>16 ± 6</td>
<td>0.0135</td>
<td>2.21</td>
</tr>
<tr>
<td>24</td>
<td><img src="image12" alt="motif" /></td>
<td>20</td>
<td>10 ± 5</td>
<td>0.0147</td>
<td>2.18</td>
</tr>
</tbody>
</table>
Identifying potential survival strategies of HIV-1 through virus-host protein interaction networks

Figure 2.15: Significantly over-represented network motifs in HIV-host regulatory network. Black nodes are HIV proteins and white nodes are human proteins. Interactions can either be activations/up-regulations (+), inhibitions/down-regulations (−), signaling/regulation (±), or both (arrow without sign). Nreal is the number of specific motifs found. Nrand ± SD is the average number and standard deviation of the motif found in one thousand randomized networks. Pvalue is the probability that Nreal or more motifs are found in the randomized networks. Zscore is the number of standard deviations Nrand differs from Nreal. Network motifs were classified as significant when Pvalue < 0.02 and Zscore > 2.
Chapter 3 Inference of Surface Membrane Factors of HIV-1 Infection through Functional Interaction Networks

S. Jaeger¹,³, G. Ertaylan², David van Dijk², Ulf Leser¹, Peter MA Sloot²

¹ Knowledge Management in Bioinformatics, Humboldt-Universität Berlin, Berlin, Germany
² Computational Science, University of Amsterdam, Amsterdam, The Netherlands
³ Algorithmic Computational Biology, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

3.1 Summary

3.1.1 Background
HIV infection affects the populations of T helper cells, dendritic cells and macrophages. Besides, it has a serious impact on the central nervous system. It is yet not clear whether this list is complete and why specifically those cell types are affected. To address this question, we have developed a method to identify cellular surface proteins that permit, mediate or enhance HIV infection in different cell/tissue types in HIV-infected individuals. Receptors associated with HIV infection share common functions and domains, and are involved in similar cellular processes. These properties are exploited by graph theory and a novel gene-ranking algorithm to predict unprecedented surface membrane proteins (SMP) potentially interacting with HIV.

3.1.2 Principal Findings
We compiled a set of SMPs that are known to interact with HIV from the HIV-1 protein interaction network. This set is extended by proteins that have direct interaction and share functional similarity. This resulted in a comprehensive network around the initial SMP set. Using network centrality analysis we predict novel surface membrane factors from the annotated network. We identify 21 surface membrane factors, among which three have confirmed functions in HIV infection, seven have been identified by at least two other studies, and 11 are novel predictions and thus excellent targets for experimental investigation.

3.1.3 Conclusions
Determining to what extent HIV can interact with human SMPs is an important step towards understanding patient specific disease progression. Using graph theory, GeneOntology and a gene-ranking algorithm we generate a set of surface membrane factors that constitutes a well-founded starting point for experimental testing of cell/tissue susceptibility of different HIV strains as well as for cohort studies evaluating patient specific disease progression. In conclusion, our findings constitute the necessary background for future research investigating the role of SMPs during infection with HIV.

3.2 Introduction
One of the important characteristics of Human Immunodeficiency Virus (HIV) is its ability to interact with many cell types and its capacity to alter the function of chemokines that otherwise work in harmony with the immune system. This interaction depends on the phenotype of the virus, the receptor type residing on the cell as well as the chemokines present in the environment. The main factor determining its complex interaction profile is HIV’s highly interactive proteome. Structurally, its genome has evolved to interact with many human proteins from
various cellular pathways, as was investigated in Chapter 2. Therefore, each infectious virion consists of viral proteins, such as Tat, Gp120 or Nef, which interact with proteins inside and outside the cell [1-3].

Another contributor to this complex behavior is the high degree of phenotypic variation in the HIV population in-vivo [4]. Interestingly, each transmission event (between individuals) introduces an evolutionary bottleneck since the majority of new infections are usually initiated with a single virus [5].

Typically, HIV infection is thought to originate from the contact of genital epithelia with the infectious virions. It has been suggested that Langerhans cells and resident dendritic cells of stratified squamous epithelia serve as the initial targets of HIV infection [6, 7]. Virions are mobilized to the lymph nodes either via attachment of the HIV Gp120 to the DC-SIGN receptor expressed on dendritic cells (DCs) [6] or by direct infection of DCs within epithelia via CD4 and CCR5 receptors [7]. In the lymph nodes virions are transferred to CD4+ T cells and macrophages.

Moreover, soluble Gp120 binds to Immunoglobulin-E on innate immune system cells, such as basophils, mast cells and monocytes, and induces the secretion of cytokines thereby causing further activation of type-2 T-helper cells (Th2), the primary targets of HIV-1 infection [8]. The system-wide activation of CD4+ T cells results in an increased number of infected cells and high viral reproduction that leads to viral peaks observed in the primary stages of the infection. This translates into virus populations, which essentially are genotypically related cloud(s) of phenotypes (or quasispecies). The infection, which has been ignited with a relatively small number of virions, then spreads to other tissue types harbouring immune system cells, such as CD4+CCR5+CCR3+ microglia and macrophages [9], or hMR+ astrocytes [10], megakaryocytes [11] and monocytes [12].

A puzzling fact is that the cell types that are targets of HIV infection have different receptor expression profiles and do not necessarily harbor main co-receptors CCR5 or CXCR4. For instance, in a clinical study with a heterozygote CCR5-∆32 (CCR5 delta 32) individual (which gives partial resistance to infection via CCR5 tropic viruses) a wide range of co-receptor usage is observed, suggesting the involvement of other surface membrane factors [13].

Furthermore, binding of HIV to cell surface factors other than CD4 and chemokine receptors does not always permit viral entry but leads to endocytosis of the viral particles. This promotes relocation of the infectious virions, future trans-infection of adjacent cells [14] and leads to the activation of the immune system. Therefore, it is imperative to bear in mind that there are surface membrane factors interacting with HIV proteins, hence affecting the course of infection indirectly.

Another important point regarding surface membrane proteins is that their interactions with HIV-1 proteins are not only restricted to the extracellular environment. Events taking place inside and outside the cell membrane are neither
decoupled processes, nor mutually exclusive. In vitro studies with HIV-1 protein Tat have shown that Tat is able to induce the intrinsic pathway of apoptosis in a number of human cell lines in addition to up-regulating the expression of co-receptor CCR5 and the interleukin-2 (IL-2) in HIV-1-infected cells. Extracellular Tat has also been shown to induce neuronal death by binding to the lipoprotein receptor-related protein (LRP) (see Romani et al. [15] for an extended review).

Although many steps of the virus life cycle have been unraveled and 24 distinct drugs targeted against HIV have been approved, all efforts to achieve an overall eradication of the virus have turned out to be ineffective [16]. However, life expectancy under highly active antiretroviral therapy (HAART) treatment has been extended to 21.5 years [17].

3.2.1 The missing piece of the puzzle

These observations lead to the following questions: What is the extent of surface membrane factors contributing to HIV-1 infection and how do they influence the outcome of the treatment?

HIV exploits the existing signaling and regulatory pathways in its host. The different receptors or surface membrane proteins that are targeted in different cell types are likely to be involved in the same (or closely related) functional pathways, because the range of processes and pathways available to the virus is limited. The complexity in finding the right factors arises from the fact that there are hundreds of surface membrane proteins expressed on a wide variety of cells.

Experimental testing of hundreds of targets from numerous pathways is not feasible. Therefore, we developed a computational approach that generates high quality hypotheses for wet-lab experiments with the aim to identify surface membrane host factors contributing to HIV-1 disease outcome. We adapt a strategy from disease gene discovery that is based on protein interaction, network centrality and functional similarity to receptors that are known to interact with HIV. We infer promising candidates using measures of centrality in the emerging network of proteins. This method reproduces reported factors, such as CCR1, CCBP2 and CD97, but also results in a list of proteins that likely affect the progression of the infection.

3.3 Materials and Methods

We designed a method to identify uncharacterized surface membrane factors interacting with HIV. We employ a ranking strategy based on network centrality that uses documented HIV receptors, human protein interaction data and protein functions. The algorithm is partially adapted from disease gene identification strategies that infer gene-disease associations from similarity networks and their properties. Its underlying principle is based on the assumption that the most central genes or proteins in a specific disease network are likely to be related to the disease [18, 19].
3.3.1 Conceptual design

For identifying novel surface membrane factors we developed a generic framework that infers candidate genes or proteins based on their similarity to a set of reported genes or gene products of interest. The general workflow of this framework, illustrated in Figure 3.1, comprises three steps. First, a seed set is defined by genes/proteins that share specific characteristics of interest that will be later used for growing a functional interaction network. This can be a set of proteins associated with a certain disease, involved in specific pathways, sharing other biological properties or transcripts that are differentially expressed in a condition of interest. In the second step, candidate proteins are extracted based on their functional similarity to the seed set and a domain-specific similarity network is generated by extending this set by all functionally related proteins. The notion of similarity is not necessarily restricted to functional annotation or interaction data but rather can cover any kind of genomic data, such as expression data, SNPs, sequences and phenotypes. Finally, in the last step network centrality analysis is performed to rank those proteins with respect to their relative importance within the network. The most central ones are presumed to be of functional importance for the specific network. Note that for simplicity we only referred to proteins in the description of the framework. However, our method is not restricted to proteins but is also applicable to genes depending on the biological question.
Figure 3.1: Conceptual design of the proposed prediction framework. The method consists of three components. I) Compiling "the seed set" from genes/proteins sharing specific characteristic of interest; II) Forming a network by including direct interaction partners described in database(s) and/or functionally similar partners; III) Network centrality analysis on the domain-specific similarity network to obtain potential interaction factors (PIFs). The final step of such an analysis is the experimental validation of PIFs. Confirmed PIFs then can be included in the seed set and the steps I-III can be repeated for identifying new PIFs.

Translating the general framework into the context of identifying surface membrane factors interacting with HIV-1 implies that proteins, which are related to known HIV receptors through functional similarity or interaction with the same ligand(s), tend to be part of the same pathway and often share the same biological function. Therefore, if a network is built based on documented surface membrane factors that are extended with related genes, yet undiscovered surface proteins should also be central in the resulting network. To study this, we build an enriched HIV receptor network from known HIV receptors and rank all its proteins according to their centrality within the network. Highly ranked proteins are further analyzed to identify potentially novel surface membrane factors.

Below we explain the details of our method for identifying surface membrane factors interacting with HIV-1. One should keep in mind that the framework is neither domain nor disease specific and can be applied for various biological questions other than the one presented in this study.

3.3.2 Data
We use a set of known HIV receptors, their functional annotations and human protein interaction data as a scaffold for building an HIV receptor network. The initial list is compiled by mining the literature and the 'HIV-1, Human Protein Interaction Database' [20]. A receptor is included if it is reported by at least two independent studies. This applies to 16 HIV receptors. However, three of them,
namely Rdc1, Gpr15 and ChemR23, are not documented in the data gathered from protein interaction databases (see below) and thus have not been used in this study. Table 3.1 shows the final list of 13 HIV receptors including protein domain information (InterPro), literature references and their role in HIV infection. The list covers established receptors such as CD4 and DC-SIGN. HIV co-receptors CCR5 and CXCR4 as well as alternative co-receptors CCR2 and CCR3. Only recently reported co-receptors, such as XCR1 [21], have not yet been included since they were not documented by the time the study was conducted. However, we use a list of cell surface proteins that are reported to interact with HIV in a broad sense. Therefore, we do not limit our prediction method to receptors that only permit the entry of HIV into the primary cells.

Table 3.1: Initial set of HIV seed receptors. List of seed HIV receptors, including the receptor type and their functional domains. Receptors are grouped according to their functional domains (see Figure 3.5(a) for the distribution of those domains). A full table including the complete list of references that indicate the association to HIV is provided in Table 3.7.

<table>
<thead>
<tr>
<th>Receptor</th>
<th>Receptor type</th>
<th>InterPro domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD4</td>
<td>Primary receptor for HIV</td>
<td>Ag2 comedic, Ig-like, Ig-like fold, Ig2, Ig2-set, Ig2n-set pub</td>
</tr>
<tr>
<td>7-TM GPCR and CCR, antigen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCR5</td>
<td>Co-receptor with CD4</td>
<td>7TM GPCR/β-hsdop, CC2 setC</td>
</tr>
<tr>
<td>CCR3</td>
<td>Alternative co-receptor with CD4</td>
<td>7TM GPCR/β-hsdop, CC2 setC</td>
</tr>
<tr>
<td>CCR2</td>
<td>Alternative co-receptor with CD4</td>
<td>7TM GPCR/β-hsdop, CC2 setC</td>
</tr>
<tr>
<td>CCR8</td>
<td>Alternative co-receptor with CD4</td>
<td>7TM GPCR/β-hsdop, CC2 setC</td>
</tr>
<tr>
<td>CCR9</td>
<td>Alternative co-receptor with CD4</td>
<td>7TM GPCR/β-hsdop, CC2 setC</td>
</tr>
<tr>
<td>CSF1R</td>
<td>Alternative co-receptor with CD4</td>
<td>7TM GPCR/β-hsdop, CXC2 setC</td>
</tr>
<tr>
<td>CEACAM1</td>
<td>Co-receptor</td>
<td>7TM GPCR/β-hsdop, CXC2 setC</td>
</tr>
<tr>
<td>7-TM GPCR and Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NPY</td>
<td>Alternative co-receptor</td>
<td>7TM GPCR/β-hsdop, NPY setC</td>
</tr>
<tr>
<td>GPR1</td>
<td>Alternative co-receptor</td>
<td>7TM GPCR/β-hsdop, GPR1 setC</td>
</tr>
<tr>
<td>Integrin-α</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITGA4</td>
<td>Co-receptor with CD4</td>
<td>Int α6α5α4, Int α6α5α4, Int α6α5α4a, Int α6α5α4b, Int α6α5α4c</td>
</tr>
<tr>
<td>C-type lectin and Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DC-SIGN</td>
<td>Receptor for HIV</td>
<td>Antithrombin, C-type lectin</td>
</tr>
</tbody>
</table>

Human protein interactions were obtained from the major public protein-protein interaction databases: DIP [22], IntAct [23], BIND [24], Mammalian MIPS [25], HPRD [26], MINT [27] and BioGRID [28]. From each database we retrieved the complete set of available human protein interactions. Table 3.3 provides the number of protein interactions obtained from each database by the time of this study. We integrated the different data sets by mapping the interacting proteins to unique protein identifiers from UniProt [29] or EntrezGene [30] and thus generating one comprehensive protein interaction map for our study. The
Inference of Surface Membrane Factors of HIV-1 Infection through Functional Interaction Networks

The integrated protein interaction set comprises 13,494 human proteins and 43,637 unique interactions observed between these proteins. Each protein included in the interaction map is associated with its respective protein domain information [31] and functional Gene Ontology (GO) annotations [32] (also retrieved from UniProt and EntrezGene).

3.3.2.1 HIV receptor network
We generate a specific HIV receptor network using known receptors as seeds (see Table 3.1). We map each seed gene to its protein(s) thus growing a network around them [33]. The network is extended by adding proteins that either directly interact with any seed or that are functionally similar to at least one seed. Functional similarity between two proteins is determined by using a semantic similarity measure proposed by Couto et al. [34]. The formal definition of functional similarity is provided in the Supporting Information at the end of this chapter. In principle, proteins are considered as functionally similar if their semantic similarity to a seed protein is above the threshold of 0.7 (averaged across the three GO subontologies: molecular function, biological process and cellular component). Thereby, we only consider close and significant biological relationships.

Functionally related proteins are integrated into the network through weighted edges to the seeds. Edge weights are assigned by combining a protein interaction and a GO score. The protein interaction score is either 1 if an interaction is documented between a protein and a seed, and 0 otherwise. The GO score ranges between 0 and 1 (see Supporting Information at the end of the chapter) depending on the similarity of the GO annotations between two proteins, whereby 1 indicates functional equality and 0 indicates maximal functional distance. Interactions and functional similarities among all non-seed proteins are also included into the network.

We exploit protein interaction because it strengthens the relationship between (similar) receptors interacting with the same ligand. Human interaction data, however, is still incomplete and will not cover the functional space for our analysis. Therefore, we also integrate functional data to capture cellular surface proteins that show significant functional similarity with the seed receptors. Nevertheless, the functional coverage is still limited and currently only a fraction of the genome is annotated with pathways, functions and phenotypes [19]. Hence, we integrate predicted functions in our framework to functionally enrich proteins that are weakly or not annotated at all.

3.3.3 Functional enrichment
To functionally enrich the HIV network we apply a network-based function prediction method to derive additional annotations. This method compares protein interaction networks across multiple species to detect evolutionarily and functionally conserved subgraphs. This involves the identification of orthologous proteins (using OrthoMCL [35]) and the detection and assembly of conserved interactions. Within each conserved subgraph we infer novel protein functions from
orthology relationships across species and along conserved interactions of neighboring proteins within a species (Jaeger et al. submitted, see [36] for early work). Predicted functions are added to the set of confirmed functions to better characterize proteins that are weakly or not annotated at all. The functional enrichment increases the final cross-validation recovery rate up to 30%.

3.3.4 HIV network centrality analysis

Network centrality analysis is particularly useful for identifying key elements in different biological processes. In general, networks are modeled as mathematical objects called graphs. A graph is an abstract presentation of a set of objects that are connected by links. In the most common sense a graph $G = (V; E)$ consists of a finite set of vertices $V$ and edges $E$ whereas an edge $e = \{u; v\}$ connects two vertices $u$ and $v$. Centrality, on the other hand, is formally defined as a function $C$ that determines a numerical value $C(v)$ for every vertex $v$ in a graph that describes its location relative to the other vertices. We are interested in the ranking of vertices of the given graph $G$, thus we follow the convention that a vertex $u$ is more important than another vertex $v$ if and only if $C(u) > C(v)$ [37].

Different centrality measures have been proposed for analyzing various types of biological networks [37]. Established measures are degree centrality, closeness centrality, betweenness centrality and PageRank centrality.

Here, we chose PageRank [38] to identify the most important factors within the HIV receptor network since the PageRank algorithm assigns numerical scores to each node to determine its relative importance within the network based on the assumption that not all relationships are equally important for determining the centrality of a node. Thus, links to high-scoring nodes contribute more to the PageRank centrality of a node than links to low-scoring nodes.

We used the PageRank centrality measure to discover novel surface membrane factors that are involved in HIV-1 infection. Accordingly, we rank all proteins with respect to their PageRank centrality within the network using the igraph library in R [39]. Clearly, we expect the seed receptors to be highly ranked in the ordered list, since our construction algorithm naturally places them in a central position.

Nevertheless, not all seed receptors are central, and many non-seed proteins are ranked high. We are especially interested in the latter since these are promising candidates for novel surface membrane factors. An appropriate ranking is essential for deciding which factors should be investigated further, e.g. in follow-up experiments.

3.3.5 Validation

We validate our method and the results as follows: First, we use leave-one-out cross-validation to assess the predictive power for finding novel surface membrane HIV factors. Second, we determine the statistical significance of our results by comparing them to a random control set. For cross-validation we remove one seed
receptor from the initial list and try to re-discover this receptor using our method. We build an HIV receptor network from the remaining receptors and rank the proteins according to their centrality within the network. Subsequently, we determine whether the left-out receptor is re-discovered and at which position of the ranked list. We repeat this procedure for each seed receptor and determine the average recovery rate across all receptors.

To determine the statistical significance of the results, we compare them to two random control sets. The first set, Set1, comprises all proteins from the human interaction network as candidates resulting in 13,494 proteins. The second set, Set2, is stricter and contains only proteins with receptor properties, simulating a more informed manual search. To generate this set we use specific GO annotations that imply a receptor activity since there is no general receptor definition indicating whether a protein is a receptor or not. Thus, Set2 is formed by filtering proteins from the interaction data that are annotated with at least one of these specific GO terms. This results in 2,512 candidates - covering 12 out of 13 seed receptors (ITGA4 is missing due to insufficient functional annotation). We randomly draw m samples from each control set, where m corresponds to the average number of proteins within the HIV network and determine whether the known receptors are among the samples. This is repeated 1,000 times and an average recovery rate is calculated which is later compared to the recovery rate from our ranking method.

3.4 Results
We have designed a framework for discovering novel surface membrane factors interacting with HIV-1. To this end, we use protein interaction, protein function, and network centrality analysis to determine yet uncharacterized surface membrane proteins based on their functional similarity and topological closeness to receptors that are known to interact with HIV.

Our strategy is based on the assumption that proteins, which are related to known HIV receptors through functional similarity or direct interaction with the same ligands(s), tend to be part of the same pathway and often share the same biological function. Therefore, an enriched HIV receptor network is built from documented surface membrane factors by populating it with functionally related proteins that either interact directly with or show significant functional similarity to any known factor. Subsequently, all proteins are ranked according to their centrality within the network. The underlying principle of the centrality analysis presumes that the most central proteins in a domain-specific network are likely to be of high functional relevance [40]. Thus, yet undiscovered but prospective surface proteins should also be central in the network. Highly ranked proteins are analyzed further to identify potentially novel surface membrane factors. The key steps of our inference method are illustrated in Figure 3.2.
Predicting novel HIV surface membrane factors

We studied the recovery rates using interaction data and GO annotation with and without functional enrichment. The comparison shows that the total number of re-discovered receptors is significantly higher when functionally enriched data is employed. Consequently, interaction data in combination with enriched functional annotation are chosen for further analysis.

The same evaluation was performed using random control sets, Set1 and Set2. The random recovery rates are compared to the network-driven recovery rate to assess the statistical significance. We determine the fraction of seed receptors that can be discovered when randomly sampling from the complete protein set (Set1) and a subset including only surface membrane proteins (Set2) (see Methods). On average, we discover 0.69 and 3.4 of the 13 seed receptors when sampling from Set1 and Set2, respectively, which results in random recovery rates of 5.3% and 26.2%. The comparison of recovery rates shows that the network driven recovery rate of 92% is clearly superior to the random recovery rates of 5.3% and 26.2%. The t-test confirms that the observed superiority over the control sets is statistically highly significant (p-value <2.2e-16) and thus underlines the advantage of our network-driven strategy over the random approach.

In the following subsections, we first evaluate the performance of the prediction method. Subsequently, we investigate the most promising predictions by exploring literature on their functional domains, expression levels and reported clinical evidence.

3.4.1 Cross-validation

Cross-validation is performed on 13 known HIV receptors to evaluate the predictive power of the method. Overall, we achieved a re-discovery rate of 92% (12 out of 13). ITGA4 was not re-discovered by our method, due to its insufficient annotation and low functional similarity to the other 12 receptors.

We studied the recovery rates using interaction data and GO annotation with and without functional enrichment. The comparison shows that the total number of re-discovered receptors is significantly higher when functionally enriched data is employed. Consequently, interaction data in combination with enriched functional annotation are chosen for further analysis.

In the following subsections, we first evaluate the performance of the prediction method. Subsequently, we investigate the most promising predictions by exploring literature on their functional domains, expression levels and reported clinical evidence.

3.4.1 Cross-validation

Cross-validation is performed on 13 known HIV receptors to evaluate the predictive power of the method. Overall, we achieved a re-discovery rate of 92% (12 out of 13). ITGA4 was not re-discovered by our method, due to its insufficient annotation and low functional similarity to the other 12 receptors.

We studied the recovery rates using interaction data and GO annotation with and without functional enrichment. The comparison shows that the total number of re-discovered receptors is significantly higher when functionally enriched data is employed. Consequently, interaction data in combination with enriched functional annotation are chosen for further analysis.
For the prediction of novel surface membrane factors, we investigate the trade-off between discovering potential candidates vs. false positives by normalizing the recovery rate by the number of proteins considered at each rank. Figure 3.3 compares original and normalized recovery rates across the prioritized protein list. The receptor-per-protein ratio is used to assess the probability to identify new HIV interacting surface proteins. The most significant discovery ratio is 29% (2/7) considering the top 1% proteins. The second best discovery ratio is achieved at 3%, where the probability of rediscovering a known surface membrane factor is 24% (5/21). Note that the probabilities are estimated from the cross-validation on known data and therefore provide lower bounds since all novel findings are counted as false positives.

![Graph](image)

**Figure 3.3:** Results of the leave-one-out cross-validation over the 13 seed receptors. The average receptor re-discovery rate is determined for the different ranks (in %) of the HIV network. The original re-discovery rate (left y-axis and solid line) is compared to the normalized re-discovery rate (right y-axis and dashed line). The red dashed line indicates the chosen cut-off. The x-axis is in log-scale to focus on the highest ranks (1% to 5%).

We used these receptor-per-protein ratios to define a cut-off to select candidates from the prioritized list. We choose 3% as threshold, since it presents a sensible trade-off between potential candidates and false positives (see above) while yielding a reasonable number of novel candidates. Thus, the top 21 proteins in the ranked list are considered as surface membrane factor candidates.
3.4.2 Predicting novel HIV surface membrane factors

Finally, we consider all 13 known HIV receptors as seeds to build an HIV receptor network with 739 proteins (726 candidates) and 80,000 functional relationships (note that during cross-validation we always removed one of the seeds). We ran the PageRank algorithm and obtained a list of centrality ranked proteins. Seed receptors are removed from the list since they are (by definition) highly ranked. We apply the chosen threshold and consider the first 21 proteins as host factor candidates. Table 3.2 presents the top-ranked candidates including their InterPro domains and cell types.
### Table 3.2: List of inferred surface membrane factors

<table>
<thead>
<tr>
<th>Receptor</th>
<th>Receptor-specific InterPro domains</th>
<th>Cell types</th>
<th>Association with HIV</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>7-TM GPCR and Other</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD79B</td>
<td>Ig-like fold, IgC2-like related, extracellular, immunoreceptor tyrosine-based activation motif (ITAM)</td>
<td>High expression: CD34, endothelial, B lymphoblasts, dendritic, myeloid, monocytes, NK, CD8 and CD4 T cells, whole blood</td>
<td>+</td>
</tr>
<tr>
<td>TNFRSF13B</td>
<td>Fas, Ig-like fold, C2-like domain</td>
<td>High expression: peripheral blood mononuclear cells, monocytes, myeloid and dendritic cells</td>
<td>+</td>
</tr>
<tr>
<td><strong>7-TM GPCR and CCR3 receptor</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCR1</td>
<td>CXC chemokine receptor 1</td>
<td>High expression: whole blood, monocytes, myeloid, dendritic cell</td>
<td></td>
</tr>
<tr>
<td>CCR2</td>
<td>CXC chemokine receptor 2</td>
<td>Uniform expression</td>
<td></td>
</tr>
<tr>
<td><strong>7-TM GPCR</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DARC</td>
<td>Duffy, PfEMP1</td>
<td>Uniform expression</td>
<td></td>
</tr>
<tr>
<td>CD2</td>
<td>Ig-like fold, IgC2-set, IgV-set, T-cell adhesion molecule (TCAM)</td>
<td>High expression: dendritic, myeloid, monocytes, NK, CD8 and CD4 T cells, whole blood</td>
<td>+</td>
</tr>
<tr>
<td>CCR1</td>
<td>FMS-like tyrosine kinase 1 ( FLT1), hematopoietin receptor 1, gp130, IgC2-like Ig-like fold</td>
<td>High expression: myeloid cells, monocyte and whole blood</td>
<td>+</td>
</tr>
<tr>
<td>L101</td>
<td>Ig-like fold, Ig-like fold, B10, B10.2, B10.7, B10.7.1</td>
<td>No expression profile available</td>
<td>--</td>
</tr>
<tr>
<td>CD79B</td>
<td>Ig-like fold, Ig-like fold, gp130, Fc gamma receptor IIb (Fc γRIIb), phosphatidylinositol 3-kinase (PI3K), and phosphatidylinositol 3-kinase-related protein (PIK3R)</td>
<td>High expression: CD34, endothelial and dendritic cells</td>
<td>+</td>
</tr>
<tr>
<td>LELT</td>
<td>FMS-like tyrosine kinase 1, gp130, IgC2-like Ig-like fold, IgC2-like Ig-like fold</td>
<td>Uniform expression</td>
<td>+</td>
</tr>
<tr>
<td><strong>TNFR_Cys rich_reg and Other</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TNFRSF5</td>
<td>Fas, Ig-like fold</td>
<td>High expression: B lymphoblasts</td>
<td>+</td>
</tr>
<tr>
<td>TNFRSF5</td>
<td>TLR4, TLR5, TLR9</td>
<td>High expression: myeloid cells and whole blood</td>
<td>+</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD97</td>
<td>EGF-like domain, ephrin-B like, Ig-like domain, IgC2-like Ig-like fold, GP130, GPCR, GPCR-like receptor, C25orf13, GPCR</td>
<td>High expression: CD34, B lymphoblasts, dendritic cells, CD8 and CD4 T cells, NK, myeloid, monocytes</td>
<td>+</td>
</tr>
<tr>
<td>G1B</td>
<td>LRR-containing, Cys-rich, Leu-rich region</td>
<td>High expression: CD34, monocytes and whole blood</td>
<td>+</td>
</tr>
<tr>
<td>GYRF</td>
<td>Glycophorin</td>
<td>High expression: (early) erythroid and endothelial cells</td>
<td>+</td>
</tr>
</tbody>
</table>

3Uniform expression in CD34, endothelial, B lymphoblasts, dendritic, myeloid, monocytes, NK, CD8 and CD4 T cells, and whole blood.

Figure 3.4 shows the subnetwork from the full HIV receptor network that exhibits only the direct functional relationships between seed receptors and predicted surface membrane factors. The analysis of the known and predicted surface membrane factors regarding their annotated KEGG pathways [41] revealed the involvement of three pathways, namely the chemokine signaling pathway (hsa04062), the hematopoietic cell lineage (hsa04640) and the intestinal immune network for IgA production (hsa04672).
Figure 3.4: Sub-network of the generated HIV receptor network. The sub-network focuses on the functional relationships between the seed receptors (red) and the predicted surface membrane proteins (yellow) within the HIV receptor network. Non-seeds and non-candidate proteins are not shown. Significantly enriched pathways within this sub-network are additionally highlighted.

3.4.3 Support for predictions
We assess the relevancy of the candidates using evidence that supports an association with HIV. We investigate the predictions with respect to functional domains, cell types, expression levels, associated SNPs and chromosomal locations.

Receptor domains: We analyze our predictions by comparing their functional protein domains to the domains of the known seed receptors assuming that overlapping functional domains indicate similar protein properties, e.g. binding the same ligand, and functional similarity [42]. Common protein domains of the seed receptors are:

- G-protein-coupled receptors (GPCR) rhodopsin-like superfamily and 7 transmembrane (7-TM) GPCR rhodopsin-like domains (7-TM GPCR)
The distribution of the domains among the seed receptors is shown in Figure 3.5(a).

Predicted surface membrane factors are grouped according to their functional domains (see Table 3.2) which results in GPCR with chemokine domains, GPCR without chemokine domains, Ig-like receptors, and receptors without any overlapping domains. The respective domain distribution is displayed in Figure 3.5(b). The largest domain overlap is found for 7-TM GPCR rhodopsin-like domains. Half of the predictions have this particular domain, which is also overrepresented in the set of seed receptors (10 of 13, see Figure 3.5(a)). In addition, CCR1 and CCBP2 share a chemokine domain, which is very frequent in the set of initial receptors (8 of 13). Moreover, five predicted surface membrane factors have Ig-like domains that match the primary HIV receptor CD4.

The amount of overlapping functional domains indicates that the functional characteristics of the initial HIV binding receptors are reflected in predicted surface factors. In particular, GPCRs have a broad usage spectrum as co-receptors by primary isolates of HIV [21] and specifically chemokine receptors are known as co-receptors for HIV [43]. Strikingly, CCR1 and CCBP2 share both 7-TM GPCR rhodopsin-like and chemokine domains and are reported as co-receptors of HIV. However, receptors without any overlapping domains might present...
unprecedented characteristics that are not documented in the initial set but are reflected in their complementary domain diversity (see Figure 3.5(b)).

Chromosomal locations: Genes with similar properties are sometimes located in the same regions of the human genome. Thus, the genomic location of a gene is often taken into account when new candidate genes are associated with a disease. The reason is that mapping those candidates to a region containing other genes associated to the same disease further supports the association. For example, HIV binding human CC chemokine receptor genes are known to cluster within the 3p21.3 region of the genome [44].

We determine the chromosomal location of the predicted surface proteins and study whether they cluster together with other candidates or known seed factors. The chromosomal location for each seed and prediction retrieved from EntrezGene is shown in Table 3.6: Chromosomal location of known and predicted surface membrane proteins. Similar chromosome regions are colored similarly. When considering the known receptors there is a group of six chemokine receptors that map to the CCR cluster within 3p21.3, and also two receptors, CCR1 and CCBP2, from the predicted set are associated to this region. However, the remaining ones are located on different chromosomes and do not map together. Only CD97 and DC-SIGN, and GPR17 and CXCR4 are mapped together to 19p13 and 2q21, respectively.

3.5 Discussion
The involvement of co-receptors and surface membrane proteins assisting HIV-1 infection and contributing to viral pathogenesis always has been underestimated [21]. Only a limited number of studies aim to elucidate the role of surface membrane factors interacting with viral proteins even though they are potential amenable drug targets for HIV therapeutics [45, 46].

We predict 21 surface membrane HIV factors that are potentially involved in the different stages of infection influencing the progression of the disease. Remarkably, among these cell surface proteins, three have confirmed functions in HIV infection, seven have been reported by at least two other studies and eleven predictions are novel findings that deserve experimental investigation. It is important to note that the high success rate of our method, as shown using cross-validation, strongly implicates that our predictions can be the missing piece of the puzzle.

3.5.1 Experimentally confirmed predictions
CCR1: The C-C chemokine receptor type 1 is a GPCR that mediates signal transduction and the recruitment of effector immune cells to inflammation sites. It is highly expressed in immune system cells, such as myeloids, monocytes, dendritic cells and whole blood. Independent studies confirmed the usage of CCR1 along with CD4 for the entry of HIV into target cells [21, 47].
72

**CCBP2:** The Chemokine-binding protein 2 is another chemokine receptor that is documented to function as alternative co-receptor for HIV [48].

**DARC:** The Duffy antigen/chemokine receptor belongs to the family of erythrocyte chemokine receptors that bind chemokines. It is highly expressed on red blood cells (RBCs). Several studies demonstrated the binding of HIV-1 to RBCs through DARC enabling RBCs to transmit HIV to peripheral blood mononuclear cells. However, binding HIV to DARC does not permit viral entry but retains the virus viability and mediates trans-infection of HIV-1 from RBCs to susceptible T cells [49, 50]. Recently, He et al. reported that the DARC -46C/C genotype is associated with an increase of 40% in the odds of acquiring HIV-1 in African Americans [50]. However, follow-up studies on different cohorts [51-53] or with correction for population stratification [54] could not establish a significant association of this DARC polymorphism and the increased risk for HIV-1 acquisition or disease progression. Although DARC’s association with HIV has been established some questions remain regarding its influence on HIV-1 acquisition and progression.

### 3.5.2 Prediction with direct and indirect experimental support

**CD97:** It belongs to the EGF-TM7 family of class II-7-TM molecules and is present on the surface of most activated leukocytes. It is broadly expressed on most hematopoietic cells, activated lymphocytes, macrophages, dendritic cells, granulocytes, monocytes and undergoes a rapid up-regulation during T and B cells activation. Recently, CD97 was identified in a large-scale genome RNAi screening as one of six uncharacterized host factors that are required for HIV replication [45] suggesting its crucial postintegration role. Furthermore, Kop et al. [55] showed that CD97 is present on the surface of all human lymphocytes in blood and lymphoid tissue and confirm its up-regulation upon cellular activation. In addition, they demonstrated significant differences in the expression levels between lymphocytes. For instance, T and NK cells possess higher levels of CD97 than B cells and memory CD4+ (but not CD8+) T cells express more CD97 than naive cells. These differences might present the missing factor that is required for active infection of naive T cells in early infection because CD97 is highly expressed inactivated memory CD4+ cells but not in naive subsets. To confirm this hypothesis longitudinal testing of in-vivo expression of CD97 is necessary in patients going through co-receptor switch.

**CSF3R:** The granulocyte colony-stimulating factor receptor is the receptor for colony stimulating factor 3 (G-CSF), a cytokine that controls the production, differentiation, and function of granulocytes. CSF3R is highly expressed on monocytes and activated T cells [56]. Its ligand modulates cytokine production in monocytes and lymphocytes. In particular, CSF3R is thought to play a role after viral DNA synthesis. The indirect influence on infection and replication in human cells has been demonstrated through the binding of recombinant G-CSF (rg-CSF). rg-CSF is able to activate replication of HIV-1 during hematopoietic stem cell mobilization in HIV-1 infected persons [57] and stimulates viral production through binding to CSF3R that is expressed on HIV-1 chronically infected cell lines [58]. The direct impact of CSF3R on HIV replication has been documented recently [46].
Besides, CSF3R has been linked to the developing congenital neutropenia [59]. This is particularly interesting since DARC has also been associated with benign ethnic neutropenia observed in people of African descent [60]. Thus, we hypothesize that in addition to the genetic predisposition of DARC, CSF3R can account for the observed differences in HIV induced neutropenia.

**TNFRSF3:** Also known as Lymphotoxin-beta receptor (LT-ßR), is a member of the tumor necrosis factor (TNF) receptor superfamily that participates in the regulation of immune and inflammatory responses by propagating signals that regulate cell survival or death through activation of NF-κB[61]. LT-ßR is expressed on myeloids, dendritic cells and monocytes, which play a critical role in the progression towards AIDS by providing a major source and reservoir of virus when the T cell population is depleted [12,62]. Signaling through LT-ßR via its ligand LT-ß stimulates viral replication within infected monocytes [63].

**TNFRSF5 (CD40):** CD40 is a type I membrane glycoprotein of TNF receptor superfamily and is expressed on B-lymphocytes. Its ligand CD40L is expressed mainly in activated CD4+ T lymphocytes. The interaction between CD40 and CD40L leads to the activation and differentiation of B-lymphocytes [64].

This mechanism constitutes a non-redundant central role in humoral and cell-mediated immunity. Early studies identified a link between CD40L expression and progression to AIDS [65]. Recently it has been demonstrated that HIV-1 promotes CD4+ T cell infection by inserting CD40L into emerging viral particles and trans-activating B cells in a CD40 dependent manner [66].

**CD2:** It is typically expressed on T cells and most CD3- Natural Killer (NK) cells. It mediates intracellular adhesion in T lymphocytes and targets cells for lysis in NK cells. CD2 has a pivotal role in activating and inducing latent HIV-1 replication in resting CD4+ T cells through the CD2 pathway [67]. The CD2 pathway is also reported to increase HIV production in-vivo [68]. Moreover, a longitudinal study on 'Highly active antiretroviral therapy' over a three-year period showed a significant increase of CD2 expression on peripheral blood mononuclear cells as well as a slight increase in viral load over the same period [69].

**IL6ST (GP130):** The Glycoprotein 130 is a transmembrane protein that controls the activity of cytokines, such as IL-6, IL-11, IL-27 and leukemia inhibitory factor (LIF) [70]. It is expressed in many tissues ranging from gut epithelia to astrocytes and T cell subsets. GP130 was associated with HIV when studying LIF's protective role against vertical transmission of HIV-1 from mother to child [71]. Both are significantly upregulated in lymphoid tissue [72] and found in high concentrations in plasma samples of patients [73] during primary HIV-1 infection.

Moreover, GP130 is involved in differentiation among T-helper cell (Th) subsets. A lack of GP130 in T cell specific conditional gp130 deficient mice models causes the activation of Th2 and regulatory T cell pathways [70]. In the case of HIV infection,
this change in T cell differentiation dynamics may be responsible for various levels of disease progression observed in different individuals. The imbalance of Th subsets is also a strong predictor of pathogenic SIV infection in primate models [74]. Similarly, successful CD4+ T cell restoration was associated with enhanced Th17 CD4+ T cell accumulation when comparing gut associated lymphoid tissue recovery rates from HIV infected individuals [75, 76].

**CD79B (B29, IGB):** CD79 is a transmembrane protein that forms a complex with the B-cell receptor (BCR) and generates a signal following recognition of an antigen by the BCR. It is expressed almost exclusively on B cells and B-cells neoplasms [76]. It is composed of two distinct chains called CD79A and CD79B. CD79B plays an important role in BCR expression in B cell development [77]. HIV Gp120 is documented to down-regulate CD79B [78] but its underlying mechanism is not yet understood. In theory, down-regulation of CD79B leads to reduced capacity of B-cells to bind antigens and more importantly to a decrease in HIV specific antibody formation [79].

We are aware of the difficulties for implicating HIV-1 strains efficiently using alternative co-receptors for infection of transfected cells. Experimental testing usually requires co-culturing of virus strains showing broad co-receptor usage [13, 21, 80] with appropriate transfected cell lines. However, we believe that this effort is necessary for unraveling potential causes underlying confounding traits of HIV-1 infection.

### 3.6 Conclusions

We use a systems biology framework that integrates protein interactions, functional annotation and protein domains for inferring surface membrane factors interacting with HIV. The analysis of our predictions confirms that surface membrane proteins, even though they are targeted under different conditions, are likely to be part of the same functional pathways.

We infer ten surface proteins that are involved in a cascade of events in HIV infection. Their involvement ranges from serving as co-receptors for cell entry (CCR1 and CCBP2), mediating transinfection (DARC), activating immune cells (CD97) to inducing viral production from latently infected cells (CSF3R, TNFRSF3 and CD2).

We also present eleven original predictions that are potential HIV interacting factors (see Table 3.2). In particular, the platelet glycoprotein Ib (GPIb) is a surface membrane protein of platelets. Mutations in the GPIb beta subunit are associated with Bernard-Soulier syndrome, which is characterized by thrombocytopenia, circulating giant platelets, and prolonged bleeding time [81]. We speculate that the prolonged interaction of blood platelet expressed GP1BB with HIV might be responsible for thrombocytopenia observed in HIV infection. Furthermore, the relaxin receptors RXFP1 and RXFP2 are expressed on the acrosome of elongated...
spermatids [82, 83]. Their intron rich gene organization indicates alternatively spliced variants. This suggests the existence of different protein isoforms that contribute to their diverse expression in-vivo. Their association with HIV might explain the different rates of evolution observed in seminal versus blood plasma of infected patients [84]. Moreover, either one or both receptors might be involved in viral hijacking of the spermatozoa in viral transmission [85].

Several seed receptors, such as CCR5, CCR2 and CX3CR1 [86, 87], have been associated with SNPs that contribute to different disease outcome. Among the 21 predicted factors, except for the controversial -46C/C in DARC, SNPs in CCR1, CCBP2, HTR6, HTR1B, HTR1E, CSF3R, IL1R1, TNFRSF5 are associated with one or more clinical phenotypes but their relation to HIV infection has not been investigated.

Thus, we encourage investigating the SNPs from the predicted surface membrane factors for association with HIV to study their potential effect on HIV infection.

Throughout the chapter we have presented a novel method and its application for identifying surface membrane factors for HIV-1. However, we emphasize that the presented framework is neither domain nor disease specific. More precisely, our approach is only depending on the initial (seed) data that is used to establish characteristic functional similarities. Thus, it can be employed for many biological questions other than the one discussed in this manuscript. Potential further applications include, for instance, clinical genetic studies for determining the downstream components of recently discovered disease genes, or drug-target testing for investigating possible effects/interactions of candidate compounds with proteins other than the intended targets. Note that, regardless of the context, it is crucial to test the novel hypotheses resulted from our algorithm with target-oriented in-vivo experiments to fully understand their impact on the system.

Consequently, in this chapter we started with the HIV-1 human protein interaction network and spatially restricted our focus to surface membrane proteins interacting with HIV. Later, we introduced, validated and applied a novel algorithm for predicting "potential missing links in our protein interaction network" based on their functional similarities with the proteins readily available. Finally we presented promising surface membrane factors that are potentially involved in HIV-1 infection using our algorithm.
3.7 References


Inference of Surface Membrane Factors of HIV-1 Infection through Functional Interaction Networks


Inference of Surface Membrane Factors of HIV-1 Infection through Functional Interaction Networks


3.8 Supplementary Material

Table 3.3: The number of human protein interactions retrieved from each database by the time of our study. The integration of the protein interactions from the different databases results in a protein interaction set with 13,494 human proteins and 43,637 unique interactions between these proteins. Note, that there is an overlap between the databases, thus the numbers do not added up to the final number unique interactions.

<table>
<thead>
<tr>
<th>PPI Database</th>
<th>Number of human protein interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIPS-MPPI</td>
<td>127</td>
</tr>
<tr>
<td>DIP</td>
<td>3045</td>
</tr>
<tr>
<td>MINT</td>
<td>3160</td>
</tr>
<tr>
<td>BIND</td>
<td>5969</td>
</tr>
<tr>
<td>BioGRID</td>
<td>12779</td>
</tr>
<tr>
<td>IntAct</td>
<td>14298</td>
</tr>
<tr>
<td>HPRD</td>
<td>19215</td>
</tr>
</tbody>
</table>

Table 3.4: Average number of proteins (network size) comprised in each network and the number of seeds that are re-discovered during cross-validation when considering different data for generating the specific HIV receptor network.

<table>
<thead>
<tr>
<th>HIV Network Type</th>
<th>Average Network Size</th>
<th>Number of recovered receptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPI Network</td>
<td>89 (± 6)</td>
<td>2 of 13</td>
</tr>
<tr>
<td>PPI+GO Network</td>
<td>418 (± 8)</td>
<td>11 of 13</td>
</tr>
<tr>
<td>PPI+ GO enrich Network</td>
<td>726 (± 16)</td>
<td>12 of 13</td>
</tr>
</tbody>
</table>
Table 3.5: List of functional annotation from Gene Ontology that are used to filter for receptor proteins for the control sets.

<table>
<thead>
<tr>
<th>GO Category</th>
<th>GO Term (GO Id)</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular Function</td>
<td>Receptor activity</td>
<td>Combining with an extracellular or intracellular messenger to initiate a change in cell activity.</td>
</tr>
<tr>
<td></td>
<td>(GO:0004872)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Co-receptor activity</td>
<td>Combining with an extracellular or intracellular messenger, and in cooperation with a nearby primary receptor, initiating a change in cell activity</td>
</tr>
<tr>
<td></td>
<td>(GO:0015026)</td>
<td></td>
</tr>
<tr>
<td>Biological Process</td>
<td>Receptor metabolic process</td>
<td>The chemical reactions and pathways involving a receptor molecule, a macromolecule that undergoes combination with a hormone, neurotransmitter, drug or intracellular messenger to initiate a change in cell function.</td>
</tr>
<tr>
<td></td>
<td>(GO:0043112)</td>
<td></td>
</tr>
<tr>
<td>Cellular Compartment</td>
<td>Receptor complex</td>
<td>Any protein complex that undergoes combination with a hormone, neurotransmitter, drug or intracellular messenger to initiate a change in cell function.</td>
</tr>
<tr>
<td></td>
<td>(GO:0043235)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Membrane</td>
<td>Double layer of lipid molecules that encloses all cells, and, in eukaryotes, many organelles; may be a single or double lipid bilayer; also includes associated proteins.</td>
</tr>
<tr>
<td></td>
<td>(GO:0016020)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Extracellular space</td>
<td>That part of a multicellular organism outside the cells proper, usually taken to be outside the plasma membranes, and occupied by fluid.</td>
</tr>
<tr>
<td></td>
<td>(GO:0005615)</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.6: Chromosomal location of known and predicted surface membrane proteins. Similar chromosome regions are colored similarly.

<table>
<thead>
<tr>
<th>Known factors</th>
<th>Chrom. Location</th>
<th>Predicted factors</th>
<th>Chrom. Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>CXCR4</td>
<td>2q21</td>
<td>CD2</td>
<td>1q13.1</td>
</tr>
<tr>
<td>GPR1</td>
<td>2q33.3</td>
<td>DARC</td>
<td>1q21-q22</td>
</tr>
<tr>
<td>ITGA4</td>
<td>2q31.3</td>
<td>HTR6</td>
<td>1p36-p35</td>
</tr>
<tr>
<td>CCR9</td>
<td>3p21.3</td>
<td>CSF3</td>
<td>1p35-p34.3</td>
</tr>
<tr>
<td>CCR3</td>
<td>3p21.3</td>
<td>IL1R1</td>
<td>2q12</td>
</tr>
<tr>
<td>CCR2</td>
<td>3p21.3</td>
<td>GPR17</td>
<td>2q21</td>
</tr>
<tr>
<td>CCR5</td>
<td>3p21.31</td>
<td>CCR1</td>
<td>3p21</td>
</tr>
<tr>
<td>CX3CR1</td>
<td>3p21</td>
<td>3p21.3</td>
<td>CCBP2</td>
</tr>
<tr>
<td>CXCR6</td>
<td>3p21</td>
<td>RXFP1</td>
<td>4q32.1</td>
</tr>
<tr>
<td>CCR8</td>
<td>3p22</td>
<td>GYPB</td>
<td>4q28-q31</td>
</tr>
<tr>
<td>APJ</td>
<td>11q12</td>
<td>IL6ST</td>
<td>5q11</td>
</tr>
<tr>
<td>CD4</td>
<td>12pter-p12</td>
<td>HTR1B</td>
<td>6q13</td>
</tr>
<tr>
<td>DC-SIGN</td>
<td>19p13</td>
<td>HTR1E</td>
<td>6q14-q15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TNFRSF3</td>
<td>12p13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GPR182</td>
<td>12q13.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RXFP2</td>
<td>13q13.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CD79B</td>
<td>17q23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CD97</td>
<td>19p13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TNFRSF5</td>
<td>20q12-q13.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NPBWR2</td>
<td>20q13.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GP1BB</td>
<td>22q11.21-22q11.23</td>
</tr>
</tbody>
</table>
Table 3.7: List of seed HIV receptors, including receptor name and type, their functional domains and references indicating an association with HIV-1 infection.

<table>
<thead>
<tr>
<th>Receptor</th>
<th>Receptor type</th>
<th>InterPro domains</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD4</td>
<td>Primary receptor for HIV</td>
<td>Ag_CD4, CD4-extracel, Ig-like, Ig-like_fold, Ig_C2-set, Ig_sub, Ig_V-set_sub</td>
<td>[1,2]</td>
</tr>
<tr>
<td>CCR5</td>
<td>Co-receptor with CD4</td>
<td>7TM_GPCR_Rhodpsn, CC_5_rcpt</td>
<td>[1-6]</td>
</tr>
<tr>
<td>CCR3</td>
<td>Alternative co-receptor with CD4</td>
<td>7TM_GPCR_Rhodpsn, CC_3_rcpt</td>
<td>[4,6,7]</td>
</tr>
<tr>
<td>CCR2</td>
<td>Alternative co-receptor with CD4</td>
<td>7TM_GPCR_Rhodpsn, CC_2_rcpt, CC_5_rcpt</td>
<td>[3,4]</td>
</tr>
<tr>
<td>CCR8</td>
<td>Alternative co-receptor with CD4</td>
<td>7TM_GPCR_Rhodpsn, CC_8_rcpt</td>
<td>[8,9]</td>
</tr>
<tr>
<td>CCR9</td>
<td>Alternative co-receptor with CD4</td>
<td>7TM_GPCR_Rhodpsn, CC_9_rcpt</td>
<td>[10,11]</td>
</tr>
<tr>
<td>CXCR4</td>
<td>Alternative co-receptor with CD4</td>
<td>7TM_GPCR_Rhodpsn, CXC_4_rcpt</td>
<td>[1,4]</td>
</tr>
<tr>
<td>CXCR6</td>
<td>Co-receptor</td>
<td>7TM_GPCR_Rhodpsn, CXC_6_rcpt</td>
<td>[12,5]</td>
</tr>
<tr>
<td>CX3CR1</td>
<td>Co-receptor with CD4</td>
<td>7TM_GPCR_Rhodpsn, CX3C_fract_rcpt</td>
<td>[12-14]</td>
</tr>
<tr>
<td>APJ</td>
<td>Alternative co-receptor</td>
<td>7TM_GPCR_Rhodpsn, APJ_rcpt</td>
<td>[11,15-18]</td>
</tr>
<tr>
<td>GPR1</td>
<td>Alternative co-receptor</td>
<td>7TM_GPCR_Rhodpsn, GPR1_rcpt</td>
<td>[19-21]</td>
</tr>
</tbody>
</table>
Inference of Surface Membrane Factors of HIV-1 Infection through Functional Interaction Networks

<table>
<thead>
<tr>
<th>Receptor</th>
<th>InterPro domains</th>
<th>Cell types</th>
<th>Association with HIV</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITGA4</td>
<td>Co-receptor with CD4</td>
<td>Int_alpha_beta-p, Integrin_alpha, Integrin_alpha-2, Integrin_alpha_C</td>
<td>[22-25]</td>
</tr>
<tr>
<td>DC-SIGN</td>
<td>Receptor for HIV</td>
<td>AntifreezeII, C-type_lectin.</td>
<td>[26-29]</td>
</tr>
</tbody>
</table>

Receptors are grouped according to their functional domains.

<table>
<thead>
<tr>
<th>Receptor</th>
<th>InterPro domains</th>
<th>Cell types</th>
<th>Association with HIV</th>
</tr>
</thead>
<tbody>
<tr>
<td>DARC</td>
<td>Duffy_cmk_rcept.</td>
<td>High expression: (early) erythroid, endothelial cells</td>
<td>+ [30-32]</td>
</tr>
<tr>
<td>CCR1</td>
<td>7TM_GPCR_Rhodpsn, CC_1_rcept</td>
<td>High expression: whole blood, monocytes, myeloid, dendritic cell</td>
<td>+ [33-35]</td>
</tr>
<tr>
<td>CCBP2</td>
<td>7TM_GPCR_Rhodpsn, CXC_4_rcept</td>
<td>Uniform expression*</td>
<td>+ [36-37] - [38]</td>
</tr>
<tr>
<td>CD97</td>
<td>EGF-type_Asp/Asn_hydroxyl_site, EGF_Ca_bd_2, GPCR_2_CD97, GPCR_2_secretin-like, GPS_dom</td>
<td>High expression: CD34, B lymphoblast, dendritic cells, CD8 and CD4 T- cells, NK, myeloid, monocytes</td>
<td>+ [39]</td>
</tr>
<tr>
<td>GP1BB</td>
<td>LRR-contain_N, Cys-rich_flank_reg_C</td>
<td>High expression: CD34, monocytes and whole blood</td>
<td>?</td>
</tr>
<tr>
<td>HTR6</td>
<td>7TM_GPCR_Rhodpsn</td>
<td>Uniform expression</td>
<td>+</td>
</tr>
<tr>
<td>HTR1B</td>
<td>5HT1B_rcept, 7TM_GPCR_Rhodpsn</td>
<td>Uniform expression</td>
<td>?</td>
</tr>
<tr>
<td>HTR1E</td>
<td>5HT1F_rcept, 7TM_GPCR_Rhodpsn</td>
<td>Uniform expression</td>
<td>?</td>
</tr>
<tr>
<td>RXFP2</td>
<td>7TM_GPCR_Rhodpsn, LDL_rcept_classA_cys-rich_rpt, Leu-rich_rpt, LRR-contain_N, Leu-rich_rpt_typical-subtyp, Relaxin_rcept</td>
<td>Low expression</td>
<td>?</td>
</tr>
<tr>
<td>RXFP1</td>
<td>7TM_GPCR_Rhodpsn, LDL_rcept_classA_cys-rich, Leu-rich_rpt, LRR-contain_N, Leu-rich_rpt</td>
<td>No expression profiles available</td>
<td>?</td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>Expression Profile</td>
<td>Notes</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>GPR17</td>
<td>7TM_GPCR, Rhodopsin, F2_purnocptor</td>
<td>Uniform expression</td>
<td>?</td>
</tr>
<tr>
<td>GPR182</td>
<td>7TM_GPCR, Rhodopsin, G10D_rcpt</td>
<td>Uniform expression</td>
<td>?</td>
</tr>
<tr>
<td>NPBWR2</td>
<td>7TM_GPCR, Rhodopsin, Neuropept_W_rcpt</td>
<td>Uniform expression - [40]</td>
<td></td>
</tr>
<tr>
<td>GYPB</td>
<td>Glycophorin</td>
<td>High expression: (early) erythroid and endothelial cells</td>
<td>?</td>
</tr>
<tr>
<td>CD2</td>
<td>Ag_CD2, Ig-like_fold, Ig_C2-set, Ig_V-set, T-cell_sdhesion_molc_CD2.</td>
<td>High expression: dendritic, myeloid, monocytes, NK, CD8 and CD4 T cells, whole blood</td>
<td>+ [41-44]</td>
</tr>
<tr>
<td>CSF3R</td>
<td>FN_III, Hematopoietin_rcpt_gp130_CS, IgC2-like_lig-bd</td>
<td>High expression: myeloid cells, monocytes and whole blood</td>
<td>+ [45,46]</td>
</tr>
<tr>
<td>IL1R1</td>
<td>Ig, Ig-like, Ig-like_fold, Ig_sub, IL1_rcpt_1, IL1R_rcpt</td>
<td>No expression profile available</td>
<td>- [47,48]</td>
</tr>
<tr>
<td>CD79B</td>
<td>Ig-like, Ig-like_fold, Ig_sub, Ig_V-set, Phos_immunorcpt_sig_ITAM.</td>
<td>High expression: CD34, endothelial and dendritic cells</td>
<td>+ [42,49]</td>
</tr>
<tr>
<td>IL6ST</td>
<td>FN_III, Hematopoietin_rcpt_gp130_CS, Ig-like_fold, IgC2-like_lig-bd</td>
<td>Uniform expression</td>
<td>+ [50-52]</td>
</tr>
<tr>
<td>TNFRSF5</td>
<td>Fas_rcpt, TNFR_Cys_rich_reg</td>
<td>High expression: B lymphoblasts</td>
<td>+ [53-55]</td>
</tr>
<tr>
<td>TNFRSF3</td>
<td>TNFR_3_LTBR, TNFR_Cys_rich_reg</td>
<td>High expression: myeloid, monocytes and whole blood</td>
<td>+ [56]</td>
</tr>
</tbody>
</table>

(*): CD34, endothelial, B lymphoblasts, dendritic, myeloid, monocytes, NK, CD8 and CD4 T cells, whole blood

(+) Predictions with literature supporting an interaction with HIV are marked as (+).

(-) Indicates predictions with negative evidences.

(?) For predictions without literature on interaction or non-interaction the association remains unclear.
3.8.1 Supporting Information

3.8.1.1 Functional similarity of proteins
We determine the functional similarity between two proteins by analyzing their GO annotations using semantic similarity. We first compute the similarity of two GO terms and extend the measure to determine the functional similarity of two proteins annotated with several GO terms. Note, the functional similarity between two proteins is computed separately for each of the GO subontologies: molecular function (MF), biological process (BP) and cellular component (CC).

Semantic similarity between GO terms:

To compute the semantic similarity between two GO terms we use the approach proposed by Lin [1]. Following Lin's definition, the information content of a GO term \( t \) is defined as follows:

\[
IC(t) = -\log \left( \frac{\text{freq}(t)}{\text{freq}(\text{root})} \right),
\]

(1)

Where the frequency of a term is defined as the number of times a term or any of its descendants occurs. Thus, less frequent terms and terms with few occurring descendants are considered more informative.

Based on this measure, the semantic similarity between two terms is defined as the ratio of the information content of their most informative common ancestor and the information contents of both concepts [1]. The information content of the most informative common ancestor is given by:

\[
\text{shareIC}(t_1, t_2) = \max \{ IC(t) | t \in CA(t_1, t_2) \},
\]

(2)

Where \( CA(t_1; t_2) \) is the set of all common ancestors between terms \( t_1 \) and \( t_2 \). The similarity between two terms is then defined as:

\[
\text{sim}(t_1, t_2) = \frac{2 \times \text{shareIC}(t_1, t_2)}{IC(t_1) + IC(t_2)}.
\]

(3)

Semantic similarity between proteins:

The semantic similarity between proteins is determined based on the similarity of their associated GO terms. Since often proteins are annotated with more than one term, the similarity of a protein \( p \) to a group \( g \) of terms is defined as the average similarity of its terms to their most similar terms in \( g \) [2] (where \( t(p) \) is the set of terms annotated to protein \( p \)):
Finally, the functional GO similarity between two proteins is defined as the average similarity of their GO terms:

$$ Sim(p, g) = \frac{\sum_{t_1 \in t(p)} \max \{ sim(t_1, t_2) | t_2 \in g \}}{|t(p)|} $$

Finally, the functional GO similarity between two proteins is defined as the average similarity of their GO terms:

$$ GO_{Sim}(p_1, p_2) = \frac{Sim(p_1, t(p_2)) + Sim(p_2, t(p_1))}{2}. $$

$GO_{Sim}$ ranges between 0 and 1 depending on the similarity of the GO annotations between two proteins, whereby 1 indicated functional equality and 0 indicates maximal functional distance. The functional similarity of all three GO sub-ontologies is added and then averaged to obtain an overall similarity score for two proteins:

$$ GO_{Sim}(p_1, p_2) = \frac{GO_{Sim_{cys}}(p_1, p_2) + GO_{Sim_{mp}}(p_1, p_2) + GO_{Sim_{ccy}}(p_1, p_2)}{3}. $$

3.8.1.2 Impact of the functional data on the outcomes of the prediction methods:
We use protein interaction data and functional annotations to generate an HIV specific receptor network. In addition, we assessed the influence of using manually curated and predicted functional annotation on our prediction method by applying it to differently compiled HIV networks.

HIV network types:

First, we only considered proteins that interact directly with any seed receptors when generating the specific HIV receptor network, which will be called PPI network. Next, we integrated proteins that interact directly with any seed and all proteins which are functionally very similar to any seed considering only manually curated functions -PPI-GO network. Third, we consider interaction data in combination with enriched functional annotation (manual curated and predicted function) –PPI-GO enrich network.

Performance comparison of the HIV network types:

We compare the ability of our framework to find novel surface membrane factors within the three different HIV networks by using cross-validation. Leave-one-out cross-validations are performed over the 13 known HIV receptors for the PPI, PPI-GO and PPI-GO enrich networks. For cross-validation, we remove one known HIV receptor from the initial list and try to re-discover this receptor by means of our method. We build an HIV receptor network by considering only the remaining receptors as seeds and rank the proteins according to their centrality within the network. Subsequently, we determine whether the left-out receptor is rediscovered and at which position of the ranked list. We repeat this procedure for each seed and determine an average recovery rate across all receptors and for each network type.
shows the average network size and the number of recovered (hidden) seeds for the three different kinds of HIV-receptor networks.

The seed re-discovery rate is very low when using only protein interaction data. Only two out of 13 receptors can be captured within the generated networks. This rate increases significantly up to 11 and 12 detected receptors when considering additionally functional annotation (PPI-GO) as well as predicted functions (PPI-GOenrich), respectively. Two receptors are not covered in the PPI-GO networks, namely DC-SIGN and ITGA4, whereas the latter one is also not detected using the enriched network, most likely due to different ligands and a lower functional similarity to the other seed receptors. In general, the number of re-discovered receptors is relatively low when considering only the top ranked proteins (e.g. \( x = 5\% \)). However, the recovery rate increases significantly the more proteins of the respective networks are examined (except for PPI), until it converges to the total number of detected seeds. Protein interaction data alone is not sufficient for finding the known receptors, since it captures similar ligands rather than functionally similar receptors. Utilizing interaction and functional annotations allows to generate more complete networks in biological sense. This is reflected in the average network size of the different network types which increases from 89 proteins to 418 and 726 for PPI-GOenrich.

Comparing the recovery rates of PPI-GO and PPI-GOenrich across the ranked list clearly shows that 'hidden' receptors are better recovered and more highly ranked within the enriched than in the non-enriched network. However, the superior performance might result from the larger size of the enriched networks, e.g. the number of proteins that is considered at the different \( x \) is twice as high for PPI-GOenrich because the networks are in average about two times larger. To ensure that the higher recovery rate is not affected by the larger amount of proteins we normalize the recovery rates by the number of proteins considered at each rank \( x \).

### 3.9 Additional Material References


Chapter 4 Promoter sequence determines the relationship between expression level and noise.

Lucas B. Carey*1, David van Dijk*12, Peter M.A. Sloot2, Jaap A. Kaandorp2, Eran Segal1

1Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel.

2Computational Science, University of Amsterdam, 1980 XG Amsterdam, The Netherlands.

*These authors contributed equally to this work.

PLOS Biology 11(4): e1001528. doi:10.1371/journal.pbio.1001528
4.1 Summary
The ability of cells to accurately control gene expression levels in response to extracellular cues is limited by the inherently stochastic nature of transcriptional regulation. A change in TF activity results in changes in the expression of its targets, but the way in which cell-to-cell variability in expression (noise) changes as a function of TF activity, and whether targets of the same TF behave similarly, is not known. Here, we measure expression and noise as a function of TF activity for sixteen native targets of the transcription factor Zap1 that are regulated by it through diverse mechanisms. For most activated and repressed Zap1 targets, noise decreases as expression increases. Kinetic modeling suggests that this is due to two distinct Zap1-mediated mechanisms that both change the frequency of transcriptional bursts. Notably, we found that another mechanism of repression by Zap1, which is encoded in the promoter DNA, likely decreases the size of transcriptional bursts, producing a unique transcriptional state characterized by low expression and low noise. In addition, we find that further reduction in noise is achieved when a single TF both activates and represses a single target promoter. Our results suggest a global principle whereby at low TF concentrations, the dominant source of differences in expression between promoters stems from differences in burst frequency, whereas at high TF concentrations differences in burst size dominate. Taken together, we show that the precise amount by which noise changes with expression is specific to the regulatory mechanism of transcription and translation that acts at each gene.

4.2 Introduction
The cellular response to environmental changes is mediated through activation of TFs and subsequent coordinated activation and repression of dozens of target genes. However, gene expression is noisy, and this limits the precision with which cells can regulate protein levels. Genome-wide, noise ($\sigma^2/\mu^2$, variance/mean$^2$) decreases as expression increases. Along this global trend, individual genes with the same average expression in the population differ in their amount of noise. The level of noise for each gene is related to its function and is determined by the mechanisms of regulation. However, the precise mechanisms by which control of noise is accomplished for native genes are not known.

Two quantities that describe the dynamics of gene expression, and have been related to the distribution of protein abundances, are burst size and burst frequency. Burst frequency is the rate at which the promoter switches from an inactive to an active transcriptional state due to TF binding and subsequent PolII recruitment (promoter on-switching). Burst size is the number of proteins produced during each promoter on-event. Native genes differ in the relative contribution of burst frequency and size to expression, suggesting that evolution can tune both parameters in order to reach an optimal level of expression and noise for each gene.
When an increase in gene expression is caused by an increase the rate of promoter on-switching (burst frequency), noise \( (\sigma^2/\mu^2) \) decreases monotonically with expression \(^{31}\). In contrast, an increase in burst size (due to a decrease in promoter off-switching rate or an increase in the transcription or translation rate) results in an increase in expression and in noise strength \( (\sigma^2/\mu) \), and no change in noise\(^{31,32}\). Mutations in the TATA box in yeast \(^{32}\) or in the ribosome binding site in \( E. coli \) \(^{33}\) both affect noise strength, but not noise. The former is thought to be involved in transcription re-initiation \(^{34}\), thus extending the time of each active state of the promoter, while the latter affects the number of proteins produced from each mRNA molecule. These observations strengthen the claim that changes in mean expression but not noise stem from molecular mechanisms that affect the number of proteins produced during each transcriptional event, but not the frequency of such events. Taken together, these data support a model of gene expression in which changes in promoter dynamics, such as changes in on-switching rates and transcription and translation rates can be deduced by measuring how noise changes with expression \(^{23,26,27}\).

Since most genes are regulated through multiple mechanisms, each of which can affect burst size and burst frequently differently, different genes should exhibit different relationships between mean expression and noise. However, measurements of a set of seven different promoters in \( E. coli \) all showed similar changes in expression and noise throughout induction \(^{35}\). Gene regulation in eukaryotes is more complex, and we thus hypothesize that burst frequency and burst size will be differentially regulated for each gene, and as a consequence, that the relationship between noise and expression will be different for different genes.

To characterize the relationship between mean expression and noise for native promoters in response to environmentally stimulated changes in TF activity, we generated a set of sixteen strains in which distinct promoters were fused upstream of a yellow fluorescent protein reporter (YFP). In each strain, we extracted a different Zap1 binding-site containing promoter from its native locus, integrated it into the \( his3 \) locus, and measured its expression and noise at 12 different zinc concentrations (induction levels). Decreasing zinc concentration increases the activity and expression of Zap1 and changes the expression of Zap1 target promoters. The resulting Zap1 dose-response curves of these targets show activation, repression, and a combination of activation and repression, consistent with previous observations \(^{36}\).

We found that for Zap1-activated targets, an increase in Zap1 causes an increase in expression and a decrease in noise. Similarly, Zap1-repressed targets exhibit the same relationship between expression and noise, whereby an increase in Zap1 causes a decrease in expression and an increase in noise. Despite this general trend that has previously been reported \(^{21-23}\), we found that the slope of noise versus expression is unique for each promoter, showing that noise is not determined by expression level alone. The most notable exception to expression determined noise is the ZRT2 promoter, which is both activated and repressed by Zap1 \(^{37}\), where we
found a different and novel relationship between mean expression level and the distribution of expression. We found that repression of ZRT2 by Zap1 results in a decrease in both expression and noise, leading to a transcriptional state of low expression and low noise that is unique among the sixteen tested promoters. This behavior is predicted by a kinetic model in which repression is due to a secondary binding event near the TATA that causes a decrease in transcription rate (burst size), thereby preventing the typical increase in noise that accompanies repression that is due to a reduction in burst frequency. These results suggest that the relationship between noise and expression is unique to each promoter and is determined by the regulatory mechanism encoded in the promoter DNA sequence and not by mean expression level alone.

Next, we hypothesized that further noise reduction will occur when activator and repressor are performed by the same TF. Using a model of noise that takes into account the sensitivity to TF level fluctuations and an experiment in which we decouple activator from repressor, we find strong evidence supporting our hypothesis that coupling between activator and repressor is a mechanism for noise reduction.

Finally, analysis of the data from all measured Zap1 targets brings forward a global principle of regulation in which the major source of differences in expression between promoters changes with induction. Our results strongly support a model in which at low Zap1 activity, differences in expression between Zap1 targets are due to variability in the frequency of transcriptional bursts, while at high Zap1 activity, differences are due to variability in the number of proteins produced during each transcriptional burst. This model suggests that such behavior is a general property of transcriptional regulation.

4.3 Results

4.3.1 Each target of a single transcription factor exhibits a unique gene-specific scaling of expression and noise in response to changes in TF activity

To study how expression of different native promoters is regulated by environmental-induced changes in TF activity, we measured promoter-driven expression in single cells for sixteen targets of the transcription factor Zap1 in response to changes in extracellular zinc. To do this we used an experimental system that we previously developed in which a promoter of interest drives YFP expression from the genomic his3 locus (Figure 4.1A) 38. We generated a set of sixteen promoter-YFP fusion strains and used flow-cytometry to perform quantitative single-cell measurements of promoter-driven expression at 12 induction points (Figure 4.1C). These promoters (Figure 4.1B) have diverse activation curves (Figure 4.1D, Figure 4.8) and, while the response of each promoter correlates with the predicted Zap1 occupancy along the promoter (Figure 4.9), the diversity of responses suggests that the way in which Zap1 alters expression is different for different promoters. In addition, we examined the
changes in noise and noise strength along the induction curves (Figure 4.1G). For most activated (11/13) and repressed (2/3) promoters, noise decreases as expression increases (Figure 4.1E, average Pearson correlation for all promoter of -0.73, Figure 4.10A), consistent with observed genome-wide trends 21-23. In contrast, noise strength changes less consistently across Zap1 targets (Figure 4.1, average Pearson of -0.09, Figure 4.10B). Surprisingly, not only do different promoters exhibit different amounts of noise at the same level of expression (Figure 4.1C), but also the way in which noise and noise strength change with expression is unique to each promoter (Figure 4.1E,G, Figure 4.11). Interestingly, a single promoter (ZRT2) that is both activated and repressed by Zap1 37 (Figure 4.1D lower right), shows very different amounts of noise at the same mean expression (Figure 4.1F). Because different molecular mechanisms of gene regulation can lead to the same change in mean expression but different changes in noise 16, these results suggest that the precise molecular mechanism by which a change in Zap1 activity causes a change in expression may be different at each promoter.
Promoter sequence determines the relationship between expression level and noise

Figure 4.1: Measuring mean promoter activity and cell-to-cell variability for a library of Zap1 target promoters. (a) The transcription factor Zap1 is induced by decreasing the concentration of zinc in the growth medium. A schematic of the site of chromosomal integration for measuring promoter-driven expression is shown. Each yeast strain has a single promoter inserted upstream of the YFP coding sequence. At the same locus a constitutively expressed mCherry is also integrated, which is used to normalize the YFP signal and correct for extrinsic cell-to-cell variability. (B) For each Zap1 target promoter the predicted locations of the major architectural features are shown. Promoters are aligned by the transcription start site (TSS) (cyan). PSSMs for the TATA box (purple) [36] and Zap1 (green) [17] were used to predict binding sites for TBP and Zap1, respectively. The width of the green bars is proportional to the predicted affinity of each.
4.3.2 A kinetic model of promoter switching replicates the experimentally observed changes in expression and noise for the ZRT1 promoter

To better understand what determines the relationship between expression and noise we used an analytical model of gene regulation (Figure 4.2A) 39 to predict changes in expression and noise in response to changes in TF activity (see Methods). We fit this model to measurements of ZRT1 expression and noise and find that the model replicates our experimental results when an increase Zap1 activity causes an increase in the promoter on-switching rate ($K_{on}$) (Figure 4.2B,C). To further challenge the model we created a set of seven start codon context mutants of the ZRT1 promoter (NNNNATG) and measured the expression distribution of these variants at twelve different levels of TF activity (Figure 4.2D-F)(only three are shown for clarity). These mutations to change translational efficiency and therefore the number of proteins produced per mRNA ($b$), without affecting promoter dynamics (Figure 4.12). We find that ATG context variants at a single induction point differ in expression but not in noise, consistent with similar experiments in E. coli 33. In support of the above hypothesis, we obtain the best fit of the model to our data when the TF induction is modeled as changing $K_{on}$ while the ATG context variants change $b$ (Figure 4.13 and Figure 4.14). Furthermore, when fitting our model to data we find that the optimal rate constants are on the order of experimentally measured promoter switching rates 28,40,41, and not TF binding/unbinding rates 42. This suggests that promoter-switching rates probably correlate with, and are partially determined by, TF concentration and binding kinetics. However, each TF binding event does not necessarily lead to transcription initiation. These results suggest that increases in TF activity increase the frequency of transcriptional bursts, while increases in translational efficiency cause an increase in the size (number of proteins produced) of each burst. We note that this

Zap1 binding site. Darker shades of grey show regions with higher predicted nucleosome occupancy. Blue lines show translation start sites. (C) Zap1 activates its own transcription, in addition to other target promoters, such as Zrt1. Shown is the measured expression (the ratio between YFP and mCherry fluorescence) of the ZAP1 promoter and the activated target ZRT1, graphed against the concentration of zinc added to the growth media. The inset shows the single-cell distribution of measured fluorescence intensities for ZAP1 and ZRT1 at two zinc levels obtained from flow-cytometry. (D) Measured promoter driven expression (quantified as the ratio between YFP and mCherry fluorescence) throughout the Zap1 induction is shown for each measured promoter. Each point shows the average of at least four biological replicates. (E & G) Noise and noise strength graphed against mean expression for each promoter that changes expression by more than two-fold. The line $\eta^{2}=\sigma^{2}$ was fit (solid lines) to the induction data per promoter, showing that different promoters show different scalings of noise and mean expression. (G) The measured expression distribution for the ZRT2 promoter at two different zinc induction levels (50.4uM and 648uM zinc, blue and red points in E & G) with the same mean expression level but different distributions. The mean expression level for each distribution is marked with a dashed line.
is in contrast to observations in *E. coli* and in yeast at the *GAL1* promoter, in which TF induction appears to change the promoter off-switching rate (Koff) but consistent with measurements of the *PHO5* promoter.

Figure 4.2: Measured and modeled gene expression of ZRT1. (a) ZRT1 expression is modeled with a kinetic scheme in which the promoter switches between a transcriptionally active (on) and inactive state (off) as a result of Zap1 (red oval) binding and unbinding. (b) Experimentally measured ZRT1 promoter driven expression changes as a function of zinc concentration (triangles). The kinetic model in (a) fits (line) the data (triangles) when zinc is assumed to change Kon (inset). (c) Noise graphed as a function of expression for the data and model from (b). (d) A schematic of the experimental system used to change translation efficiency through mutations of the ATG context. (e) Measured expression distributions for two ATG context variants at three zinc induction levels shows that changing expression via induction or ATG context has a different effect on the shape of the expression distribution. Measured (f, squares) and fit (f, solid lines) of noise as a function of mean expression for three ZRT1 promoter mutants (f, colors) that each has a unique 4 base-pair sequence immediately upstream of the ATG. A model (f, solid lines) in which the only difference between ATG context variants (different colors) is in the number of proteins produced per mRNA (f) fits the experimental data (squares) better than any alternative model.
4.3.3 Repression of ADH1 and ADH3 by Zap1 is likely due to a decrease in the frequency of transcriptional bursts

In addition to increasing expression of target genes, Zap1 can also act as a repressor. Zap1 represses two targets (ADH1 and ADH3) by binding upstream of the core promoter and inducing intergenic transcription through the core promoter, probably promoting dissociation of the activating transcription factor Rap1. Two mechanisms have been proposed for repression by transcriptional interference: dislodgement of TFs and the Pol II pre-initiation complex by RNA Polymerase, and competitive binding, one form of which is deposition of nucleosomes in the otherwise nucleosome free region where the activating TFs and Pol II bind. We hypothesized that deposition of nucleosomes would result in occlusion of the activating binding site, the TATA box, and PolII binding, thus reducing the effective TF concentration and lowering the frequency of transcriptional activation. We model this mechanism as a reduction in $K_{on}$. Alternatively, passage of RNA polymerase may dislodge already bound Rap1, TBP, and/or the RNA polymerase pre-Initiation Complex. This would shift the promoter from the ‘on’ into the ‘off’ state, thus reducing the length of each transcriptional on state and therefore the number of mRNA molecules produced during each transcriptional burst (Figure 4.3C). We model this mechanism as an increase in $K_{off}$. To determine the ability of dislodgement by Pol II (TD) or occlusion of TF binding by nucleosomes (NO) to explain our experimental data we fit each model to our data. We find that the NO model fits our data better than the TD model (Figure 4.3D) (see Methods). Furthermore, the NO model consistently fits the data better in the case in which we vary each parameter by up to two-fold. The increased robustness (Figure 4.3E) and decreased sensitivity (Figure 4.15) of the NO model gives us further reason to favor a model in which repression by Zap1 at the ADH1 and ADH3 promoters occurs by inducing intergenic transcription and nucleosome deposition over the core promoter and/or Rap1 binding site.
Promoter sequence determines the relationship between expression level and noise

**Figure 4.3:** Measured and modeled gene expression for ADH1. We model ADH1 expression using a two-state kinetic scheme (a) in which Kon and Koff are determined by the binding of transcriptional activators (blue circle) or a repressor (red circle). (b) Two mechanisms have been proposed for repression by upstream interfering transcription: TF dislodgment, in which an alternative transcript dislodges the bound activator, and nucleosome occlusion, where transcription through the promoter results in an occluding nucleosome that prevents binding of the activator. Hence, we assume that TF dislodgment increases the dissociation rate of the activator and that nucleosome occlusion results in a decrease in the binding rate of the activator. (c) We fit the model such that either Kon (black) for nucleosome occlusion or Koff (blue) for TF dislodgment changes as a function of [zinc]. (d) Measured mean expression versus noise (triangles) and fits (lines) of both model variants show that the nucleosome occlusion model has a better fit to the data (Δ is distance of fit to data). (e) To compare the robustness of each model, each parameter was independently perturbed 50 times over a two-fold change from the fit value, and the distance of each model to the data was computed. Shown are the cumulative distributions of these distances. The narrower distribution of the nucleosome occlusion model (black) shows that it is significantly more robust to parameter variation than the TF dislodgment model (blue).
4.3.4 ZRT2 achieves a state of low expression and low noise due to a repression mediated mechanism of intrinsic noise reduction

Uniquely among Zap1 target promoters, ZRT2 responds non-monotonically to an increase in Zap1 activity, whereby its expression first increases then decreases in response to increasing Zap1 activity 37. In the activating regime of ZRT2, noise decreases as expression increases, suggesting a Kon (burst frequency) dominated change that is similar to the purely activated targets. However, in contrast to the repressed targets ADH1 and ADH3, where noise increases with the decrease in expression, in the regime where ZRT2 expression decreases noise remains constant. These results suggest that the decrease in ZRT2 expression is a result of a decrease in burst size (see below), with the consequence of having induction points that have the same mean expression level but different expression distributions (Figure 4.1F). At high induction, the distribution is less noisy (Figure 4.1F, blue) than at low induction (Figure 4.1F, red). Thus, the ZRT2 promoter reaches a state that is unique amongst Zap1 targets that is characterized by both low expression and low noise. Taken together, these findings suggest that although ADH1, ADH3 and ZRT2 are all repressed by Zap1, the mechanism by which ZRT2 is repressed is unique.

4.3.5 A Zap1 binding site near the TATA box is both necessary and sufficient for repression through Zap1 mediated burst size reduction

In response to increasing Zap1, ZRT2 expression first increases and then decreases. The activation by Zap1 is a result of Zap1 binding at activating binding sites 250-300bp upstream of the start codon, while the repression is due to the presence of repressive Zap1 binding sites near the TATA box (between -90 and -112) 37. In addition we made a variant (ZRT2-zre) of the ZRT2 promoter that lacks the repressive binding sites (Figure 4.4B). We hypothesized that a model of the ZRT2 promoter should include promoter states in which Zap1 is bound as an activator, as a repressor and both as activator and repressor (Figure 4.4A). Based on experimental evidence 37, we model the binding site affinity for the repressive site as weaker than that of the activating site. We assume that binding of Zap1 to the repressive site turns off the promoter but does not affect the transition probabilities between states. When the model was simultaneously fit to both the ZRT2-WT and ZRT2-zre experimental data, we find that the model obtains a good fit to data when, like with ZRT1, an increase in Zap1 activity increases Kon and does not affect any other parameters. Interestingly, we find that the repressed state (state 4, Figure 4.4A) is not fully off, but has a small, but not insignificant, transcription rate relative to the transcription rate of the active state (state 2, Figure 4.4A). Notably the only parameter change required to change from ZRT2-WT to ZRT2-zre is setting Koﬀ to be very high, mimicking the mutation of the repressive binding sites (Figure 4.4C,D). These experimental and modeling results suggest that binding of the transcriptional activator Zap1 to a binding site between the TATA box and TSS is necessary to generate a promoter state with low transcriptional activity. Notably, a very simple promoter model is able to replicate a
Promoter sequence determines the relationship between expression level and noise

Promoter sequence determines the relationship between expression level and noise. Furthermore, it suggests that the reason that the ZRT2 promoter is able to reach a state of low expression and low noise purely through transcriptional regulation is due to a promoter state with high burst frequency (due to binding of activating Zap1) and low burst size (due to binding of repressive Zap1).

These results suggest that in the ZRT2 promoter, an increase in Zap1 both increases the frequency and decreases the size of transcriptional bursts. Therefore, our simple kinetic model shows that adding a repressive binding site for the activating TF is sufficient for explaining both ZRT2 expression and noise as a function of induction.

Repression of ZRT2 is accompanied by a decrease in noise strength, suggesting that repression occurs via a decrease in burst size. We therefore hypothesized that addition of a repressive Zap1 binding site to a native Zap1 target that lacks repression would cause a decrease in expression and burst size. To test this hypothesis, we added a consensus Zap1 binding site (ACCTTAAGGT) upstream of the transcription start site of ZRT1 (Figure 4.4E, ZRT1pr+ZRE). Consistent with our hypothesis that this repressive site reduces expression through a decrease in burst size, this additional site results in a constant ~2 fold decrease in expression, a decrease in noise strength and no change in noise (Figure 4.4F). A model identical to the ZRT2 model (Figure 4.4A), except that the repressive site has a higher affinity to Zap1 than the activating site, replicates the experimental data (Figure 4.4F). Interestingly, we find that while both models require the repressed state to be partially active, the repressed state of the ZRT1 promoter has higher activity (in model and data) than for the ZRT2 promoter. This may be because ZRT2 has at least two repressive Zap1 binding sites, while we only introduced a single repressive binding site into ZRT1. Nevertheless, these results show that the presence of a Zap1 binding site between the TATA box and transcription start site is both necessary and sufficient for repression mediated by a decrease in burst size.
Figure 4.4: A repressive Zap1 binding site is both necessary and sufficient for repression in ZRT2. (a) We model ZRT2 expression with a 4-state kinetic scheme that represents four promoter configurations as a result of binding and unbinding of Zap1 to two different binding sites. One binding site is activating (blue square) the other repressing (purple square) and as a result we assume that each configuration can have different transcriptional activity (see Methods for a detailed description of the model). (b) Promoter architectures are shown in terms of Zap1 binding sites (green), TATA box (purple), TSS (light blue) and nucleosome occupancy (white to grey for increasing occupancy) for wild-type ZRT2 and a ZRT2 mutant (-zre) in which the repressive Zap1 binding site was removed (at the arrow). (c) Measured (triangles and squares) and modeled (lines) mean expression as a function of [zinc] for wild-type ZRT2 (black) and the -zre mutant (blue). (d) The same measured data and model from (c) are shown for mean expression versus noise. The ZRT2 model was simultaneously fitted to the wild-type (c,d, black line) and the mutant (c,d, blue line) with the assumption that the only difference between wild-type and mutant is that the Kd of the mutant is infinite, to model the removal of the repressive binding site. Intrinsic noise (d, inset) measured in a dual reporter assay shows the same mean to noise scaling. (e) The promoter architectures are shown for the wild-type ZRT1 promoter and a +zre mutant in which a repressive Zap1 binding sites was added around the TSS/TATA (at the arrow). (f) Measured mean expression and noise for the ZRT1 wild-type (green circles) and the +zre mutant (red triangles), and mean expression versus noise strength (inset). The ZRT2 model was simultaneously fitted to both wild-type ZRT1 (green line) and +zre
Promoter sequence determines the relationship between expression level and noise

4.3.6 Mutation of additional repressive Zap1 binding sites suggests that a combination of activation and repression may be common

A computational search for Zap1 binding sites between the TATA box and the transcription start site identified three weak Zap1 binding sites in the ZRT3 promoter (Figure 4.5A). A closer look at the ZRT3 induction curve at very low zinc concentrations showed that expression of ZRT3 decreases slightly at high Zap1 induction (Figure 4.5B, inset). To determine if these weak Zap1 sites were functional, we mutated them and measured expression of the wild-type and mutant ZRT3 promoters. Consistent with our hypothesis that Zap1 binding sites around the TSS are repressive, removal of the presumptive Zap1 binding sites increased expression (Figure 4.5B), in particular at higher induction, consistent with our model in which repression is a function of repressor activity. This suggests that low-affinity binding sites may be functional at high TF concentration, perhaps mostly at promoters that have additional high-affinity binding sites.

Figure 4.5: Removal of a predicted repressive Zap1 binding site increases expression of ZRT3. (a) Promoter architectures are shown for wild-type ZRT3 (wt) and a ZRT3 mutant (-zre) in which a potential repressive Zap1 binding site was removed (at the arrow).
Shown are Zap1 binding sites (green), TATA box (purple), TSS (light blue) and nucleosome occupancy (white to grey for increasing occupancy). The potential repressive binding site was predicted using a bioinformatics search. (b) Consistent with this prediction mean expression is higher for the -are mutant (blue) compared to the wild-type (black). The difference in expression appears only to exist at higher induction, consistent with the idea that repression is a function of Zap1 induction. Further induction of wild-type ZRT3, at very low zinc levels (inset), appears to decrease expression, consistent with a ZRT2-type repressive mechanism in which expression first goes up and then down with increasing TF levels.

4.3.7 Activation and repression by the same TF as a mechanism for reduction of extrinsic noise due to fluctuating TF levels

The ZRT2 promoter presents a case in which the activator and repressor are the same TF. We were intrigued by this mechanism and wondered whether this affects the noise properties. Many promoters in yeast are regulated by the binding of both activators and repressors to different binding sites in the promoter. The activator and repressor can be different proteins (e.g., ADH1 is activated by Gcr1 and Rap1 and repressed by Zap1) or the same protein (such as ZRT2 that is both activated and repressed by Zap1) (Figure 4.6A). We hypothesized that the sensitivity to TF fluctuations for a promoter that is both activated and repressed depends on the coupling between activator and repressor. For example, we expect that when activator and repressor are done by the same TF, in a regime where a change in activator binding has the exact opposite result on expression as the same change in repressor binding, the promoter is insensitive to any fluctuations in TF levels. To study this hypothesized phenomenon, we used our kinetic model of ZRT2 and simulated the case where activator and repressor are different (decoupled) and where they are the same TF (coupled). We then calculated the contribution of TF fluctuations to expression noise throughout the induction (Figure 4.6B) (See section Materials and Methods for a detailed description of the model). Coupling of the activator and repressor reduces the sensitivity to TF fluctuations throughout induction and places the point of minimal sensitivity to TF fluctuations at the point of maximum target gene expression (Figure 4.6B blue line).

Our model predicts that the total sensitivity to TF fluctuations is reduced throughout the induction curve, and that this reduction is greatest at maximal promoter expression (Figure 4.6B, point 1). To test this we measured extrinsic noise (the contribution of variance in all factors, e.g. ribosomes, Zap1, PolII) for the native ZRT2 promoter using a dual-reporter. We find that extrinsic noise is constant across the induction (Figure 4.6D, purple). However, when we remove as much global extrinsic noise as possible using a very narrow forward and side scatter gate (Figure 4.1E) we hypothesize that we are left with mostly pathway-specific noise, e.g. noise due to TF level fluctuations. In support of this hypothesis, we find that pathway-specific noise is not constant, but rather varies greatly (around 10 fold) with induction. We find that this signal, which we expect to be dominated by
changes in TF sensitivity, does indeed drop around the point of maximal expression (Figure 4.6D, blue), consistent with our model. In fact, the extrinsic noise replicates quite well the general predicted change in TF sensitivity with induction.

Finally, our model predicts that decoupling of activator and repressor will increase total noise as the sensitivity to TF fluctuations is increased. To test this we replaced the two activating Zap1 binding sites of ZRT2 with two Gal4 binding sites (Figure 4.6A) and measured expression and noise throughout the repressive regime (at high Gal4 induction as to reduce its noise and remove any bias for comparison with native ZRT2). Consistent with our model, the Gal4-Zap1 regulated ZRT2pr variant has higher noise than the wild-type promoter (Figure 4.6C). These results show that, while repression is able to reduce expression and keep noise constant a transcriptional regulatory motif, in which the activator and repressor are the same protein, is capable of reducing noise even further. This suggests that the coupling of activator and repressor can be a mechanism to regulate gene expression with less variability.

Figure 4.6: Activation and repression by the same TF as a mechanism for noise reduction. (a) A promoter that is both activated and repressed can be regulated by two different TFs (decoupled, e.g. Gal4-act and Zap1-rep) or one TF (coupled, e.g. Zap1) that functions as both an activator and repressor. (b) A simulation of noise as a result of fluctuations in TF concentration is shown for a coupled (blue) and decoupled (red)
The Y-axis shows noise as a result of TF fluctuations as a function of promoter induction (mean on-switching rate, $K_{on}$) for the coupled (blue) and decoupled (red) system. In addition, the mean expression at each induction level is shown (dashed line). Noise from TF fluctuations was quantified by sampling the model at different TF concentrations (i.e. $K_{on}$ values) that were drawn from a gamma distribution (See Methods for a detailed description of the model). The model predicts that coupling of activator and repressor (e.g. if they are the same molecule) reduces noise. Notably, reduction is maximal where mean expression peaks (arrow 1). (c) Noise measurements, at various zinc induction levels, of native ZRT2 (blue) and a mutant that has two Gal4 UASs upstream of a repressive Zap1 site (red). The coupled system (wild-type Zrt2) has consistently lower noise than the decoupled system (Gal4-act Zap1-repr), as is predicted by our model. (d) Measurement of extrinsic noise from a dual-reporter assay is shown as a function of zinc induction. Non-stringent gating on cell-size (through forward and side scatter) shows an extrinsic noise that is constant with induction (purple). However, strict gating (through a small forward and side scatter gate) significantly reduces the extrinsic noise and reveals a signal that changes with zinc induction (blue). We hypothesize that this signal is determined by noise from TF fluctuations, which according to our model has specific behavior as a function of induction. As predicted by our model we find a reduced noise where mean expression (dashed line) is maximal and sensitivity to TF changes is minimal (b,d, arrow 1), and minimal reduction (maximal extrinsic noise) where mean expression is most sensitive to changes in TF concentration (b,d, arrow 2).

**4.3.8 The dominant source of differences between promoters in expression and noise changes with TF concentration**

We hypothesized that the source of differences in expression between genes might change with TF concentration. At low TF concentrations, promoters will be inactive most of the time, and differences in expression may depend mostly on differential recruitment of the TF. In this case, the major source of differences in expression between promoters should stem from the frequency with which transcriptional bursts occur. Alternatively, at saturating concentrations of activating TF, the promoter should be ‘on’ most of the time and the major difference in expression between promoters should arise from the transcription and translation rates of each promoter. Thus, as the concentration of TF changes from negligible to saturating, we expect the transcription and translation rates of each promoter to become more important in determining expression differences between genes.

To determine whether burst frequency or burst size dominate the differences in expression between promoters, for each induction level, we measured the correlation between expression and noise or noise strength across promoters. Consistent with the above hypothesis, across all promoters, noise is highly correlated with expression at low levels of Zap1 activity ($R=-0.66$, $P<0.01$), while noise strength is uncorrelated ($R=-0.02$, $P<0.94$) (Figure 4.7A). This suggests that at low TF concentration, burst frequency determines the differences in expression across promoters. Conversely, at high levels of Zap1 activity, noise strength is
Promoter sequence determines the relationship between expression level and noise

correlated with expression (R=0.63, p=0.01), and noise is slightly less correlated
(R=0.55, p=0.04) (Figure 4.7A). Overall, we found a continual increase in the
correlation between noise strength and expression with increasing TF activity (data
not shown). To test the hypothesis that these differences are due to a change in the
dominant source of expression difference between promoters, we generated 50
random genes in-silico that differ only in their rates of promoter on-switching ($K_{on}$)
and translation ($K_{tl}$). We then performed an induction by increasing $K_{on}$ for each
promoter to 20 times its original value. This results in a mean to noise and mean to
noise strength scaling that is strikingly similar to what we observed for the native
Zap1 targets (Figure 4.7B). Taken together, our results suggest that as a set of
targets of the same TF are induced, the major source of expression differences
between them changes from being dominated by burst frequency to a combination
of burst frequency and burst size.

Figure 4.7: The correlation of noise and noise strength with expression changes with TF
concentration. (A) Scatter plots of noise (top) and noise strength (bottom) graphed
against expression for each promoter at low (left side) and high (right side) Zap1
induction points. A line fit to each set of points using linear regression shows that, across
promoters, noise strength is uncorrelated with expression at low TF concentration, but
is positively correlated with expression at high TF concentration. (B) Noise and noise strength graphed against expression for high and low TF as in (A) but for in-silico promoters that differ in both KON and KTL. The change from low to high TF was simulated by multiplying the initial KON of each promoter by 20.

4.4 Discussion

We have measured the dose response curve, in terms of expression and noise, for a set of native yeast promoters that are all targets of the same TF, yet are regulated by that TF via at least three distinct transcriptional mechanisms: activation, repression by binding between the TATA box and TSS, and repression by induction of an upstream interfering transcript. Although noise generally decreases with increased expression, the quantitative scaling of noise with expression is specific to each promoter and depends on the mechanism by which the TF regulates the promoter.

4.4.1 The promoter sequence determines how activation by Zap1 affects noise and expression

As in the global trend seen in expression 22, our data suggests that changes in expression of individual promoters are dominated by differences in burst frequency. This is consistent with Zap1 binding to promoters being limiting for transcriptional activation, especially at low Zap1 concentrations, and with the proposal that the rate-limiting step in transcription for yeast is promoter firing-rate, which is determined by TF search times49. However, the observation that different activated targets have different scaling between noise and expression suggests that while activation by Zap1 acts only through burst frequency at most activated promoters, it may act partially or even completely through burst size at other activated promoters. This is entirely reasonable; Zap1 is not the only TF acting at these promoters, and the promoters differ in both nucleosome organization and the presence and location of TATA boxes. Experiments that placed a tetO sequence at different locations within the FLO11 promoter suggest that the same TF can have different effects on promoter dynamics, depending on the location of binding sites within the promoter41. Unfortunately, there are not enough strongly induced Zap1 targets in S. cerevisiae to identify the promoter architecture features that determine the source of the promoter-specific slope. It will be interesting to perform dose-response curves for a larger set of promoters from other yeasts, or on synthetic promoters, in order to identify promoter architecture that determine the promoter-specific slope.

4.4.2 Different mechanisms of regulation by the same TF can cause similar changes in expression but different changes in noise

Our observation that repression by production of an upstream interfering transcript causes an increase in noise, while repression when the TF binds near the TATA box causes a decrease in noise, suggests that different dynamics occur at each
promoter during repression. This, along with previous observations, suggests that the mechanism of regulation by any TF is determined in cis by the promoter architecture. Binding sites between the TATA box and TSS decrease burst size, binding sites within a few hundred bases upstream of the TATA box increase burst frequency, and binding sites further upstream, with a nearby downstream TATA box, repress through a reduction in burst frequency. These data show for the first time that different promoter architectures can cause a similar change in expression in response to changes in TF activity, but exhibit different changes in noise.

4.4.3 High burst frequency and low burst size is a strategy to produce low abundance proteins with low noise
If the genome-wide scaling of expression and noise extends to proteins with very low expression, then a large fraction of cells will have zero molecules of protein. Single-molecule studies have confirmed this: many cells have zero molecules of proteins with low levels of expression. However, many proteins expressed at low levels are essential. This raises the question: how does the cell maintain a low level of both expression and noise for essential proteins, so that all cells have the minimum number of proteins. Our results showing that burst size regulation can reduce expression without increasing noise suggest a way out of this trap. Lowly expressed genes tend to be bound by many transcriptional regulators, both activators and repressors. Low levels of an activating TF result in low expression and high noise. Notably, a motif in which weak transcription but efficient translation generates high noise may exist at the comK gene in B. subtilis. In contrast, combinatorial regulation that results in high burst frequency and low burst size (approaching the Poisson limit) provides a regulatory motif through which cells can produce low levels of protein with low cell-to-cell variability. Our identification of this same regulatory motif in the ZRT3 promoter suggests that this motif may be common. This regulatory strategy may be used to prevent some cells from having zero molecules of protein when expression is low.

4.4.4 Coupling of activator and repressor as a mechanism for reducing extrinsic noise
The concentrations of TFs, like those of all other proteins, vary greatly from cell to cell. We therefore expect that these variations have a significant impact on the cell-to-cell variability of target gene expression, and therefore wondered how cells deal with this source of noise. Interestingly, we find that ZRT2 is able to reduce noise through its reduced sensitivity to fluctuations in TF levels, as a result of activator and repressor being the same molecule. Mechanisms for extrinsic noise reduction have been previously reported. However, to the best of our knowledge, we are the first to propose theoretically and confirm experimentally a mechanism for desensitizing promoters to TF noise. Noise as a result of TF fluctuations has been proposed theoretically in several studies. In fact, Bai et al. propose a dual-reporter experiment to investigate extrinsic noise resulting from TF fluctuations, which we have performed in this work (Figure 4.6D). We note that noise from TF fluctuations is a special case of noise propagation in a gene network, where the
noise of a downstream gene is a function of its intrinsic noise and the noise from any upstream genes\textsuperscript{54}. An alternative mechanism for a similar reduction in sensitivity would be the regulation by multiple different decoupled TFs. We hypothesize that as the number of different TFs increases, target sensitivity (and therefore noise) decreases, if the TFs are sufficiently de-correlated. This potential mechanism, as well as the general characterization of the effect of TF noise on target noise, would make the subject of a meaningful follow up study.

4.4.5 The dominant type of noise changes with TF concentration
The observed change in the scaling between noise and expression throughout the increase in TF concentration suggests that variability between promoters in burst size (transcription efficiency, translation efficiency, and promoter off-switching rate) becomes more important as TF concentration is increased. This suggests that differences in promoter architecture play different roles at low and high TF concentrations. In the presence of limiting TF, promoter architecture may determine expression by determining TF search time, through the number of accessible TF binding sites. However, at high TF concentration, promoters are mostly bound by TFs, and the transcription and translation efficiency of each gene may play a greater role in determining expression. This idea is supported by the positive correlation between noise strength and expression at high TF concentration, as would be expected from theory\textsuperscript{16}. In addition, differences in burst frequency cannot account for the measured single-cell expression distributions at high TF concentration. These data suggest that the dominant sources of gene-to-gene variability in expression change with TF concentration: at low TF concentration burst frequency (the ability of the promoter to recruit TF) differences dominate, whereas at high TF concentration burst size (transcriptional and translational efficiency) differences dominate.

Overall, our results show that the relationship between expression and noise is highly dependent on the promoter architecture. One implication of this finding is that using only a single TF, evolution can implement diverse expression profiles with unique noise properties. The fact that repression of \textit{ZRT2} by Zap1 is evolutionarily conserved suggests that there is an advantage to this ability.

4.5 Materials and Methods
4.5.1 Yeast strains
Construction of promoter-YFP strains was performed as described previously\textsuperscript{38}. In brief, a master strain, \textit{his3}:TEF2pr-mCherry-YFP-NatMX4 was created in the background strain Y8205\textsuperscript{55} by homologous recombination. Each promoter-YFP strain was created by integration of a PCR product containing the native promoter along with \textit{URA3} as a selection marker. Integration by homologous recombination upstream of YFP was confirmed by DNA sequencing and by identical expression and growth of multiple independent transformants.
4.5.2 Creation of promoter variants
To introduce, alter and remove elements within ~150bp of the ATG we developed a method in which an existing URA3-promoter-YFP cassette is amplified over multiple rounds of PCR. In each subsequent round a new primer is used that further extends the product towards the YFP and optionally introduces designed mutations. Thus multiple site-directed mutations can be tiled onto the 3’ end of the promoter. All promoter variants were confirmed by DNA sequencing.

4.5.3 Yeast growth and expression and noise measurements
Yeast strains were grown overnight to saturation in YPD, resuspended in low zinc medium 37, and grown overnight to saturation in media lacking zinc. Cultures were then diluted 1:40 in water and 6ul of this dilution was inoculated in 130ul of low zinc media supplemented with various concentrations of zinc. Cells were grown in round-bottom 96well plates shaking at 30°C a minimum of 12 hours, to approximately 5*10^6 cells / ml prior to expression measurements. Flow cytometry was performed on a BD LSRII. YFP and mCherry were excited using 488nm and 561nm lasers and emitted light was collected with 525/50nm and 610/20nm band-pass filters, respectively. There is no detectable spillover of YFP or mCherry into the other channel using these filters and lasers. Expression and noise measurements were collected calculated using the ratio of YFP over mCherry for each cell. To obtain expression and noise measurements from each well a relatively homogenous subpopulation of mostly G1 cells was chosen by gating on forward and side scattering. Wells containing fewer than 500 cells after gating or with obvious contamination were excluded from further analysis. Noise was quantified as the variance over the mean squared and noise strength as the variance over the mean.

4.5.4 Kinetic model of activation and repression
We model stochastic promoter state switching, transcription and translation using the master equation following the approach described by Sanchez et al. 56 which in turn is an adaptation of previous derivations of the master equation for gene regulation 56-58.

In this description of promoter regulation, transcription factor (TF) binding and unbinding events determine the transitions between promoter states. A change in transcriptional activity occurs when a transition is made to a state with differing transcription rate. Each promoter state is modeled to have a low (including zero) or relatively high transcription rate to describe in-active (“off”) or active (“on”) states respectively. Translation occurs in bursts with the probability of a burst described by a geometric distribution. The master equation (in matrix notation) takes the form:

\[
\frac{d}{dt} \hat{p}(n) = \left[ \hat{K} - \frac{b}{1+b} \hat{R} - n\hat{I} \right] \hat{p}(n) + \hat{R} \sum_{\beta=1}^{n} h(\beta) \hat{p}(n-\beta) + \hat{R} \delta I \hat{p}(n+1) \tag{1}
\]
Where \( \tilde{P} \) is the vector of probabilities of having \( n \) proteins in the cell for each promoter state. \( \frac{d}{dt} \tilde{p}(n) \) describes the time evolution of these probabilities. \( \hat{K} \) is the matrix of promoter state transition rates, where \( \hat{K}_{ij} \) is the rate of transitioning from state \( j \) to state \( i \) and \( \hat{K}_{ii} \) is the sum over all outgoing rates from \( i \) times -1. \( \hat{R} \) is the diagonal matrix of transcription rates with \( \hat{r}_i \) on the diagonal (\( \hat{R}_{ii} = \hat{r}_i \)), where \( \hat{r}_i \) is the transcription rate of state \( i \). \( \hat{I} \) is the identity matrix. \( b \) is the average burst size (proteins produced per mRNA). \( \delta \) is the protein degradation rate. \( h(\beta) \) describes a geometric distribution and is the probability of producing a burst of size \( \beta \).

To derive the mean protein abundance and variance we solve this system at steady state, thus for \( \frac{d}{dt} \tilde{p}(n) = 0 \). We get mean protein abundance:

\[
\langle n \rangle = \frac{b \hat{r} \tilde{m}_{(0)}}{\delta}
\]  

(2)

Where \( \tilde{m}_{(0)} \) is the zeroth partial moment of the distribution of mRNA abundance and is the solution to:

\[
0 = \hat{K} \tilde{m}_{(0)}
\]  

(3)

We can get noise \( (\sigma^2/\mu^2) \) and noise strength \( (\sigma^2/\mu) \) by deriving:

\[
\langle n^2 \rangle = (1 + b) \langle n \rangle + \frac{b \hat{r} \tilde{n}_{(1)}}{\delta}
\]  

(4)

Where \( \tilde{n}_{(1)} \) is the first partial moment of the distribution of protein abundance and is the solution to:

\[
0 = (\hat{K} - \delta \hat{I}) \tilde{n}_{(1)} + b \hat{R} \tilde{m}_{(0)}
\]  

(5)

Variance \( (\sigma^2) \) is:
Promoter sequence determines the relationship between expression level and noise

\[ \text{Var}(n) = \langle n^2 \rangle - \langle n \rangle^2 \]  

(6)

Therefore noise \((\sigma^2/\mu^2)\) becomes:

\[ \langle \eta^2 \rangle = \frac{(1+b)\langle n \rangle - \langle n \rangle^2 + \frac{b \bar{\rho} \bar{n}_{(i)}}{\delta}}{\langle n \rangle^2} \]  

(7)

And noise strength \((\sigma^2/\mu)\):

\[ \langle F \rangle = \frac{(1+b)\langle n \rangle - \langle n \rangle^2 + \frac{b \bar{\rho} \bar{n}_{(i)}}{\delta}}{\langle n \rangle} \]  

(8)

We solve the master equation for a number of different promoter architectures, where we define \(\hat{K}\) and \(\hat{R}\) for each system to describe the specific promoter states, the transitions between them and the transcriptional activity of each state.

Case ZRT1:

We describe gene expression and regulation of ZRT1 using a two state kinetic scheme that represents switching between an active (ON) and inactive (OFF) promoter configuration. We assume that on switching (with rate \(K_{on}\)) and off switching (with rate \(K_{off}\)) are a function of binding and unbinding respectively of the Zap1 transcriptional activator at the ZRT1 promoter. The Zap1 bound (ON) state is transcriptionally active (with rate \(r1\)) and we allow the unbound (OFF) state to have some (leaky) transcriptional activity (with rate \(r2\)). \(\hat{K}\) and \(\bar{r}\) thus become:

\[
\hat{K} = \begin{pmatrix}
-K_{off} & K_{on} \\
K_{off} & -K_{on}
\end{pmatrix} \]  

(9)

\[
\bar{r} = \begin{pmatrix}
r1 \\
r2
\end{pmatrix} \]  

(10)
The promoter off rate (Koff) is a function of Zap1 binding affinity and therefore assumed to be constant. We describe the promoter on rate (Kon) with a Hill-equation (see Eq.11) with Kmin and Kmax as the minimum and maximum possible rates respectively, [Zn] as the zinc concentration (induction level), [Zn]mid as the zinc concentration that gives half maximal induction and H as the Hill-coefficient, i.e. the sensitivity of Kon to changing zinc.

\[
Kon = K_{min} + \frac{(K_{max} - K_{min})}{1 + ([Zn]_{ind})^H}
\]  

(11)

Case ZRT2:

ZRT2 is both activated and repressed by Zap1. We model Zap1 binding (Kon), unbinding at the activating binding site (Koffact) and unbinding at the repressive binding site (Koffrep). We describe Kon with a Hill-equation as a function of [Zinc] (see Eq.11). Transcription can occur from each state (i.e. expression is leaky), however we assume that the state where the activator, and not the repressor, is bound has the highest transcriptional activity (with rate r1). \( \hat{K} \) and \( \vec{r} \) therefore become:

\[
\hat{K} = \begin{pmatrix}
-(Koff^{\text{act}} + Kon) & Kon & 0 & Kon^{\text{act}} \\
Koff^{\text{rep}} & -2Kon & Kon^{\text{rep}} & 0 \\
0 & Kon & -(Kon + Koff^{\text{rep}}) & Kon^{\text{rep}} \\
Kon & 0 & Kon & -(Koff^{\text{act}} + Koff^{\text{rep}})
\end{pmatrix}
\]

(12)

\[
\vec{r} = \begin{pmatrix}
r1 \\
r2 \\
r3 \\
r4
\end{pmatrix}
\]

(13)

Case ADH1:

ADH1 is activated by Rap1 and repressed by Zap1. We model the switching between the active (Rap1 bound) and repressed (Zap1 bound) states with the two-state kinetic scheme that we used to model ZRT1 activation by Zap1. Repression of ADH1 by Zap1 occurs through (one of) two hypothesized mechanisms: nucleosome occlusion and transcription factor dislodgement. We model both mechanisms by a subtle difference in the dynamics of repression.

Case ADH1, nucleosome occlusion:

Zap1 mediated intergenic transcription may repress ADH1 by causing nucleosome deposition in the otherwise nucleosome free core promoter region, thus preventing
the activator (Rap1) or PolII from binding. We model this mechanism by changing $K_{on}$ as a function of induction, as occlusion (accessibility) effectively changes the on-switching rate of the promoter. More specifically we model $K_{on}$ as a sigmoid that is a function of [Zinc] (see Eq.14), where increasing zinc increases $K_{on}$ (as Zap1 decreases) while $K_{off}$ is constant.

$$K_{on} = K_{min} + \frac{(K_{max} - K_{min})}{1 + (\frac{[Zn]}{[Zn]_{mid}})^n}$$ (14)

Case ADH1, transcription factor dislodgement:

In the hypothesized TF dislodgement mechanism repression occurs as the interfering transcript (caused by upstream Zap1 binding) dislodges the already bound activator (Rap1) or the PolII holoenzyme. This would effectively change the rate at which the promoter switches from on into the off state, hence we model this by changing $K_{off}$ as a (sigmoidal) function of [Zinc] (see Eq.15), while keeping $K_{on}$ constant.

$$K_{off} = K_{min} + \frac{(K_{max} - K_{min})}{1 + (\frac{[Zn]}{[Zn]_{mid}})^n}$$ (15)

4.5.5 Model fitting and robustness analysis

We fit the kinetic scheme’s analytical solutions of mean and noise of protein abundance to the measured mean and noise of fluorescence intensity (See supplementary information for a detailed description of the fitting procedure and parameter constraint). The goodness of fit is measured by the root mean squared error (distance, Δ) of both mean and noise.

To investigate the hypothesized effect of promoter mutations we simultaneously fit the model to wild type and mutant promoters while only one parameter is allowed to change between the fits.

We distinguish between two hypothesized ADH1 regulatory mechanisms by fitting two models to the measured data. While ADH1 nucleosome occlusion gives a better fit than TF dislodgement, both models give have a good fit to the data. To investigate if nucleosome occlusion is a significantly better fit to the data, for each fit found by optimization we perform 2-fold perturbations on each parameter. By looking at the distribution of fits after perturbation we get an idea of which model is more robust and as a result is more likely to be the correct model. We find that the NO model is significantly more robust than the TD model.
4.5.6 Sensitivity analysis of the kinetic model

To measure the sensitivity of the kinetic model to variations in each parameter, we performed a rigorous sensitivity analysis procedure described by Marino et al. that uses the Latin Hypercube Sampling based Partial Rank Correlation Coefficient (LHS-PRCC). First, we uniformly sampled 10,000 instances of the model (without fitting), each with a unique parameter setting, sampled from the entire allowed parameter space using Latin Hypercube sampling, and evaluated each of these models by measuring the distance to the experimental data. Next, we calculate the Partial Rank Correlation Coefficient of the parameter value to the model score (goodness of fit) to measure the sensitivity of that parameter. We find that the NO model is significantly less sensitive to parameter variation than the TD model (See Figure 4.14 and Figure 4.15 for sensitivity analyses of the ZRT1 and ADH1/3 models respectively).

4.5.7 Simulating coupled and decoupled activation and repression

To investigate the effect on gene expression noise of activation and repression by the same TF (coupled) versus activation and repression by two different TFs (decoupled), we extended the ZRT2 kinetic model to incorporate fluctuations in the concentration of TF. We use a kinetic scheme in which on-switching rates for activator and repressor can be changed independently (Konact and Konrep, see Eq.16). These rates are determined by the distributions of activator TF and repressor TF, respectively. We therefore assume that the on-switching rates have Gamma distributions with a constant shape parameter (burst size) and varying scale parameter (burst frequency) as the activator and repressor are induced. The means of the on-switching rates were chosen to be in the range of our model fits (10^{-3} to 10^{3}), which are in accordance with previously determined promoter switching rates. Next, we calculate the shape parameter of the distribution of Kon using the ratio between the mean of the measured protein distribution and the chosen mean of the on-switching rates (ratio of ~10^{3}), which we apply to the measured noise strength (~10^{3}). This gives a shape parameter value of around 10^{-1}. Because the product of shape and scale is equal to the mean, we can compute the values of the scale parameter (10^{2} to 10^{6}). We note that the qualitative result of predicted noise reduction (Figure 4.6B) is robust to 10-fold changes (up and down) of both shape and scale parameter of the distributions of Konact and Konrep. Finally, to simulate coupled and decoupled activation and repression we sampled the on-switching rates of the activator and repressor from a bivariate gamma distribution with a normalized covariance of zero (decoupled) or one (coupled). Each sample represents a single cell with some amount of activator and repressor, and therefore some Konact and Konrep. We then computed the mean expression for each ‘cell’ using the analytical solution of the ZRT2 model and calculated the predicted noise that results from fluctuations of activator and repressor as the squared coefficient of variation (sensitivity to TF fluctuations, η^2_{TF}).
Promoter sequence determines the relationship between expression level and noise

\[ \hat{K} = \begin{pmatrix} -(K_{off}^{\text{rev}} + K_{on}^{\text{rev}}) & K_{on}^{\text{rev}} & 0 & K_{off}^{\text{rev}} \\ K_{off}^{\text{rev}} & -K_{on}^{\text{rev}} - K_{on}^{\text{rev}} & K_{off}^{\text{rev}} & 0 \\ 0 & K_{on}^{\text{rev}} & -(K_{on}^{\text{rev}} + K_{off}^{\text{rev}}) & K_{off}^{\text{rev}} \\ K_{on}^{\text{rev}} & 0 & K_{on}^{\text{rev}} & -(K_{off}^{\text{rev}} + K_{off}^{\text{rev}}) \end{pmatrix} \] (16)

4.6 Financial Disclosure
This work was supported by grants from the European Research Council (ERC) and the US National Institute of Health (NIH) to E.S. E.S. is the incumbent of the Soretta and Henry Shapiro Career Development Chair. P.S., J.K. and D.D. were supported by FP7 FET Open project ‘DynaNets’ (EU Grant Agreement no. 233847), and ‘ViroLab’ (EU project no: IST-027446). D.D. was also supported by an EMBO Short-Term Fellowship and a travel grant from the BioRange program of the Netherlands Bioinformatics Centre (NBIC) and Netherlands Genomics Initiative (NGI). L.B.C was supported by a Weizmann Institute postdoctoral fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

4.7 Acknowledgments
We thank Leeat Keren and Maya Lotan-Pompan for yeast strains and members of the Segal lab, Marten Postma, Michael Dinitz and Asaf Farhi for useful discussions and suggestions; Iftach Nachman, Josh Rest and Yair Field for commentary on the manuscript; and the flow cytometry core facility for help with expression measurements.

4.8 Author Contributions
The authors have made the following declarations about their contributions:
Conceived and designed the experiments: LBC DD ES.
Performed the experiments: LBC DD.
Analyzed the data: LBC DD.
Conceived the models: DD LBC.
Wrote the paper: LBC DD ES PS JK.
4.9 References

4.10 Supplemental information

4.10.1 Flow cytometry with a dual-reporter dual-gene system can accurately measure promoter-specific changes in intrinsic noise

It is important to note that quantification of changes in burst size and burst frequency through changes in noise assumes an intrinsic measurement of expression in which variability in expression between cells is a result of stochasticity at the promoter and not at the general transcriptional/translational machinery. Any extrinsic influences will distort the noise-burst frequency and noise strength-burst size relationships. In order to correct for ‘extrinsic noise’ (expression capacity) we used a dual-gene system with the promoter of interest driving YFP and a constitutively expressed mCherry, and use the ratio between YFP and mCherry as the measured expression level. Because mCherry expression does not change with zinc concentration (Figure 4.8), this normalization can be used to remove significant amounts of extrinsic variability consistently throughout induction. To further eliminate extrinsic variation we perform stringent gating on forward scatter, side scatter, and mCherry expression, which results in the removal of ~70-90% of the cell population. Finally, we note that we quantify changes in noise, not absolute noise, and therefore we only need to accurately measure changes in intrinsic noise, and not intrinsic noise itself. For these reasons we therefore expect that the reported YFP/mCherry measurements accurately reflect a combination of pathway specific and promoter specific noise. Furthermore, we believe that pathway specific noise represents a small fraction of the reported noise, and that this does not affect our results. The strongest evidence is that the noise for all promoters does not change uniformly with zinc or with measured ZAP1 promoter driven YFP. For example, noise of ADH1 and ADH3 is high at low zinc concentrations, while noise of ZRT2 and ZRT1 is low at low zinc concentrations. Furthermore, our ability to measure differences in noise between different promoter and ATG context mutants (which, presumably, do not affect pathway-specific noise), strongly suggests that intrinsic noise is the dominant source of noise in our measurements.

4.10.2 Flow cytometry with an Elowitz style dual-reporter measures intrinsic noise similar to our dual-gene dual-reporter system

To validate our claim that the dual-gene dual-reporter system measures intrinsic noise, we constructed an Elowitz style dual-reporter for the ZRT2 promoter. Figure 4.4d (inset) shows that intrinsic noise as a function of induction gives the characteristic curve that we observed using our dual-gene system.

4.10.3 Sensitivity analysis

To investigate the robustness and sensitivity of the parameters of our model we perform a rigorous sensitivity analysis procedure described by 60. First, we sample 10,000 instances of the model, each with a unique parameter setting. This sampling is done using a Latin Hypercube, to ensure uniform sampling in the multidimensional parameter space of our model. Next we do a rank transformation
on the sampled instances using the distance of the model instances to the measured data (for both mean expression and noise) and quantify the correlation that each parameter has with the goodness of fit measure. Figure 4.14 shows the scatter plots and correlation coefficient for each parameter of the ZRT1, ADH1 NO/TD and ZRT2 models.

We note that the NO model is significantly less sensitive than the TD model. The most sensitive parameter in the model is the scaling parameter. This parameter converts measured fluorescence to number of proteins and therefore we would expect this parameter to influence model outcome hugely.

4.10.4 Robustness analysis
Barkai et al. describe a method of model validation that is based on the robustness of a model to parameter perturbations. The idea is that a free parameter model will often have a parameter setting that fits the data, regardless of it being the right model. However, biological systems are noisy and parameter values are not fixed but fluctuate around some mean. These fluctuations do not prevent the biology from working. Therefore a model of a biological system should be robust to parameter perturbations. A model that is not robust is likely not the right model.

We proposed two different models for ADH1 repression. We found that although the NO model gave a better fit, both models have good fits to the measured data. To investigate which of the models is more robust we performed a perturbation analysis. For each of 100 best fits, and for each parameter, we sampled 10,000 perturbations in the range of -2 to 2 fold change. We did this for both alternative models. Figure 4.3E shows the cumulative distribution of the goodness of fit (distance of model to data) for both models. We find that the NO model is significantly more robust than the TD model and therefore more likely to be the correct model for ADH1 repression.

4.10.5 Notation and choice of parameter values
While we do not know the precise values of many of the parameters in our model, biologically realistic bounds can be estimated for most parameters. We chose to set the upper and lower bounds of each free parameter as loose as reasonable, and we obtained good fits to data for a wide range of parameters. This robustness to parameter variation shows that the promoter-switching model accurately replicates the experimental data despite both uncertainties in the values of individual biochemical values and in experimental error.

Here follows a description of all parameters used in the model and their respective value constraints:

(All rate parameters are in minute⁻¹)

**Protein degradation (δ).** YFP is highly stable. We therefore assume that protein degradation comes only from dilution. While the population doubling time does
increase with as the zinc concentration drops below 150μM, per-cell rates of protein production also decrease with decreasing growth rate. For reasons of simplicity we chose to fix δ. This approximation does not affect our conclusions. We set the degradation rate (δ) to \( \ln(2) / 90 = 0.0077 \). Where 90 is the average measured doubling time (in minutes) in our experiments.

**Burst size (b).** Burst size is the average number of YFP molecules produced from each mRNA molecule. This parameter combines both translation and mRNA degradation assuming geometrically distributed protein abundance. Lu et al. found that ~80% of yeast proteins have between 100-10,000 proteins per mRNA. We therefore set the bounds for expression of the yeast codon-optimized YFP to be 100 – 10,000 proteins/mRNA.

**Transcription rate (r1, r2, r3, r4).** The transcription rate is the rate of production of stable mRNAs while the promoter is in each respective state. Experimentally measured transcription rates combine both on and off promoter states. Therefore, the total rate of transcription for a gene is the sum, for all states, of the fraction of time spent in that state times the transcription rate in that state. The upper bound for expression rate is between 4 transcripts/minute and 10 transcripts/minute. We note that these rates represent the combination of on and off promoter states. We set the transcription rate bounds to be 0-1 for off states, and 2-8 for on states.

**Promoter switching rates (Kon, Koff).** To estimate the minimum rate with which a promoter can switch between off and on we used expression data from the repressed GAL1 promoter. We assume that this promoter is off most of the time, but occasionally switches on. Cai et al. measured a burst frequency of 0.2 events per cell-cycle for the GAL1 promoter on a plasmid with a copy number of 10-40 \( ^{26} \). We therefore set the minimum switching rate (\( K_{\text{min}} \) see Eq.11,14,15) to be \( 10^{-4} \) (0.2 transcripts/cell-cycle / 90 minutes/cell-cycle / 20 plasmid copies/cell). High Kon values for a TF means that the TF is essentially always bound, while high Koff values mean that the TF is never bound. We know that the off-switching rate for a promoter can be relatively rapid, as transcription of POL1 occurs as a series of uncorrelated events, each resulting in the production of a single mRNA molecule. This implies that the promoter off switching rate can be high enough so that only a single transcription event occurs each time the promoter switches on. We therefore set the maximal switching rate (\( K_{\text{max}} \)) to be 10.

**Promoter switching as a function of [zinc].** We describe promoter switching with a hill equation as a function of [zinc]. This function has four parameters: \( K_{\text{min}} \) and \( K_{\text{max}} \) (described above), and [Zn]mid and H, which is the [zinc] of half maximal K value (threshold) and Hill-coefficient (sensitivity) respectively. We allow [Zn]mid to take on any value that is within the input [zinc] range – between 1 and 3000 uM zinc. H can be any positive value, where zero is a linear response and any non-zero positive value a sigmoidal response.
Scaling of YFP to proteins/cells (S). To convert measured fluorescence into molecules of YFP we determined how the measured YFP/mCherry ratio for each promoter at 3mM zinc compares to measured protein molecules / cell (Wang:2012eu). We find a normalized fluorescence to protein scaling of: \( \text{YFP/mCherry} \times 0.0146 = \text{Protein (molecules/cell)} \). We allow the scaling to be a free parameter, within a range of 10 fold up and down from the determined scaling value.

### 4.10.6 Parameter values of best fits

**ZRT1 (ATG):**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kmin</td>
<td>10^-5.5014</td>
</tr>
<tr>
<td>Kmax</td>
<td>10^-0.9406</td>
</tr>
<tr>
<td>ZNmid</td>
<td>10^3.8179</td>
</tr>
<tr>
<td>H</td>
<td>10^0.5778</td>
</tr>
<tr>
<td>Koff</td>
<td>10^-0.7332</td>
</tr>
<tr>
<td>r1</td>
<td>10^0.0867</td>
</tr>
<tr>
<td>r2</td>
<td>10^-0.3211</td>
</tr>
<tr>
<td>b1 (each ATG variant has its own b)</td>
<td>550</td>
</tr>
<tr>
<td>b2</td>
<td>843</td>
</tr>
<tr>
<td>b3</td>
<td>1100</td>
</tr>
<tr>
<td>b4</td>
<td>904</td>
</tr>
<tr>
<td>b5</td>
<td>1430</td>
</tr>
<tr>
<td>b6</td>
<td>1312</td>
</tr>
</tbody>
</table>

**ADH1 NO:**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kmin</td>
<td>10^-3.3292</td>
</tr>
<tr>
<td>Kmax</td>
<td>10^-0.8158</td>
</tr>
<tr>
<td>ZNmid</td>
<td>10^1.7836</td>
</tr>
<tr>
<td>H</td>
<td>10^2.2554</td>
</tr>
<tr>
<td>Koff</td>
<td>10^-0.6999</td>
</tr>
<tr>
<td>r1</td>
<td>10^0.0331</td>
</tr>
<tr>
<td>r2</td>
<td>10^2.0000</td>
</tr>
<tr>
<td>b</td>
<td>10^2.7879</td>
</tr>
</tbody>
</table>

**ADH1 TD:**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kmin</td>
<td>10^-2.8635</td>
</tr>
</tbody>
</table>
Promoter sequence determines the relationship between expression level and noise

Kmax $10^{-0.5724}$
ZNmid $10^{1.9606}$
H $10^{3.2449}$
Kon $10^{-1.8496}$
r1 $10^0$
r2 $10^2$
b $10^2.4002$

ZRT2:

Kmin $10^{-6.0000}$
Kmax $10^{5.3546}$
ZNmid $10^{2.0913}$
H $10^{0.1965}$
5 Kon_rep (ratio Kon_act to Kon_rep) $10^{1.2548}$
Koff_act $10^{-0.4506}$
5 Koff_rep (ratio Koff_act to Koff_rep) $10^{-0.1878}$
r1 $10^{0.3281}$
r2 $10^5.0000$
r3 $10^0$
r4 $10^{-0.2273}$
b $10^{-3.1694}$

4.10.7 Supplementary Table 1. Cloned promoter sequences lengths

<table>
<thead>
<tr>
<th>Promoter</th>
<th>Cloned sequence length (upstream of ATG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADH1</td>
<td>1000</td>
</tr>
<tr>
<td>ADH3</td>
<td>1000</td>
</tr>
<tr>
<td>ADH4</td>
<td>1000</td>
</tr>
<tr>
<td>DPP1</td>
<td>1000</td>
</tr>
<tr>
<td>ENO1</td>
<td>1000</td>
</tr>
<tr>
<td>ENO2</td>
<td>448</td>
</tr>
<tr>
<td>HXT1</td>
<td>1000</td>
</tr>
<tr>
<td>IZH1</td>
<td>654</td>
</tr>
<tr>
<td>MCD4</td>
<td>1000</td>
</tr>
<tr>
<td>PGM2</td>
<td>694</td>
</tr>
<tr>
<td>TKL1</td>
<td>898</td>
</tr>
<tr>
<td>TKL2</td>
<td>1000</td>
</tr>
<tr>
<td>ZAP1</td>
<td>1000</td>
</tr>
<tr>
<td>ZRC1</td>
<td>1000</td>
</tr>
</tbody>
</table>
4.11 Supplemental figures

Figure 4.8: Zap1 regulated targets change expression in response to changes in zinc concentration. Shown is measured mean expression for three different Zap1 targets (ZAP1, ZRT1, ADH4 and TKL2) showing quantitatively different basal (high zinc, low Zap1) expression levels as well as different induction curves. Also shown is expression of YFP from a truncated ENO2 promoter that lacks the two Zap1 binding sites, and is thus insensitive to changes in zinc concentration. For all YFP data, error bars show the standard deviation from at least three biological replicates. Shown in red is measured expression of TEFpr-mCherry from all of the data from all strains shown.
Promoter sequence determines the relationship between expression level and noise

Figure 4.9: Predicted Zap1 occupancy of each promoter is predictive of each activated promoter’s change in expression in response to increasing Zap1. The range of expression (lowest to highest measured values) graphed against the predicted number Zap1 molecules bound to a 500bp window from -600 to -100 relative to the transcription start site for each promoter. A thermodynamic model of promoter TF occupancy [37] shows that the measured Zap1 dose-response curve for each promoter highly correlates with the predicted occupancy of Zap1 at each promoter, suggesting that the number and affinity of Zap1 binding sites plays a large role in determining each promoter-specific dose-response curve. To identify sequence features in each promoter, we used a thermodynamic model in which Zap1 binding along the promoter sequence is determined by the concentration of the TF, its measured sequence specificities, and competition with nucleosomes [37]
Figure 4.10: The correlation of noise and noise strength with mean expression for all measured Zap1 targets throughout the induction. The correlation between expression and noise (A) and expression and noise strength (B) is shown for all measured expression data for each Zap1 target promoter. For each target expression is normalized between zero and one. The mean Spearman correlation coefficient for all promoters is shown.
Promoter sequence determines the relationship between expression level and noise.

Figure 4.11: Gene-specific slopes are significantly different from each other. Distributions show the mean-noise slopes obtained by bootstrapping all biological replicate experimental measurements of noise ($\eta^2$) and mean expression ($\mu$) and fitting the line $\eta^2=c\mu^k$ 1000 times for each promoter. The distributions show that the slope ($k$) values are significantly different and that slope value estimation is robust for most promoters.
Figure 4.12: Only part of the change in expression due to ATG context variants can be explained by changes in mRNA level. Shown are the fold differences in protein and mRNA between start codon context variants of the ZRT1 promoter. In order to determine if changes of the four nucleotides upstream of the ATG lead to differences in mRNA levels we performed RT-qPCR on three ATG context variants plus the WT ZRT1 promoter. Because mCherry is expected to be constant between the different strains, YFP/mCherry ratios are used as measurements for both fluorescence and mRNA. In addition, because both measurements are in arbitrary units, we cannot compare numeric values directly. However, both measurements are linear, and therefore ratios relative to a common control (the CTTT strain) can be compared. We find a 2.1 fold change in protein levels and a 1.3 fold change in mRNA levels. In addition we observe no correlation between mRNA and protein. These results are consistent with previous data [Yun, Laz, Clements & Sherman MM 1996] showing that the start codon context can change protein expression without affecting mRNA levels.
Promoter sequence determines the relationship between expression level and noise

Figure 4.13: A model in which zinc concentration changes promoter Kon, and ATG context variants change $b$, best explains the experimental data. There are many possible regulatory mechanisms by which zinc concentration and ATG context may change expression of ZRT1. In order to determine which regulatory mechanism best fits our data we fit the model represented by kinetic scheme in Fig 2a to our data 12 times. In each time, we mandated that a different pair of regulatory mechanisms be used to fit the induction of the ATG context variants. We find that a model in which the induction increases Kon, and ATG context variants change $b$ (first column) obtains the best fit to data. We note that similar regulatory mechanisms in which ATG context variants change burst size (columns 2 and 3) obtain fits that are almost as good. In contrast, a model in which zinc changes promoter off switching rate ($K_{off}$) never obtains as good a fit to the data, and is only capable of obtaining a reasonable fit to the data when ATG context variation changes the rate of transcription.
Figure 4.14: Sensitivity analysis of the ZRT1 model. To determine how sensitive the ZRT1 model is, we used LHS-PRCC (see materials and methods). We find that the fit of the model to data is highly only to sensitive only to $S$, a scaling factor we use to convert measured YFP/mCherry per cell signal to number of YFP protein molecules per cell. (A) Density scatter plots from LHS sampling of the parameter space show how the fit to data (y-axis) changes as a function of each parameter (x-axis). Correlations of the data in (A) are shown together for comparison in (B).
Figure 4.15: Sensitivity analysis of both proposed ADH1 models shows that the nucleosome occlusion (NO) model is less sensitive to parameter variation than the TF dislodgment (TD) model. To determine how sensitive each of the ADH1 models are we performed LHS-PRCC sensitivity analysis. (A) Density scatter plots from LHS sampling of the parameter space show how the fit to data (y-axis) changes as a function of each parameter (x-axis). Correlations of the data in (A) are shown together for comparison in (B). The NO model is far less sensitive to variation in biological parameters.
Figure 4.16: Gating of the ZRT2pr dual reporter causes a 10-fold reduction in extrinsic noise. For each cell, mCherry is plotted against YFP for one induction point from the ZRT2pr dual reporter. Extrinsic noise was reduced by gating using either a very wide (black) or very narrow (red) gate on forward and side scatter.
Chapter 5 Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

Maya Dadiani1, David van Dijk2, Barak Segal1, Yair Field1, Gil Ben-Artzi1, Tali Raveh-Sadka1, Michal Levo1, Irene Kaplow1, Adina Weinberger4 and Eran Segal1

1Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel.

2Computational Science, University of Amsterdam, 1980 XG Amsterdam, The Netherlands.

Genome Research (2013). doi:10.1101/gr.149096.112
5.1 Summary

Individual cells from a genetically identical population exhibit substantial variation in gene expression. A significant part of this variation is due to noise in the process of transcription that is intrinsic to each gene and is determined by factors such as the rate with which the promoter transitions between transcriptionally active and inactive states, and the number of transcripts produced during the active state. However, we have a limited understanding of how the DNA sequence affects such promoter dynamics. Here, we used single cell time-lapse microscopy to compare the effect on transcriptional dynamics of two distinct types of sequence changes in the promoter that can each increase the mean expression of a cell population by similar amounts but through different mechanisms. We show that increasing expression by strengthening a transcription factor binding site results in slower promoter dynamics and higher noise as compared to increasing expression by adding nucleosome-disfavoring sequences. Our results suggest that when achieving the same mean expression, the strategy of using stronger binding sites results in a larger number of transcripts produced from the active state, whereas the strategy of adding nucleosome disfavoring sequences results in a higher frequency of promoter transitions between active and inactive states. In the latter strategy, this increased sampling of the active state likely reduces the expression variability of the cell population. Our study thus demonstrates the effect of cis-regulatory elements on expression variability and points to concrete types of sequence changes that may allow partial decoupling of expression level and noise.

5.2 Introduction

Fluctuations in the process of transcription can generate considerable cell-to-cell variability in the expression level of genes across isogenic cell populations\textsuperscript{1-4}. This variability can result in phenotypic diversity and thus lead to potentially important effects on many cellular and developmental processes. Consequently, unraveling the sources that underlie expression variability has been a topic of considerable interest. Several studies demonstrated that genes are transcribed in bursts\textsuperscript{1,5-6}, such that the expression variability due to transcription is determined by the frequency with which the bursts occur (burst frequency) and the number of transcripts produced per burst\textsuperscript{9} (burst size). These parameters are determined in part by fluctuations in factors that are extrinsic to the promoter, such as the concentrations of the regulating transcription factors and RNA polymerases\textsuperscript{3,10}. The remaining variability is gene-specific intrinsic noise that was shown to vary greatly across promoters\textsuperscript{1,12}, and which is at least partly encoded by cis-acting regulatory elements embedded within the DNA sequence of each promoter\textsuperscript{12,13}.

Several types of regulatory elements were shown to affect transcriptional noise. Mutating a TATA box in two yeast promoters reduces expression variability, in a manner that is consistent with an effect of TATA boxes on the size of transcriptional bursts\textsuperscript{12-14}. The number and affinity of transcription factor binding sites was also shown to affect expression variability\textsuperscript{6,15,16}, with one study demonstrating that using two sites for the mammalian transcription factor NF-Y instead of one, or using a
higher affinity NF-Y site, increased expression variability in an artificial promoter construct. Similar to the effect of TATA boxes, the increase in expression variability observed with two NF-Y sites or with a higher affinity site was mediated primarily by an increase in the average burst size, whereas the burst frequency was largely unaffected. Chromatin regulation has also been linked to expression variability, with two studies showing that adding nucleosome-disfavoring sequences to a yeast promoter resulted in lower nucleosome occupancy and reduced expression variability. However, in contrast to the effect of TATA boxes and TF sites, chromatin was suggested to mainly affect the frequency with which promoters transition between transcriptionally active and inactive states. In yeast, statistically significant genome-wide associations were found between higher expression variability and the presence of TATA boxes, the number of TF binding sites, and the encoding of high nucleosome occupancy.

A recent study showed that the quantitative increase in the mean expression over a cell population that results from adding nucleosome-disfavoring sequences can be comparable to the increase in expression achieved when increasing the affinity of a transcription factor (TF) binding site. Several lines of evidence suggested that the effect of the nucleosome-disfavoring sequences was likely mediated by the lower nucleosome occupancy and thus increased accessibility that nucleosome-disfavoring sequences confer over the nearby promoter elements, such as TF sites. Thus, combined with the above studies, an intriguing hypothesis is that the two distinct types of sequence changes in either TF sites or in nucleosome-disfavoring sequences provide a genetic mechanism that may allow partial decoupling of mean expression level and transcriptional noise. Specifically, since mean expression is the product of burst frequency and burst size, and noise (under some assumptions) is the inverse of burst frequency, then increasing burst frequency is expected to result in lower noise compared to a similar increase in mean expression that is due to an increase in burst size.

Here, we set out to test the above hypothesis, as well as, to directly compare the effects that changes to transcription factor binding sites have on expression variability, to the effects induced by changes to sequences that affect nucleosome occupancy. To associate the observed changes in expression variability with one of these types of sequence changes, we always performed the two types of sequence changes being compared within the same promoter context. By monitoring fluorescent intensity in single cells using time-lapse microscopy, we compared the dynamics of promoters in which expression is increased by adding or lengthening nucleosome-disfavoring sequences to the dynamics induced by changes to the binding affinity of sites for two different transcription factors. Our results demonstrate that these two distinct DNA-encoded strategies for increasing expression indeed have opposing effects on promoter dynamics, allowing the same mean expression level to be achieved but with lower noise when the strategy of nucleosome-disfavoring sequences is employed. These results hold for the two
different transcription factors that we tested and are further supported by independent flow cytometry measurements that we performed.

Thus, our results show how mean expression and expression variability may be partially decoupled, providing a potentially useful tool for synthetic biology, and raising intriguing hypotheses regarding the extent to which evolution of native promoters has utilized these genetic mechanisms to partly decouple mean expression and noise levels.

5.3 Results

5.3.1 Experimental design and measurement of promoter dynamics
We sought to compare the effect on promoter dynamics of two distinct strategies for increasing expression, strengthening transcription factor binding sites and adding or lengthening nucleosome disfavoring sequences, since we hypothesized that each type of sequence change would affect different aspects of promoter dynamics. Specifically, in the classical scenario in which the binding of a transcription factor induces expression via its stabilizing effect on polymerase, we expect that strengthening the affinity of a TF site may increase the time that the factor remains bound at its site, resulting in more transcripts produced during the active state, and thus increasing expression by increasing the burst size (Figure 5.1A, bottom). In contrast, since addition of poly(dA:dT) tracts, which disfavor nucleosome formation\textsuperscript{24}, was shown to result in lower nucleosome occupancy and more accessible DNA in the vicinity of the tract\textsuperscript{20,25}, we hypothesized that this would result in faster on/off binding dynamics of a TF to a nearby site, and thus increase expression by increasing the burst frequency (Figure 5.1A, middle). Thus, we predicted that the strategy of increasing expression by lowering the encoded nucleosome occupancy of a TF binding site would result in faster promoter dynamics and thus lower expression variability across a cell population, as compared to the strategy of increasing expression by strengthening the TF binding site.

To test this hypothesis, we employed a controlled setting in which only the elements hypothesized to have differential effects on promoter dynamics are altered. To this end, we used yeast strains in which different promoters are separately integrated into the same genomic location and upstream of the same yellow fluorescent protein (YFP) reporter\textsuperscript{26}. To control for experimental variability and as a marker for segmenting and tracking the cells in all promoter variants, an mCherry fluorescent reporter is also integrated into the same genomic location in all strains, downstream of a constant promoter (TEF2, a translation elongation factor). We used a set of synthetic variants that were generated in a recent study, in which poly(dA:dT) tracts with different lengths and binding sites with different affinities for two distinct transcription factors were integrated into the same
Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

promoter context20 (Figure 5.1B). The promoter context is derived from the native yeast HIS3 promoter, and contains two poly(dA:dT) tracts flanking a single site for the transcriptional activator Gcn4. Measurements of mean expression levels of these promoter variants showed that variants with longer poly(dA:dT) tracts exhibit higher YFP expression. Consistent with an effect mediated by changes in nucleosome organization, variants with longer poly(dA:dT) had lower nucleosome occupancy over the nearby transcription factor binding site20. Important for our purposes, this recent study showed that these two distinct strategies of increasing expression, once by modulating binding site affinity and once by modulating poly(dA:dT) length, can achieve comparable quantitative increases in the mean expression level of the cell population.

To test our hypothesis that these distinct genetic strategies for increasing expression have opposing effects on promoter dynamics, we utilized two independent approaches that can each provide estimates of promoter dynamics (Figure 5.1C,D). The first monitors YFP fluorescence in single cells over time and thus provides more direct measurements of promoter dynamics. To this end, we grew cells in a microfluidic plate that maintains continuous supply of media, and used an automated microscopy system to image fluorescence continuously over 15 hours and at a resolution of 3 minutes (Figure 5.2A, S5.6). We developed analysis software that uses the mCherry fluorescence signal to segment the cells and track the segmented cells over time, resulting in time-traces of mCherry and YFP signal intensities (Figure 5.2A,B, Supplemental methods). We followed each cell over multiple cell doublings, which was identified by the marked decrease in the fluorescence signal that is caused by diffusion of the fluorescent reporter into the daughter cell (Figure 5.2B, Supplemental methods). For each cell, we then computed the rates of its mCherry and YFP production over time, by differentiating the dilution-corrected signal tracks (Figure 5.2B, Supplemental methods). To compare promoter dynamics across different strains, we normalized the YFP signal by the mCherry signal (Figure 5.2B), as this accounts for many extrinsic factors that are correlated with the measured protein production rates, such as variations in cell size, time within cell cycle, and abundance of general components of the transcriptional machinery (Supplemental methods). As a second independent method from which indirect estimates of promoter dynamics can be extracted23, we used flow cytometry to measure the distribution of single cell fluorescent intensity at a single time point (Figure 5.1D). Notably, in both measurement methods, the mean expression of the cell population increases with both the length of the poly(dA:dT) tract and the affinity of the transcription factor binding sites, consistent with previous measurements20 (Figure S5.1).

5.3.2 Lengthening poly(dA:dT) tracts and strengthening transcription factor binding sites have opposing effects on the rate of promoter dynamics

As a direct and unbiased approach to compare dynamics of promoter activity across our various promoter variants, we first computed the autocorrelation function of the YFP production rates as measured by time-lapse microscopy of every variant27.
Higher correlations indicate production rates that are maintained across longer time periods and thus suggest slower changes in dynamics of promoter activity. To perform the analysis, we extracted the time traces of YFP production of thousands of individual cells from each promoter variant, computed the autocorrelation of each time trace at different time lags in six minute resolution, and for every time lag, averaged the autocorrelation values of the cells of each variant.

Comparing the autocorrelations of five different variants with poly(dA:dT) tracts of length 0, 5, 12, 17, and 22bp, we found significant gradual reductions in the autocorrelation of the variants as the length of the tract increases (Figure 5.2C). In contrast, we found an opposite trend of an increase in the autocorrelation values of three different promoter variants in which the affinity of the binding site for Gcn4 is strengthened (Figure 5.2D). In both cases, the differences between variants were significant across several lags, and the most significant lags correspond to lags that are less than 20 minutes and thus much shorter than the average cell cycle time (~90 minutes).

To verify that these effects are not specific to the tested variants and to Gcn4, we repeated the analysis in three additional sets of variants consisting of a set of five variants in which the above poly(dA:dT) tracts were added to a different promoter background; a set of three variants with increasing binding site affinities for the transcriptional activator Gal4, which is not known to regulate the native promoter from which our variants were derived; and another set of the same three Gal4 sites added to a modified promoter background. Consistent with our above results, we found that variants with longer poly(dA:dT) tracts within the modified promoter background had lower autocorrelation values than shorter tracts (Figure 5.2E), and variants with higher affinity Gal4 sites in both promoter backgrounds had higher autocorrelations than lower affinity sites (Figure 5.2F, S5.2). Moreover, comparing the two sets of Gal4 site variants between the two promoter backgrounds that differ in the presence of a poly(dA:dT) tract, we found that the variants within the poly(dA:dT)-containing promoter background had lower autocorrelation values, providing further support to the reduced autocorrelation that occurs upon addition of poly(dA:dT) tracts.

Our results suggest that these two distinct strategies for increasing expression, namely lengthening poly(dA:dT) tracts or strengthening binding site affinity, do so with opposing effects on the time-dependent dynamics of promoter activity.

5.3.3 Longer poly(dA:dT) tracts induce faster promoter transition rates
To obtain a more direct visual view of the above results at the single cell level, we used our data to estimate the frequency with which each of the above promoter variants transitions between high-expressing and low-expressing states. To this end, we defined the high and low expressing states by whether the amount of YFP produced during the examined time window was above or below some arbitrary threshold, respectively. We again used the time traces of YFP production of thousands of individual cells from each promoter variant, and for a given YFP...
production threshold, classified every 3 minute time window of each cell’s YFP time trace into high and low expressing states (Figure 5.3A). From these classifications, we then computed for each promoter variant, its probability of transitioning between high and low expressing states, as well as the fraction of all of its cell cycle time traces that had \( k \) transitions, for all possible values of \( k \). Thus, this analysis examines how the high and low expressing time windows are distributed across the time traces of each variant. For example, a fast transitioning variant in which 50% of its time windows are high expressing would transition between high and low expressing states every 3 minutes, whereas an extremely slow variant would be continuously high expressing in half of all of the cell cycles examined and continuously low expressing in the other half. To ensure that the results are not sensitive to the choice of threshold, we performed the analysis across a wide range of threshold values.

Notably, we found that variants with longer poly(dA:dT) tracts had higher probabilities of transitioning between high and low expressing states and a higher fraction of cell cycles with more than five transitions (Figure 5.3B,D). Conversely, variants with higher affinity binding sites had lower transition probabilities and a higher fraction of cell cycles with at most one transition between high and low expressing states (Figure 5.3C,E, S5.3). These results were robust across all tested threshold choices (Fig. S5.4). Thus, although our data does not directly measure high and low expressing states, it clearly demonstrates that across the broad range of threshold choices for the high expressing states, increasing the length of poly(dA:dT) tracts results in faster transitions between the classified states, whereas increasing the affinity of a transcription factor binding site results in slower transitions. This suggests that these two distinct strategies for increasing the mean expression level of a cell population are mediated by opposing effects on promoter dynamics, consistent with the above autocorrelation analysis.

### 5.3.4 Agreement between time-lapse promoter dynamics and static expression measurements

As another experimentally independent way to estimate promoter dynamics, we also measured the single cell YFP and mCherry fluorescence intensity distributions of each promoter variant at a single time point of mid-log phase using flow cytometry. Under certain assumptions, previous studies have shown that by fitting the static distribution of such single cell fluorescence intensities to a gamma distribution, the two fitted parameters can be interpreted as the burst frequency with which genes are transcribed and the size of the bursts\(^{23,28}\). To remove much of the extrinsic noise that is due to variations such as cell size, we normalized the YFP intensity of each cell by the measured mCherry intensity to get a pathway specific measure of noise. Although this measure does not fully correspond to intrinsic noise in our experimental system, the distributions of normalized YFP fluorescent intensities were well fitted to a gamma distribution, thus supporting the use of this framework for estimating promoter dynamics (Fig. S5.5).
Examining the fitted parameters of the gamma distribution for each promoter variant, we found that increasing the affinity of the transcription factor binding site resulted in higher values for the inferred burst size parameter, and in little effect on the inferred burst frequency parameter (Figure 5.4A,B). We found this behavior in four different sets of variants that correspond to changes in the affinity of sites for both Gcn4 and Gal4, each in two different promoter backgrounds. In contrast, we found that increasing the length of a poly(dA:dT) tract resulted in higher values for the burst frequency parameter, and in little change in the burst size parameter (Figure 5.4C). Here too, we found similar behavior across two different promoter backgrounds in which the length of the poly(dA:dT) tract was increased. Thus, these results, obtained by an independent experimental system that measures the population-level distribution of fluorescent intensities, suggest that longer poly(dA:dT) tracts increase expression primarily by inducing a higher frequency of promoter transitions to the active state, whereas higher affinity binding sites do so primarily by increasing the size of the bursts from the active state. These results are in accord with the more direct measurements of promoter dynamics that we obtained using time-lapse microscopy.

5.3.5 Achieving similar mean expression levels with predictably different noise levels
Finally, analytical models predict that when two promoters achieve the same mean expression level but with different burst frequency and burst size, then the promoter with the higher burst frequency will exhibit less noise, defined as the standard deviation of the expression across single cells divided by the mean expression level of the population. We used our promoter variants to test this intriguing prediction, since according to our above estimates of promoter dynamics, our variants represent a case in which the same increase in mean expression can be achieved through different effects on burst frequency and burst size. To this end, we examined promoter triplets that each consist of three promoter variants in which promoters are modified by either adding/lengthening a poly(dA:dT) tract or strengthening a transcription factor binding site such that these modifications increase the mean expression level to similar levels. For each variant in every set, we then used our above flow cytometry measurements to compute its noise level. In all cases, the promoter in which a poly(dA:dT) tract was added had lower noise than the promoter in which the similar expression level was reached by increasing the binding site strength. This result is consistent with the prediction we set out to test, if indeed the same mean expression levels were achieved with the poly(dA:dT)-containing promoter exhibiting a higher burst frequency (Figure 5.5A). Notably, these sets included sites for both Gcn4 and Gal4, suggesting that the results may hold generally for more transcription factors.

We next examined two additional sets that also consist of three promoter variants each in which starting from a promoter with a poly(dA:dT) tract and a high affinity transcription factor binding site, deleting the poly(dA:dT) tract or lowering the site affinity reduced expression to similar mean levels. In both sets, the promoter in
which the poly(dA:dT) was deleted had higher noise, which is again consistent with the original prediction, if indeed the promoter without the tract achieved the same mean expression with a lower burst frequency (Figure 5.5B). Here too, the two sets represent sites for two different transcription factors (Gcn4 and Gal4).

Taken together, our results suggest that we can achieve similar quantitative effects on the mean expression of a cell population by either adding poly(dA:dT) tracts or strengthening the affinity of a transcription factor binding site, but the promoter with the poly(dA:dT) will have a higher burst frequency, lower burst size, and, as predicted by analytical models, lower noise, as compared to the promoter with the stronger site. Since the effects of these manipulations are predictable, these two distinct strategies may allow for partially decoupling mean expression level and noise.

5.4 Conclusions and discussion
Understanding the dynamical process by which promoters transition between active and inactive states is central to an understanding of transcriptional regulation. Here, we focused on the role of promoter DNA sequence in this process, and compared the effect on single cell expression of two different DNA-encoded strategies that are capable of increasing the mean expression level of a cell population by similar magnitudes. Compared to strengthening transcription factor binding sites, we found that increasing expression by adding nucleosome disfavoring sequences results in faster promoter dynamics and lower expression variability across an isogenic cell population, suggesting that mean expression level and transcriptional noise can be partially decoupled. Notably, these results held in several different contexts and for two distinct transcription factors. Previous studies showed that gene expression and noise can be decoupled by mutating the TATA box or by introducing negative autoregulation. Our work suggests that such decoupling can also be achieved through chromatin.

Although our experimental system does not directly measure the frequency with which promoters transition between active and inactive states or the number of transcripts produced from each transcriptional burst, both our time-lapse microscopy and flow cytometry measurements suggest that the expression increase of each of the two types of sequence changes that we tested is mediated by distinct mechanisms. In the case of adding poly(dA:dT) tracts, our results suggest that the increase in expression is primarily achieved by a higher frequency of transitions between the inactive and active states. We propose that the lower nucleosome occupancy induced by the poly(dA:dT) tracts increases the accessibility of the nearby promoter region, thereby reducing the time that it takes the cognate transcription factor to find the site and leading to a higher frequency of successful binding events and promoter activation. This suggestion is supported by a study showing that promoter firing rate is dictated by the time required for a transcription factor to find its gene within the nucleus. In contrast, when strengthening a transcription factor binding site, our results suggest that the
increase in expression is primarily achieved by an increase in the average number of transcripts produced during the active state. We propose that a stronger binding site reduces the dissociation rate between the binding site and its cognate factor, resulting in more stable binding of the transcriptional machinery and thus a higher probability of transcription re-initiation and longer and more sustained bursts, similar to the effects suggested for promoter TATA boxes. This suggestion is supported by recent findings that measured bursting kinetics in mammalian genes, and found increased mean burst sizes for artificial promoters designed with a higher affinity binding site for a transcriptional activator.

Finally, comparing the expression of five pairs of promoters in which the promoters in each pair have similar expression levels, we found that in every pair, the promoter with the longer poly(dA:dT) tract and lower affinity transcription factor binding site always had lower noise. Since noise is the inverse of burst frequency (allowing for some assumptions), and mean expression is the product of burst frequency and burst size, these results are also consistent with longer poly(dA:dT) tracts causing a higher burst frequency. We note that our experimental system provides a measure of pathway specific noise rather than intrinsic noise. However, stringent filtering gives us a measure that correlates well with intrinsic noise. Our results are consistent with a model in which the promoter elements affect promoter state switching (see Supplemental methods). Although we cannot rule out that other mechanisms may affect the observed transcription rates, our results suggest that the frequency and length of transcriptional activity change more than the magnitude of the activity itself.

Taken together, our results suggest genetic mechanisms by which partial decoupling of mean expression and noise can be achieved. It will be interesting to test the effect of other types of sequence changes on promoter dynamics and to identify cases where these signals may have been utilized in evolution for achieving biologically important single cell behaviors.

5.5 Methods

Yeast strains. The set of promoters analyzed consists of 22 promoters that were each genomically integrated into a shared master strain upstream of the yellow fluorescence protein (YFP) reporter and the HIS3 proximal promoter (100bp upstream of the ATG). The integrated region also contains a fixed control promoter for the TEF2 gene (a translation elongation factor) upstream of an mCherry fluorescent protein. The mCherry reporter serves as an internal control for normalization and cell segmentation. Promoter variants consist of different binding sites for the Gcn4 and Gal4 transcriptional activators and different lengths of poly(dA:dT) tracts upstream of the factor binding sites. Variants of the Gcn4 binding site are as described. Variants of the Gal4 binding sites were designed to span a range of weak medium and strong affinity as follows (strong site,
Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

CGGAAGACTCTCCTCCG; medium, AGGAAGACTCTCCTCCG taken from the GAL1-GAL10 UASg site 3, and weak, CGGATTAGAGCCGCGCG, taken from the GAL1-GAL10 UASg site 1.

Growth conditions. Yeast strains were grown at 30°C in synthetic complete medium supplemented with 2% glucose, in a 96-well plate for 48 hours to saturation. For fluorescence microscopy of single cells, cells were diluted to reach an optical density of 1 prior to the experiment in the desired medium and loaded into the microfluidic imaging plate (The ONIX™ Microfluidic perfusion system, CellASIC Corporation, CA). Cells were loaded into a 4 µm trap region that ensures monolayer growth of the cells for optimal maintenance of focal plane. A constant flow of media was kept, by applying a stable pressure of 1 psi that enabled constant flow at a rate of 2.5 µl/hour. For the flow cytometry measurements, cells were grown to saturation and then diluted to reach mid log phase (optical density of ~0.2) in the specified conditions. Conditions during the time-lapse experiment were SC supplemented with 2% glucose.

Time-lapse Microscopy. Time-lapse experiments were conducted using a commercial fully automated inverted fluorescence microscope (Nikon TiE) equipped with a motorized stage (Ti-S-ER, Nikon), hardware based focus maintenance system (PFS, Nikon), fast external shutters (SUTTER), 60X objective lens, and a cage incubator. Commercial filter sets were optimized for detection of YFP and mCherry (YFP ex500/20 em535/30 mCherry ex572/35 em632/60, Chroma). High-resolution images (effective pixel size of 0.216 µm) were acquired for 15 hours at a resolution of 3 minutes using a cooled charge-coupled device camera (EMCCD; DU-888E Andor). Approximately 3000 cell tracks were obtained for each imaging area.

Flow Cytometry. Flow cytometry experiments were conducted using the Becton-Dickinson LSRII machine and standard protocols. Four channels were acquired: forward-scatter, side-scatter, YFP and mCherry (excitation wavelength was 350nm for YFP and 740nm for mCherry). About 150,000 cells were collected from each well at a flow rate of 1µl/s. To reduce cell variability, cells were gated in the forward-scatter and side-scatter channels.

Data Analysis. Automated image analysis and cell tracking was performed using a modified version of CellProfiler and Matlab to analyze and process single-cell data. We devised an automated and robust framework to analyze and filter the cell data tracks, leaving only high quality cell data tracks. The framework included six modules that deal with image corrections, cell segmentation and tracking of the cells, segmentation and tracking post processing, cell lineage analysis, data masking and filtering, correction for cell cycle dilution that stems from diffusion from the mother to the daughter cell, and calculation of production rates.
5.6 Figures

A

- Transcription Factor
- TF site
- Nucleosome
- Rate of promoter activation
- Rate of promoter inactivation

Inactive Promoter state

Active Promoter state

Add poly (dA:dT) tract: Reduce search time

Poly dA:dT

Increase TF site affinity: Decrease TF dissociation rate

Higher BS affinity

B

- Genome
- Poly(dA:dT) Length modifications
- Binding site strength modifications

C

- Time-lapse Microscopy
- Flow Cytometry

D

- Single-cell expression analysis
- Time-course data

Higher k_{on}

Lower k_{off}
Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

Figure 5.1: Promoter variants and experimental setup. (A) Illustration of hypothesized promoter dynamics induced by two distinct DNA-encoded strategies for increasing expression and a matched stochastic simulation of promoter switching, transcription and translation (right panel) for each strategy. For a simplified model in which promoters transition at some rate between transcriptionally inactive and active states (top), we hypothesize that addition of nucleosome disfavoring elements such as poly(dA:dT) tracts would increase the accessibility of the transcription factor binding site, thereby reducing the time that a factor molecule spends in search of its site (middle). In contrast, we hypothesize that increasing the affinity of a factor binding site would reduce the factor’s dissociation rate. Note that both types of sequence changes result in a higher mean expression over the cell population, but with distinct hypothesized effects on promoter dynamics. Using the Gillespie algorithm, we simulate the kinetic scheme in three scenarios. In the simulation runs we record, as a function of time, the promoter state (black-red line), mRNA levels (blue line), protein levels (black line) and the protein production rate (green line), which is the derivative of the protein levels with respect to time. In addition, we record protein production bursts (red line), when production is positive. The “normal” promoter (a) represents the reference point for the parameter changes. The fast promoter (b) has an increased Kon to simulate an increase in promoter accessibility (added polyT). The “slow” promoter (c) has a decreased Koff (with respect to the normal promoter) to simulate an increase in TF binding site affinity. The three example runs shown illustrate that although both parameters can increase the overall expression level, Kon increases the frequency of production bursts, while Koff increases the length of the bursts. (B) Illustration of promoter variants employed in this study. All promoters are genomically integrated upstream of a yellow fluorescent protein (YFP) reporter and into a region that also contains an mCherry fluorescent protein driven by a constant TEF2 promoter. Promoter variants differ in the presence and length of two poly(dA:dT) tracts and in the affinity of the transcription factor binding site for either the Gcn4 or Gal4 transcriptional activators. (C) Representative YFP time-lapse microscopy images of four promoter variants, imaged in a microfluidic platform that supplies a continuous flow of medium. Each cell was followed over time, and its lineage, YFP, and mCherry signal intensity were extracted (Methods). (D) Single cell flow cytometry data collected for the corresponding promoter variants from (C), shown as a histogram of normalized YFP values (top) and as a scatter plot of YFP (x-axis) against mCherry fluorescence (bottom).
Figure 5.2: Opposing effects on promoter dynamics for lengthening poly(dA:dT) tracts and strengthening transcription factor binding sites. (A) Representative time lapse microscopy images of one imaging area at five different timepoints, displaying YFP fluorescence (top), mCherry fluorescence (middle), and automatically segmented cells. (B) Representative time-lapse traces of YFP and mCherry fluorescence of a single cell over time (top), along with YFP and mCherry production rates (middle), and normalized YFP production rates (bottom; normalization done by mCherry, see Methods). Blue circles denote cell cycles. (C) Longer poly(dA:dT) tracts result in faster promoter dynamics. Shown is the average autocorrelation of normalized YFP production rates.
Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

across thousands of different cell traces for each of five different promoter variants with poly(dA:dT) tracts of length 0, 5, 12, 17, or 22bp. Bars denote standard errors. (D) Higher affinity binding sites result in lower promoter dynamics. Same as (C), for three promoter variants that differ only in the affinity of the Gcn4 site. (E) Same as (C), but where the poly(dA:dT) tract variants were inserted into a different genetic background in which the right poly(dA:dT) tract is deleted (R0). (F) Same as (D), but for three Gal4 sites that differ in their affinity.

Figure 5:3: Lengthening poly(dA:dT) tracts and strengthening transcription factor binding sites have opposing effects on the rate of promoter transitions between active and inactive states. (A) Illustration of our analysis of promoter transition rates. For each cell cycle of every cell, we classify its trace of normalized YFP production rate (blue trace illustrated here for one cell cycle of one cell) into active (red) and inactive (green) states

- **A** Threshold for active state
- **B** Graph showing promoter dynamics
- **C** Expression levels
- **D** Active state vs Inactive state
- **E** Time vs Cell number

---

152 Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

across thousands of different cell traces for each of five different promoter variants with poly(dA:dT) tracts of length 0, 5, 12, 17, or 22bp. Bars denote standard errors. (D) Higher affinity binding sites result in lower promoter dynamics. Same as (C), for three promoter variants that differ only in the affinity of the Gcn4 site. (E) Same as (C), but where the poly(dA:dT) tract variants were inserted into a different genetic background in which the right poly(dA:dT) tract is deleted (R0). (F) Same as (D), but for three Gal4 sites that differ in their affinity.

Figure 5:3: Lengthening poly(dA:dT) tracts and strengthening transcription factor binding sites have opposing effects on the rate of promoter transitions between active and inactive states. (A) Illustration of our analysis of promoter transition rates. For each cell cycle of every cell, we classify its trace of normalized YFP production rate (blue trace illustrated here for one cell cycle of one cell) into active (red) and inactive (green) states

- **A** Threshold for active state
- **B** Graph showing promoter dynamics
- **C** Expression levels
- **D** Active state vs Inactive state
- **E** Time vs Cell number
according to whether they are above or below a predefined arbitrary threshold (red horizontal line), respectively. (B) Increasing the length of a poly(dAdT) tract results in a higher rate of transitions between active and inactive states. For promoter variants that differ in the length of a poly(dAdT) tract, shown is the fraction of all of its measured cell cycle traces in which the number of transitions between active and inactive states was at most 2 (slow transitions, left bar graph) or at least 5 (fast transitions, right bar graph). The comparison of these different promoter variants was done at a threshold in which the fraction of all inactive states in each variant was 70% (since absolute expression levels vary across variants, the absolute threshold value is different for each variant). See Figure S4 for similar analyses at a range of thresholds from 50-90%. (C) Increasing the affinity of a transcription factor binding site results in a lower rate of transitions between active and inactive states. Same as (B), but for variants that differ in the affinity of a Gcn4 binding site. (D) Visual illustration of cell cycle traces corresponding to the bar graphs from (B) in which the length of poly(dAdT) tracts was varied. For each promoter variant, shown are 200 rows that each correspond to a time trace of one cell cycle of one cell with colored entries representing active (red) or inactive states (green) at a threshold in which 70% of all states were inactive. Rows are sorted according to the number of transitions between active and inactive states, and the 200 rows were sampled from all cell cycle traces such that they accurately represent the same probability distribution of number of transitions across all cell cycle traces. (E) Same as (D), but for the bar graphs from (C) in which the affinity of Gcn4 sites was altered.
Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

Figure 5.4: Flow cytometry validation of opposing effects of lengthening poly(dA:dT) tracts and strengthening transcription factor binding sites. (A) Strengthening the affinity of a transcription factor binding site mainly affects burst size. For two different sets of promoter variants each with three different binding site affinities for Gcn4, shown are the values of the two parameters from the Gamma function when fitted to the normalized YFP intensities of each promoter variant measured over the cell population at a single time-point using a flow cytometer. Under certain assumptions, these two parameters correspond to the burst frequency (left graph) and burst size (right graph). Note the larger variation in burst size across these variants. (B) Same as (A), for variants in which the affinities of Gal4 binding sites were varied. (C) Lengthening a poly(dA:dT) tract mainly affects burst frequency. Same as (A), for variants in which the length of a poly(dA:dT) tract was varied. In contrast to (A) and (B), note the larger variation in
burst frequency across these variants. (R0, right poly(dA:dT) deleted; L0, left poly(dA:dT) deleted).

Figure 5.5: Adding poly(dA:dT) tracts and strengthening transcription factor binding sites have opposing effects on transcriptional noise. (A) Adding a poly(dA:dT) tract results in a similar increase in mean expression level but in lower noise compared to strengthening the affinity of a transcription factor binding site. For three sets of promoter variants (three different graphs), shown are the median YFP expression (x-
Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

axis) and expression noise (standard deviation of expression divided by the mean expression) of a starting promoter (leftmost promoter in each graph with lowest median YFP expression) and two promoters representing modifications to the starting promoter, in which either a poly(dA:dT) tract was added (bottom right promoter in each plot) or the binding site was strengthened (top right promoter in each plot). Note that both promoter modifications result in similar mean expression levels but the promoter in which the poly(dA:dT) tract was added always has lower noise. Bars denote standard error. (B) Same as (A), but where the starting promoter contained a poly(dA:dT) tract and a strong transcription factor binding site, and the modifications either deleted the poly(dA:dT) tract (top left promoter in each plot) or weakened the binding site (bottom left promoter in each plot). As in (A), both promoter modifications result in similar mean expression levels but the promoter with the poly(dA:dT) tract always has lower noise.
5.7 References

Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise


5.8 Supplementary information

**Figure S5.1. Promoter variants and expression levels.** Shown are the promoter variants used in this study and the median YFP expression of each variant, as measured by flow cytometry. Note the increase in expression with both the lengthening of poly(dA:dT) tracts (upper two promoter sets) and the strengthening of transcription factor binding sites (four bottom promoter sets).
Figure S5.2. Higher affinity binding sites result in lower promoter dynamics. Shown is the average autocorrelation of normalized YFP production rates across thousands of different cell traces for each of three different promoter variants that differ only in the affinity of the Gal4 site. Bars denote standard errors.
Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

Figure S5.3. Increasing the affinity of a transcription factor binding site results in a lower rate of transitions between high and low expressing states. Same as Figure 3B and Figure 3D, but for three promoter variants with differing affinities for the Gal4 binding site. The bar graph shows the fraction of cell cycle traces of each variant in which the number of transitions between high and low expressing states was at most 2 (slow transitions, left bar graph) or at least 5 (fast transitions, right bar graph). The comparison of these different promoter variants was done at a threshold in which the fraction of all low expressing states in each variant was 70% (since absolute expression levels vary across variants, the absolute threshold value is different for each variant). The left heatmap shows a visual illustration of the cell cycle traces from the bar graphs where for each promoter variant, shown are 200 rows that each correspond to a time trace of one cell cycle of one cell with colored entries representing high (red) or low expressing states (green) at a threshold in which 70% of all states were low expressing. Rows are sorted according to the number of transitions between high and low expressing states, and the 200 rows
were sampled from all cell cycle traces such that they accurately represent the same probability distribution of number of transitions across all cell cycle traces.

Figure S5.4. Lengthening poly(dA:dT) tracts and strengthening transcription factor binding sites have opposing effects on the rate of promoter transitions between high and low expressing states. For every promoter variant, we extract each cell cycle of every one of its cells, and classify for each cell cycle the trace of normalized YFP production rate into high and low expressing states according to whether they are above or below a predefined arbitrary threshold, respectively. From these numbers, we can then compute the probability of transitioning between high and low expressing states for every variant. For promoter variants that differ in the length of a poly(dA:dT) tract, the left graph shows the probability of transitioning between high and low expressing states (y-axis) across a broad range of thresholds (x-axis) on the fraction of all low expressing states in every promoter variant. The right graph shows the same computation but for three promoter variants that differ in the affinity of a Gcn4 binding site. Note that across the broad range of thresholds examined, lengthening a poly(dA:dT) tract (left graph) results in a higher rate of transitioning between states, whereas strengthening a transcription factor binding site (right graph) results in a lower transition rate.
Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

Figure S5.5. The static distribution of YFP fluorescence accurately fits a Gamma distribution. Promoter activity histograms measured by flow cytometry for two representative variants and their relative fits to several distributions: gamma, lognormal and Gaussian. For all promoter variants, the Gamma distribution best fit the data and thus justify the calculations of promoter dynamics.
5.8.1 Supplemental Experimental Procedures

5.8.1.1 Image acquisition
To ensure monolayer growth of cells and to keep a stable focal plane, cells were imaged in a yeast microfluidic plate (CellASIC). This plate enables constant flow of media throughout the experiment and thus provides "chemostat-like" growth. Images were acquired by a fully automated inverted fluorescence microscope. Raw images were segmented and objects were tracked to yield a single track of fluorescence intensity over time (Fig. S6).

5.8.1.2 Image analysis pipeline
In order to obtain high quality data on the cells imaged in the microscopy system, the images were analyzed in a modular framework (Fig. S7), with six modules that deal with image corrections, cell segmentation and tracking of the cells, segmentation and tracking post processing, analysis of the cell lineage, masking and filtering the data, and dilution correction and production rates. The main components of the module framework are the flat field correction module which first deals with removing image artifacts, followed by the background correction module that removes cell fluorescence that did not originate from the respective cell. The cells are then segmented and tracked using a modified version of CellProfiler (Carpenter et al. 2006). Problems in the segmentation and tracking and several other issues in the image area are then corrected using the post processing modules. The Cell Lineage module is used to obtain cell lineage assignments for each cell born. The Masking and filtering module is used to filter problematic or low quality cell data tracks. Finally, the dilution correction and production rates module is used to create tracks corrected for dilution effects, and produce production rates from those tracks.

Figure S5.7. Modules used in the image analysis pipeline.
5.8.1.3 Image Corrections

**Flat field correction:** A flat field correction map is calculated to account for the observed non-uniformity of the images. A set of images was acquired at the beginning of each experiment from a field of view devoid of cells that contains only a homogenous medium (SCD). These images were taken with increasing exposure times of: 0, 10, 50, 100, 200, 400, 800, 1200, 1600, and 2000 milliseconds. Usually, 6 fields of view repeats were done, in which the intensities at these exposure times were measured. A linear regression over the exposure times and their respective pixel intensities was calculated for each pixel, in order to obtain the pixel’s gain (the linear regression slope - amount of signal given by the detector as a function of the amount of light) and offset (y axis intersection, auto fluorescence of medium). This resulted in correction maps: one specifying the gain correction over an entire image and one specifying the offset correction over an entire image. The gain and offset correction maps were then smoothed using a Gaussian filter with a 32 and 16 filter size, respectively, in order to remove noise effects in the flat field images. Finally, a normalized gain map was calculated by dividing the gain map by the average gain over all pixels. Each image in the experiment was then corrected by subtracting the offset of each pixel and dividing by the gain. Any pixel values that were reduced below zero were set to zero.

**Background Correction:** After flat field correction, we corrected the images for fluorescence background resulting from auto-fluorescence of both medium and surrounding cells. To remove this fluorescence without removing fluorescence intrinsic to the cell, the fluorescence intensities in proximity to cells were measured and subtracted from the cells. In order to subtract the background, each field corrected image was roughly and widely segmented according to the mCherry to obtain a cell mask (using CellProfiler) with a wide margin around the cells (Fig. S8). A subtraction mask was created by dividing each image (mCherry and YFP separately) into 32 by 32 pixels blocks. For blocks with very few cell pixels (at least 85% of the pixels within it occupied by background pixels, pixels not inside the identified cell mask) the median value of the background pixels was assigned as the value of the subtraction mask block. For the rest of the blocks (blocks containing mostly cells), the background was computed according to the mean of the computed blocks surrounding it. In every case, the next block to be computed was chosen as that with the most computed blocks surrounding it. Thus, blocks containing mostly cells were assigned subtraction values composed of the subtraction values of blocks close to them, since such blocks contain very few background pixels that cannot be extracted reliably. Finally, in order to create a smooth subtraction, the subtraction mask was smoothed using a Gaussian filter of size 32. The subtraction mask was then subtracted from the original image. Any pixel values that were reduced below zero were set to zero. Notably, the
background fluorescence intensity represents a rather small fraction (~2%) of the cells’ fluorescence intensity.

**Bleaching correction:** A photobleaching curve was acquired to model the effects of light exposure on the fluorescence trajectories. To obtain a photobleaching curve, a field of view of cells expressing YFP and mCherry were subjected to constant exposure of fluorescent light. Resolution of image acquisition was 1.5 second. The mean fluorescence over a region of interest was plotted against time and this data was fitted to an exponential function to obtain the decay rate, γp, for YFP and mCherry, separately, according to Cookson et al. (Cookson et al. 2005). We obtained a negligible correction term, of less than 0.015% for YFP and 0.02% for mCherry (the decay rate γp equals 0.0015 for YFP and equals 0.002 for mCherry). Therefore, photobleaching correction was not applied to the data.

![Image](image.png)

**Figure S5.8. Background fluorescence subtraction.** To remove background fluorescence, we used the mCherry flat field corrected images (a) and roughly and widely segmented to create a cell mask (b). Images were divided into blocks, calculating median background fluorescence in blocks that contain little (less than 15%) cell fluorescence (c) and obtaining subtraction values of cell blocks from the values of blocks surrounding them (these subtraction masks were then smoothed) (d). The images were subtracted according to the subtraction masks resulting in background corrected images (e).
5.8.1.4 Cell Segmentation and tracking

Image analysis was performed using a modified version of the CellProfiler software, which provides high quality image segmentation and tracking capabilities. Cells were segmented using CellProfiler's automatic adaptive thresholding of the mCherry fluorescent signal, followed by CellProfiler's watershed algorithm to separate clumped cells. Background corrected mCherry fluorescent signal was used for segmentation, as this fluorophore is driven by the same promoter in all of our strains. Modifications to CellProfiler include the handling of misshapen cells (usually as a result of segmenting two cells as a single cell) by iteratively increasing the strength of the mCherry threshold and watershed attempts on the result. Other modifications were used to reduce computer memory usage. A fixed set of segmentation parameters for all experiments were empirically chosen. Following segmentation, the tracking of cells in CellProfiler was performed by pixel overlap of cells across time. The measured parameters for each cell in every time point were: mean, median and total fluorescence intensity (both for mCherry and YFP), pixel area, cell eccentricity (defined as the ratio of the distance between the foci of the ellipse, with the same second moments of the cell, and the ellipse's major axis length), and x and y centroid locations.

5.8.1.5 Segmentation and tracking post processing

We developed further processing steps to overcome problems that remained with the tracking and segmentation process above. The major modules of the post-processing in their order of operation are:

- Trim invasions module – used to remove data tracks of cells invading from outside the current field of view, as well as to remove data tracks of cells affected by such invasions. This module also filters cells that are improperly tracked.
- Split module – merges cells that are incorrectly split into two cells in the segmentation process, mainly due to vacuoles. An example for a split correction is seen in Figure S5.9.
- Merging module – corrects data tracks of two cells that incorrectly merge for a short time into a single cell.
- Interpolate intensity drops module – identifies outlier fluorescent intensity values and interpolates them using intensities from neighboring time frames.
- Tracking errors correction module – disconnects cell tracks that are incorrectly merged together in time in the tracking process (or incorrectly merged in the gapping module in the second run of the module).
- Clean outliers module – cleans outliers in data tracks due to segmentation errors based on cell area data tracks.
- Gapping module – connects data tracks of two cells which are actually a single cell with incorrect tracking.
- Filtration modules – several modules that filter problematic data tracks.
Remove cells with no neighbors: cells that appeared in some time frame without any neighboring cells next to them in their birth frame were removed (as a cell cannot be born without a nearby parent). This may happen due to poor medium trapping, resulting in cells that drifted far from their previous location in a single frame and therefore were not tracked correctly.

Remove cells with high birth signal intensity: when a bud is formed, it has a relatively lower area and mCherry signal intensity than an adult cell. Cells that first appear with a high birth intensity and large area are usually not born at that frame, but are usually a continuation of poorly tracked cells. Therefore, if a cell had a significantly high total mCherry intensity and a large area at its point of birth compared to the rest of the cells at their birth frame (Z-Score of total mCherry birth intensity higher than 2 and Z-Score of cell area at birth higher than 2), its data tracks were completely filtered.

Remove cells on edges: cells in close proximity to the edge of the field of view may produce poor segmentation and incorrect signal as usually part of their volume is outside of the field of view. Therefore, the part of the data tracks in which a cell appears too close to an edge (defined by a 14 pixel margin) was removed to avoid incorrect data tracking in those frames.

Remove cells that appear in too few frames: Cells appearing in only one or two frames were completely filtered, as these cells are usually leftovers of other cell tracks that were not combined with these tracks.

5.8.2 Cell Lineage
One of the prominent advantages in dealing with single cell data is the ability to track cell lineages. This information was obtained by discovering the most likely parent cell assignment for each new bud in each field of view using the Cell Lineage algorithm. The basic idea of the algorithm (Fig. S5.9) is to compute scores for each of the possible parents, and then assign the best scoring parent for each of the buds, taking into account the effect of the assignment on the scoring of the other bud-parent pairs. This is done by optimizing for the best scores over many parent-bud pairs.
Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

Figure S5.9. Parent assignment scheme. Scores are computed for each mother daughter pair. These scores are optimized to obtain the best pairs. These pairs are later used to create better scores.

Computing the scores: For each bud, a score was calculated for each of its potential parents (its neighboring cells at the time of its birth). The scores were based on the cell cycle of the parent cell. As a bud emerges, it is tracked separately from the parent cell, and thus, the fluorescent protein produced by the parent decreases due to diffusion to the emerging bud. This happens from the synthesis stage, through the G2 phase until mitosis, when the parent and bud disconnect. At the start of the next G1 phase, the parent cell again accumulates new fluorescent proteins. This scheme was illustrated well by Cookson et al. (Cookson et al. 2005) and is shown in Figure S10 over real mCherry total intensity. The S phase (synthesis) and the M phase (mitosis) were used as reference points to define the cell cycle of a parent. These two points were identified around the birth of each bud for each of its possible neighbors at the time of its birth.
Figure 5.10. **Cell cycle intensity fluctuations.** Shown are mother and daughter cells and their total mCherry signal. At the start of the mother’s cell cycle, the signal increases. From the S phase (synthesis) to the M phase (mitosis), mCherry diffuses towards the daughter cell.

The behavior of these neighbors around these points helped to create scores defining how likely these neighbors were as the parent of the emerging bud. In order to identify the possible cell cycle points in each neighbor, the total mCherry signal was modified to better display the cell cycle trends. Outliers in the signal were smoothed as depicted in the post processing but using a smaller moving window of 5 frames. The signal was then smoothed using a moving window of 5 frames. Next, the period in which the cell itself buds from its parent was removed (filtering until the point in which the cell’s area Z-Score at the first 150 minutes of life rises above -1), as in this period a cell cannot produce a bud, and the swift rise in total intensity at that period would have strongly influenced the global smoothing described next. The general signal trend of the cell was then obtained using a strong local regression smoothing (loess smoothing). This general signal was subtracted from the signal to produce a signal that was not affected by the general trends of the cell, but mostly by the cell cycle. To make the signal comparable between cells, the result was then scaled according to the signal’s mean and standard deviation. This signal was then used to find local maxima peaks around the birth of a bud to find the S phases (54 minutes prior to the birth and up to 27 minutes after the birth) and minima peaks to find the M phases (23 minutes prior to the birth and up to 70 minutes after the birth). Note that due to segmentation issues, it may happen rarely that the M phase occurs before the birth, or the S phase occurs after the birth, but an S phase occurring prior to the M phase was not allowed. Also, a 36 minutes minimum time was set from the cell’s birth until its first S phase.

Several parameters were used to create the score for each of the possible parents:
Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

- Time from S phase to bud birth.
- Time from bud birth to M phase.
- S phase peak shape – Scaled signal values of the potential parent from 20 minutes prior to the S phase peak until 20 minutes after the S phase peak.
- Eccentricity of potential parent cell around the S phase (at 20 minutes prior to the S phase peak until 20 minutes after the S phase peak), as the shape of a parent cell is somewhat eccentric close to the budding point.
- Area of potential parent around the S phase (at 20 minutes prior to the S phase peak until 20 minutes after the S phase peak), as a parent cell slightly expands prior to the S phase and contracts slightly after the S phase.
- M phase peak shape – Scaled signal values potential parent from 20 minutes prior to the M phase peak until 20 minutes after the M phase peak.
- Neighboring time - Fraction of time bud and potential parent neighbor each other (less than 5 pixels distance between cell edges) from bud birth to M phase (as the parent and bud are attached until the M phase, this number should be very close to 1, and therefore a neighboring fraction less than 0.4 was not accepted).
- Eccentricity of potential parent from 20 minutes before the bud birth until its birth.
- Ratio of the area between the potential parent and the bud from 15 minutes prior to the M phase up to the M phase. There is a certain ratio between the parent and the bud as the bud grows which this property tries to capture.
- Total mCherry intensity ratio (outlier cleaned and smoothed) between the potential parent and the bud from 15 minutes prior to the M phase up to the M phase. This ratio should increase in time as the protein diffuses into the bud, resulting in a decrease in the intensity of the parent and an increase in the intensity of the bud.

To use the above parameters to identify the real parent out of the neighboring cells, we sought to use a Naïve Bayes classifier, which models the distribution of these parameters for real parents and non-parent neighboring cells. The score was defined as:
The probability distribution of all parameters was taken to be a Normal distribution. Some parameters were computed from multiple time points (such as the S phase peak shape defined across 40 minutes), by geometrically averaging the probabilities over all time points. The distribution in each point was modeled over averaged values (median and standard deviation over 75% of the values around the median).

As the positive set, we took parent-bud pairs whose bud only had one neighbor during their birth (and thus this neighbor must be the parent). In addition, we required that neighbor to be a neighbor for at least 80% of the period between the bud’s birth and the M phase, and we required that it was not weaker in intensity than the bud at any point during that time. In the choice of the negative set, it is important not to enter true parents into it. Therefore, only neighbors of buds with 4 or more neighbors were considered into the negative set. The best scoring neighbor out of those was not entered into the negative set, considered as the true parent, and the rest were entered. The probability of each neighbor to be a parent was then calculated using the positive set distribution (score calculated only as: \( P(X|M = +) \)). Thus, the probability of entering a true parent into the negative set in each choice was at most 0.25 (if we consider the classifier according to the positive set to be at least as good as a random classifier).

Several penalties were used to penalize empirically unlikely events:

- If the neighboring ratio parameter was lower than 0.6, 2 standard deviations were added to the score (standard deviations according to the middle 90% of the scores).
- If the distance from birth to M phase parameter had a Z-Score greater than 3, a single standard deviation was added to the score.
- If the distance from S phase to birth parameter had a Z-Score greater than 3, a single standard deviation was added to the score.
In order to decrease the search space for possible parents, parents that had a Z-Score (built from all the scores) above 4 were filtered. Also, only up to 4 possible parents were allowed per bud (the best scoring ones were kept). A score for a no-parent assignment was set as a score having a Z-Score above 5. That way, only when there was no other choice, no parent was assigned.

Score optimization: Since the optimal score does not always reflect the correct assignment, a certain optimization on the scores was needed, instead of greedily picking the best scoring parent for each bud, as depicted in figure S5.11. The optimization was not global across the entire movie, since this was not computationally intractable. Therefore, a local optimization was done for each new bud. The optimization considered parent to bud assignment scores, and also other nearby bud scores that might be affected by the parent assignment (up to a certain degree). This way, for each bud born, we found the best parent assignment, which also did not hinder the parent assignment of the buds that could have been affected by its assignment. For each new bud (denoted as the center bud), buds affected by it were chosen as buds born up to an hour after its birth and in layers of influence around it. Layers of influence refer to layers of bud assignments around the bud that might be affected by the center bud. The first layer consisted of buds that shared the possible parents of the center bud. The second layer consisted of buds sharing the possible parents of buds in the first layer, and so on, up to 5 layers of influence. This resulted in a search space consisting of all the possible parent assignments for each of the buds affected up to 5 layers from the center bud.

A valid assignment was defined as an assignment in which no two buds were assigned the same parent at the same time. The optimal valid assignment was found for the entire search space, and only the center bud was assigned a parent. After assignments were made, the scores were calculated again, but with the positive set taken as the assigned parents, and the negative set as the incorrect parent of those same buds.

**Figure S5.11. Parent assignment possibilities.** An example of possible parent cell assignments. (a) Scores for each of the possible parent assignments for cells 3 and 4 (the lower the score, the better). (b) A simple greedy assignments – assigning best possible parent for cell 3, and then the best possible parent for cell 4 (cell number 1 is already assigned to cell 3, and as a cell can only have one bud at a time, cell 4 is
assigned no parent), resulting in a score of 12. (c) Optimization on the scores, to assign the best possible scores for each bud, assigning cell number 2 as the parent of cell number 3, and cell number 1 as the parent of cell number 4, resulting in a score of 4.

In order to further improve the scores, the distribution of the time between cell cycles and time from the birth of a cell until its first cell cycle were collected from the previous assignments. Empirical probability density functions were created for both measures for the positive set. For the negative set, a uniform density function was created as these measures are meaningless for a negative set. Thus, in the second assignment iteration, the scores for parents budding for the first time were multiplied by a constant divided by the empirical probability density function collected previously. The same was done for the time between cell cycles according to time distance to assignments already made. This algorithm was tested on ~200 manually tagged mother-daughter pairs, and had an accuracy of 0.96.

5.8.2.1 Masking and Filtering
A conservative framework for filtering the data was created to ensure that only high quality data is kept in the end. The following are the standard filtrations:

- Mask last known cell cycles: As cell tracking ends abruptly due to the end of an experiment, the last cell cycle of all cells is incomplete. For the majority of our analyses, we required complete cell cycles and therefore the last cell cycle of every cell was masked.
- Mask initial growth: The initial growth stage of a bud while it is attached to its parent was usually of no interest to us, as we attributed the protein production at that time to the parent cell. Therefore, the time a cell spent as a bud was masked.
- Mask abnormally large and strong cells: Rarely, there are cells that grow to abnormal sizes. These cells usually begin their life as normal cells, and at some point in time grow to large proportions. To avoid such extremities in the analyses, these cells were filtered from the point of their abnormal growth. These points were found by finding cells with an area Z-Score larger than 2 and a total fluorescence intensity Z-Score larger than 5 for more than 5 consecutive frames.
- Mask missed cell cycles: As some cell cycles are not detected by the lineage, the cell cycles that follow appear as very long cell cycles. Usually during these incorrectly long cell cycles, the missing S phase is detectable. Within these cell cycles the probability for an additional S phase peak was calculated. If an additional S phase was found with high probability, the two cell cycles were filtered.

Long cell cycles were found by finding cell cycles longer than 1.5 times the median cell cycle length in the population. Cell cycles longer than 2 times of the median cell cycle length were automatically filtered as being too long. The cell cycles in between
these lengths were checked for an additional S phase peak. The probability for an S
phase peak was built from several parameters used in the Lineage algorithm:

- S phase peak shape
- Eccentricity of cell around the S phase
- Area of cell around the S phase

The probabilities according to these three parameters were sampled from S phases
set by the Lineage algorithm. The median and standard deviations were taken from
the top 80% sample in order to consider only high probability S phases, which are
more likely to be true. Then, the threshold for S phase peak was set as the median
minus two standard deviations. Long cell cycles were filtered in cases in which the
probability for the additional S phase to occur passed the threshold, within its
expected time frame in the cell cycle. The expected period for the additional S phase
was set from the time of the previous M phase plus the minimum empirically
time checked G1 time and until the time of the S phase of the cell cycle minus the
minimum empirically checked interphase time.

**Mask cell cycle arrest:** Due to high cell density within a field of view many cells
enter cell arrest. These occurrences were found and filtered by searching for cell
cycles longer than 3 times the median cell cycle length in the population. In a cell
arrested field of view, it is probable that cells not in a cell arrest also suffer in some
way from the conditions in the field of view. Therefore, in field of views in which
more than half of the cells were in a cell arrest, all the cell cycles from that point
onwards were filtered.

**Mask negative protein production cell cycles:** Another strong indication of missed
cell cycles are cell cycles that have a long period of negative mCherry production
(see no dilution section). As the mCherry has very slow degradation time, a
negative protein production can only happen due to protein dilution to the bud.
This dilution should be accounted for in the calculation of the non-diluted total
protein fluorescence, and the primary reason for such a drop, is a missed cell cycle.
Protein dilution occurs during the bud formation between the S phase and M phase,
and therefore the typical drop length was calculated as the median length of time
between the two phases. If within 33% of a window of such size negative
production occurred, the cell cycle in which the drop occurred was filtered.

**Mask low quality cell tracks:** During the post processing framework some cells have
more problems than most, e.g., cells with large vacuoles might split many times. The
quality of these cell tracks is questionable as the post processing framework cannot
always find a perfect solution. Such cells, exceeding certain thresholds (defined
below) on the amount of problems that occur in the post processing, were
considered as low quality cells and filtered from further analysis. For a cell to be
considered as a low quality cell, a cell needed to pass any of the following criteria:

- Split at least three times (found in the Splitting module).
• Disappear and reappear at least three times (in the Gapping module).
• Having at least 3 long intensity drops (of at least 3 frames).
• A tracking error of at least 3 frames length occurred.

**Mask growth rate:** In order to achieve relatively constant conditions, we aimed at keeping the growth rate at certain limits within a strain, filtering periods in field of views where the median growth rate differed from a reference growth rate obtained from all the field of views. We defined the growth rate at a certain point in time as the median of the cell cycle times at that point in time. To obtain a reference growth rate, we calculated the growth rate for each field of view at a time reference between 150 and 420 minutes into the experiment, which is usually a time of stable growth and low cell density. The reference growth rate was set as the median over the median growth rate over time of each field of view. The limit for field of view growth rate was set as up to 15% distant from the reference growth rate. Only time periods within a field of view in which the growth rate was within the limits for a time period were kept. The length of these periods needed to be at least twice the reference growth rate so as to keep only large periods of almost constant growth rates. For periods shorter than 4 times the reference growth rate, between periods of growth rates within the limits, slight deviations of up to 22.5% from the reference growth rates were allowed. Note that in order to complement this work in terms of comparisons between repeats, different strains and different experiments; we also manually checked that the resulting growth rates were similar.

**5.8.2.2 Dilution correction and production rates**

In order to perform analysis on protein production rates, we defined the protein production rate as the average amount of protein accumulated over a certain period of time. However, the total protein fluorescence of a cell provides information only on the protein accumulated in a cell without accounting for the protein and transcripts lost through dilution to its buds (mCherry and YFP degradation are negligible in relation to dilution. This was verified by observing negligible degradation in mCherry and YFP after addition of a translation inhibitor to the media). As the bud transcripts and cell environment are transported from the parent through active transportation and diffusion, we considered the total protein accumulated in the bud until the mitosis separation as proteins produced by the parent. Although the bud is a part of the parent until the mitosis point, it is considered as a new cell by the image segmentation and tracking. Therefore, the protein accumulated in the bud until the mitosis point was added to the protein accumulated in the parent as depicted in Figure S5.12, to obtain the protein accumulated in the parent until the mitosis point while correcting for dilution to the bud. The corrected accumulated protein was then smoothed for outlier points in the signal (as done in the post processing, in the ‘Clean outliers module’, but with a smoothing window of size 5).
After calculating the corrected accumulated protein in a cell for each time point, the calculation of the protein production rate from time $T$ to time $T+\Delta T$ for a single cell was defined as:

$$\text{Production rate} (T \ldots T + \Delta T) = \frac{\text{Accumulated protein}(T+\Delta T) - \text{Accumulated protein}(T)}{\Delta T}$$

Prior to the production rate calculation, the accumulated protein signal was smoothed separately in each of the time windows in which the protein production rate was calculated using Matlab's Loess smoothing, with a span set to 0.4 of the data points in the window, such that data from a certain time window did not leak to its adjacent windows.

![Graph showing corrected accumulation of protein over time](image)

**Figure S5.12. Correcting for dilution.** An example of how the effects of dilution are corrected. The bud total intensity values (dark blue) are added to the total intensity of the parent (red) from the point of the bud’s appearance until the mitosis point (first blue dot) to produce the parent’s total intensity values, corrected for dilution (dark red). The rest of the lines are additional daughter cells, and are also used for dilution correction.

### 5.8.2.3 Yeast Strains

A shared master strain that contains the YFP reporter gene, the HIS3 proximal promoter (100bp upstream of the ATG), and the mCherry fluorescent protein downstream of the TEF2 promoter, was used to construct the library. The variable part of each promoter were synthesized by Biomatik and inserted into a master strain by genomic integration, as previously described (see [Raveh-Sadka et al. 2012](#) for details). Additional strains with mutated Gal4 binding sites were designed to span a range of weak medium and strong affinity to Gal4, based on previous studies by Ptashne et al. as follows:
5.8.2.4 Flow Cytometry data analysis and Gamma distribution analysis

An automatic data analysis pipeline was applied for gating and filtering the data and to remove outliers. Wells were considered outliers having abnormal forward scatter, side scatter or mCherry distributions. Cells collected in the first or last 0.5 seconds in each well were discarded from further analysis, as well as cells for which negative or saturated values were measured in one of the parameters. Cells collected over periods that show flow instability (bubbles, etc.) were also removed. Automated gating based on FSC and SSC values was implemented to minimize cell heterogeneity and reduce extrinsic variability. The gating algorithm selects for the lower population that is enriched for G1 cells. The final gated YFP values were normalized by mCherry.

5.8.2.5 Fluorescent proteins

A direct and accurate measure of promoter activity over time was achieved by real-time monitoring the fluorescent protein levels as a measure of promoter activity. The YFP protein was very stable over the acquisition time (>15 hours), showing no degradation, also after addition of cycloheximide to block translation. Protein accumulation, corrected for dilution from mother to daughter cells, showed a constant increase (Fig. S5.13) and was considered as an accurate measure for promoter activity.
Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

Figure S5.13. Individual cell tracks, corrected for dilution between the mother and the daughter cells. Cell tracks are shown for 2 strains: strong binding site affinity (purple) and weak binding site affinity (pink). The tracks show robust continuous increase in signal intensity, and were differentiated to yield the production rate (as described in the main text).

5.8.2.6 Normalization between experiments and wells
As the mCherry distribution should be equal across all wells in all experiments, we made sure that all mCherry rate distributions have the same mean mCherry rate. This was done by measuring the mean mCherry rate from a reference well, and multiplying all mCherry rates of a well by a factor such that its new mean mCherry rate will have the same mean mCherry rate as the reference well. A multiplication factor between the different mean mCherry rate distributions reflects a global factor that should also be reflected in YFP changes as well. Therefore, the YFP rates were also multiplied by the same factor multiplying the mCherry. Displayed in figure S5.14 are corrected mCherry rate histograms and their common distribution (a distribution created from an equal amount of data points taken from each strain). It can be seen that the mCherry rate histograms of different strains are very similar, and fit the common distribution.
Figure S5.14. Corrected mCherry rate histograms. Representative mCherry rate histograms normalized between wells and experiments. Depicted in red is a distribution of mCherry rates of all strains in the example.

5.8.2.7 Autocorrelation of promoter production rates
Normalized YFP rates were calculated for each cell at a resolution of 3 minutes. Negative rates were filtered. The top and bottom 0.5% values of the data for both mCherry and YFP were considered as outliers. Cell cycles containing at least 50% unfiltered YFP and mCherry rates were used. A cell cycle not containing at least 30 minutes of data was not used. Autocorrelations were calculated at a resolution of 6 minutes, averaging the temporal information to reduce noise. Autocorrelation at 3 minutes resolution (an example shown in Figure S5.15) yielded lower autocorrelation values, however the relation between the different strains was kept the same. Autocorrelation lag information smaller than 14 data points was not used. Then, all autocorrelations across all cells were averaged. Correlations between consecutive time windows were calculated by averaging the values across each resolution (e.g. for a resolution of 9 minutes the values of windows i, i+1, i+2 were averaged), and then computing the correlation of the consecutive windows in a single cell. Then, all the values across all cells for each resolution were averaged. Values filtered were not used for the correlation. Correlations with less than 28 data points divided by the current resolution (where a resolution of 3 minutes is considered as 1 as this is the system’s minimal resolution), or less than 4 data points, were not used. To verify that the strains differ in their autocorrelation of YFP rates and not the mCherry rates, we calculated the autocorrelation also for YFP and mCherry rates alone. We clearly observed that for all the strains there was no
significant change for between the autocorrelation curves of the mCherry, as opposed to the autocorrelation curves of the YFP rates.

**Figure S5.15. Autocorrelation of normalized YFP production rates at 3 min resolution.** The autocorrelation was calculated across thousands of different cell traces for each promoter variants. Bars denote standard errors.

**5.8.3 Model of stochastic gene expression**
We use a two-state kinetic scheme in which the promoter switches between an active and inactive promoter state, and has transcription, translation, mRNA degradation and protein degradation (Fig. S5.16).

**Figure S5.16: Kinetic model of gene expression.** The scheme represents a promoter that switches between a transcriptionally inactive and active state. On- and off switching happens with rate $K_{on}$ and $K_{off}$ respectively. Transcription, translation, mRNA and protein degradation rates are not shown, but are taken into account in all simulations.

To analyze the temporal dynamics of gene expression we use a stochastic simulation of this model (Gillespie 1977). For steady-state gene expression and cell-to-cell variability (noise) we use an analytical solution of this same model (Sanchez et al. 2011), which was solved using the master equation.
For both models we use the following rate parameters: (All rate parameters are in minute⁻¹)

**Protein degradation.** YFP was found to be highly stable (see fluorescent protein paragraph). We therefore assume that protein degradation comes only from dilution. Since we measured the doubling time of our strain (~90 min), we chose to fix the protein degradation rate. We set the degradation rate to \(\ln(2)/90 = 0.0077\).

**Translation rate.** Yeast, under fast growing conditions, has a protein production rate of between 6500 and 19500 proteins/cell/sec (von der Haar 2008). Yeast has around 60,000 mRNAs/cell (Zenklusen et al. 2011). This gives a translation rate between 6.5 and 19.5 proteins/mRNA/min. We set the translation rate to 10 proteins/min.

**Transcription rate.** The transcription rate is the rate of production of stable mRNAs while the promoter is in each respective state (on or off). Experimentally measured transcription rates combine both on and off promoter states. Therefore, the total rate of transcription for a gene is the sum, for all states, of the fraction of time spent in that state times the transcription rate in that state. The upper bound for expression rate is between 4 transcripts/minute and 10 transcripts/minute (Pelechano et al. 2010). We note that these rates represent the combination of on and off promoter states. We set the transcription rate to 5.

**mRNA degradation.** The median mRNA half-life was measured to be ~20 min (Wang et al. 2002). We therefore set the mRNA degradation rate to \(\ln(2)/20\).

**Promoter switching rates (Kon, Koff).** Promoter switching is a function of TF binding and unbinding, however each (un)binding event does not necessarily switch the promoter to another state. For this reason promoter switching is much slower than TF binding kinetics. Promoter switching was measured to be in the range of 1e-3 to 1e1 (Zenklusen et al. 2011; Octavio et al. 2009; Tan and van Oudenaarden 2010). For the “normal” promoter we use \(2^{-3}\) min⁻¹ and \(2^{-2}\) min⁻¹ for Kon and Koff respectively. For the “fast” promoter we use \(1^{-2}\) min⁻¹ and \(2^{-2}\) min⁻¹, for the “slow” promoter we use \(2^{-3}\) min⁻¹ and \(2^{-3}\) min⁻¹ for Kon and Koff respectively.

**5.8.3.1 Stochastic simulation of fast and slow promoter dynamics**

We simulate three regimes: normal, slow and fast dynamics. The “fast” regime has a higher Kon to simulate an increased accessibility with the reduction of nucleosome coverage as a result of the addition of a poly(dA:dT) element. The “slow” regime has a lower Koff to simulate the increase in transcription factor binding affinity. Figure 5.1 shows three example runs in which the promoter switches between the ON and OFF state, and transcription and translation occur
Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

when the promoter is ON. The protein production shows the bursting behavior of gene expression. The figure illustrates that protein production is correlated with promoter dynamics, however also that protein production has a delay compared to promoter on-switching and that, due to chance, not every on state results in a protein production burst.

5.8.3.2 Autocorrelation analysis of simulated protein production

We found that, in experiment, increasing the binding affinity or decreasing the poly(dA:dT) length both increase the autocorrelation (Fig. 5.2). To investigate if this change in autocorrelation is expected given the simple model of promoter state switching, we performed the same autocorrelation analysis on simulated data obtained from the Gillespie simulations. We model the increase in binding affinity by a decrease in Koff and the decrease in polyT length by a decrease in Kon. Figure S5.17 shows that indeed we observe the same increase in autocorrelation when we decrease Koff (increase affinity) or decrease Kon (decrease polyT length).

![Figure S5.17: Autocorrelation analysis on simulated protein production.](image)

We use the stochastic simulation data to look for autocorrelation changes when changing either Koff (a) or Kon (b). We observe that decreasing either Koff or Kon significantly increases the autocorrelation, in line with our experimental observations. Values shown for Koff and Kon are in min⁻¹.
5.8.3.3 Limits of measuring promoter dynamics through protein production

In order to systematically investigate how promoter dynamics relate to protein production bursting we use the above stochastic simulation and gradually increase either Kon or Koff and quantify protein production bursting. We quantify the burst frequency as \( P(\text{on}|\text{off}) \) and burst size as \( P(\text{off}|\text{on})^{-1} \). \( P(\text{on}|\text{off}) = \frac{N_{\text{off} \rightarrow \text{on}}}{t_{\text{off}}} \) and \( P(\text{off}|\text{on}) = \frac{N_{\text{on} \rightarrow \text{off}}}{t_{\text{on}}} \). Where \( N_{\text{off} \rightarrow \text{on}} \) is the number of off to on event and \( N_{\text{on} \rightarrow \text{off}} \) vice versa. \( t_{\text{off}} \) and \( t_{\text{on}} \) is the total time in the off or on state respectively.

To simulate our microscopy experiments as best as possible we sample the protein level every 6 min. We note however that our simulation does not capture every process that affects the delay between on-switching, transcription and translation, such as mRNA maturation and protein folding. However, adding such delays would not change the qualitative result of our simulations.

Figure S5.18 shows how changing promoter ON and OFF switching (\( \text{Kon} \) and \( \text{Koff} \) respectively) affects mean expression and the derived promoter dynamics from protein production bursting. The figure shows that there is a regime in which promoter \( \text{Kon} \) and protein burst frequency change linearly with respect to each other. However, at higher \( \text{Kon} \) the protein burst frequency saturates. For \( \text{Koff} \) the same happens, but now both at very low and very high \( \text{Koff} \). There appears to be only a limited regime where changes in \( \text{Koff} \) are linear to changes in the observed protein burst size. However, in the biologically meaningful regimes of \( \text{Kon} \) and \( \text{Koff} \), the simulation shows that changes in \( \text{Kon} \) and \( \text{Koff} \) can be detected, at least qualitatively, by measuring changes in protein production bursting. The fact that there exists a regime in which this measurement is non-linear could explain why, in our microscopy experiment, we measure a smaller change in bursting than is expected from the change in mean expression.
Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise.

Figure S5.18: Measured protein production bursting as a function of promoter switching, in a stochastic simulation. We measure mean expression, protein production burst frequency and protein production burst size as a function of changing \( K_{on} \) or \( K_{off} \). Increasing \( K_{on} \) increases the mean expression level (a) and increases the burst frequency, while burst size remains constant (b). However, there is a regime, at higher \( K_{on} \) (b, \( K_{on} > 2 \times 10^{-2} \)), where measured burst frequency increases non-linearly with increasing \( K_{on} \), while mean expression continues to increase (c). For changing \( K_{off} \) we observe a similar phenomenon. Decreasing \( K_{off} \) increases the mean expression level (d), but at higher and very low \( K_{off} \) (e, \( K_{off} > 1 \times 10^{-4} \) and \( K_{off} < 2 \times 10^{-2} \)) burst size changes non-linearly with \( K_{off} \), while expression continues to increase (f). This analysis shows that changing promoter dynamics can be measured qualitatively through protein production bursting, but
might be limited quantitatively in regimes where protein production bursting changes non-linearly with changing promoter dynamics.

5.8.3.4 Exploring the relationship between mean expression level and noise using an analytical model

We have changed the mean expression and noise using two different biological parameters, namely promoter accessibility (adding polyT) and binding affinity (TF binding site sequence). Using various analyses we show that each biological parameter changes the dynamics of gene expression in a different way. Accessibility changes the frequency of expression bursts and affinity changes the length of expression bursts. Using a stochastic simulation (Gillespie 1977) we show how promoter kinetics (on-off switching) can be modulated to give changes in protein production bursting. Various studies have connected this “bursting” to the noise properties of gene expression. Most notably is the gamma model by Friedman et al. (Friedman et al. 2006), where noise \( (\text{std}^2/\mu^2) \) is inversely related to the burst frequency and the burst size is related to the noise strength \( (\text{std}^2/\mu) \).

To understand how the expression and noise changes that we observe are related to promoter dynamics we use an analytical solution of the kinetic scheme that we used in the stochastic simulations (see below for the derivation), which gives us the mean expression level, noise and noise strength at steady state.

We set the parameters to biologically meaningful values (see above) and assume that changing accessibility changes \( K_{on} \) and affinity changes \( K_{off} \), where an increased affinity has a decreased \( K_{off} \) and an increased accessibility has an increased \( K_{on} \).

Figure S5.19 shows the mean versus burst frequency (and noise) plot of either changing \( K_{on} \) or \( K_{off} \). Increasing \( K_{on} \) increases expression and decreases noise (increases burst frequency) (Fig. S5.19b,c,e,f blue lines), which is in accordance with our experimental observations (Figure 5.4). Interestingly, for changing \( K_{off} \) we observe two different qualitative responses, depending on the rate of on switching (\( K_{on} \)). When \( K_{on} \) is relatively slow decreasing \( K_{off} \) increases expression and increases noise (Fig.S5.19b,c red lines). When \( K_{on} \) is relatively fast decreasing \( K_{off} \) increases expression and decreases noise (Fig.5.1 and S5.9e,f red lines). In our measurements we observe that increasing the binding affinity increases expression and increases noise. The model therefore predicts that if we would increase \( K_{on} \) (TF concentration and/or activity) that noise would decrease with increasing binding affinity. Figure S5.19d shows expression and burst frequency from gamma fit (inverse noise) for increasing the binding affinity, however now in activating condition (SCD-AA). When starved for amino acids the Gcn4 TF is induced. Exactly as the model predicts, burst frequency increases (or noise decreases) now with increasing expression.
Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

Figure S5.19: Measured and predicted change in burst frequency and noise.
We compare the measured change in expression and burst frequency with the predicted change in expression and burst frequency from the analytical solution of the kinetic scheme shown in figure 1. Burst frequency (a,b,d,e) is quantified as noise$^{-1}$ (c,f) (Friedman et al. 2006). In non-activating condition, where Gcn4 is lowly induced, we measured an increase in noise, or decrease in burst frequency, with increasing expression as a result of increasing the binding affinity of the Gcn4 site (a). The analytical model predicts that such behavior happens when two conditions are met: 1) the promoter is leaky, i.e. the "OFF" state has low, but significant, transcriptional activity, and 2) promoter switching is slow, i.e. $K_{on}$ is relatively low (b,c $K_{off}$ change: red lines). Changing $K_{on}$ in the model always results in an increase in burst frequency and therefore a decrease in noise (b,c,e,f $K_{on}$ change: blue lines). Given these theoretical conditions, the model predicts that the decrease in burst frequency, or increase in noise, disappears when promoter switching is relatively fast, i.e. $K_{on}$ is high (e,f $K_{off}$ change: red lines). In accordance with this prediction, when we repeat the experiment in activating condition (amino acid starved, to induce Gcn4), we find that burst frequency now increases (d), and noise decreases, with increasing expression as a result of increasing the binding affinity.
5.8.3.5 Frequency spectrum analysis

When the lifetime of mRNAs is short relative to the lifetime of the protein, as is the case for yeast, protein production can be assumed to occur in random uncorrelated events (Cai et al. 2006; Friedman et al. 2006). If protein production is indeed uncorrelated, we don’t expect to observe any frequencies, other than cell cycle, in the production over time. Figure S5.20 shows the frequency spectrum analysis of the YFP production in time. As expected the only frequency that we observe is around 1/90 per min, which is the frequency of the cell-cycle (measured doubling time). We note however that when mRNA production occurs in bursts, and that re-initiation occurs at some fixed time interval, we would expect to see this in the frequency spectrum. Since re-initiation is probably faster than the interval at which we measure the protein production and due to stochastic translation, we would not be able to pick up this re-initiation frequency.

Figure S5.20: Frequency spectrum analysis of YFP production in time. Shown is a frequency spectrum analysis performed using the Fourier transform of the autocorrelation of the YFP production in time, which was measured in the microscope. The shown spectra quantify the relative contribution of each frequency to the total signal. We find that the only frequency that is significantly present in the data is around 0.01 min⁻¹, which very likely stems from the frequency of the cell cycle (around 90 min, a,b arrow). Changing the polyT length (a) or
Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

binding affinity (b) does not change the frequency spectrum. For comparison spectra are shown for random sequences with different means (c).

5.8.3.6 Analytical solution of the kinetic model

We model stochastic promoter state switching, transcription and translation using the master equation (Sanchez et al. 2011). In this model promoter transcription factor (TF) binding and unbinding events determine the transitions between promoter states. Transcriptional activity changes when the promoter switches to a state with different transcriptional activity. The promoter states can have low (including zero) or high transcription rate, to describe in-active ("off") or active ("on") states respectively. Translation occurs in bursts with the probability of a burst described by a geometric distribution. The master equation (in matrix notation) takes the form:

$$\frac{d}{dt} \tilde{p}(n) = \left[ \hat{K} - \frac{b}{1+b} \hat{R} - n\delta \hat{I} \right] \tilde{p}(n) + \hat{R} \sum_{\beta=1}^{n} h(\beta) \tilde{p}(n-\beta) + (n+1)\delta \hat{I} \tilde{p}(n+1)$$  \hspace{1cm} (1)

Where $\tilde{p}$ is the vector of probabilities of having $n$ proteins in the cell for each promoter state $d/dt \tilde{p}(n)$ describes the time evolution of these probabilities. $\hat{K}$ is the matrix of promoter state transition rates, where $\hat{K}_{ij}$ is the rate of transitioning from state $j$ to state $i$ and $\hat{K}_{ii}$ is the sum over all outgoing rates from $i$ times -1. $\hat{R}$ is the diagonal matrix of transcription rates with $\tilde{r}_i$ on the diagonal ( $\hat{R}_{ii} = \tilde{r}_i$), where $\tilde{r}_i$ is the transcription rate of state $i$, $\hat{I}$ is the identity matrix. $b$ is the average burst size (proteins produced per mRNA), $\delta$ is the protein degradation rate. $h(\beta)$ describes a geometric distribution and is the probability of producing a burst of size beta.

To derive the mean protein abundance and variance we solve this system at steady state, thus for $d/dt \tilde{p}(n) = 0$. We get mean protein abundance:

$$\langle n \rangle = \frac{b \tilde{r} \tilde{m}(0)}{\delta}$$  \hspace{1cm} (2)
Where $\tilde{m}(0)$ is the zeroth partial moment of the distribution of mRNA abundance and is the solution to:

$$0 = \hat{K} \tilde{m}(0)$$

(3)

We can get noise ($\sigma^2/\mu^2$) and noise strength ($\sigma^2/\mu$) by deriving:

$$\langle n^2 \rangle = (1 + b) \langle n \rangle + \frac{b \bar{r} \tilde{n}(1)}{\delta}$$

(4)

Where $\tilde{n}(1)$ is the first partial moment of the distribution of protein abundance and is the solution to:

$$0 = (\hat{K} - \delta \hat{I}) \tilde{n}(1) + b \hat{R} \tilde{m}(0)$$

(5)

Variance ($\sigma^2$) is:

$$Var(n) = \langle n^2 \rangle - \langle n \rangle^2$$

(6)

Therefore noise ($\sigma^2/\mu^2$) becomes:

$$\langle \eta^2 \rangle = \frac{(1 + b) \langle n \rangle - \langle n \rangle^2 + \frac{b \bar{r} \tilde{n}(1)}{\delta}}{\langle n \rangle^2}$$

(7)

And noise strength ($\sigma^2/\mu$):

$$\langle F \rangle = \frac{(1 + b) \langle n \rangle - \langle n \rangle^2 + \frac{b \bar{r} \tilde{n}(1)}{\delta}}{\langle n \rangle}$$

(8)

We describe gene expression using a two state kinetic scheme that represents switching between an active (ON) and inactive (OFF) promoter configuration (Figure S5.16). The ON state is transcriptionally active (with rate $r_1$) and we allow the OFF state to have some (leaky) transcriptional activity (with rate $r_2$). $\hat{K}$ and $\bar{r}$ thus become:
Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

\[
\hat{K} = \begin{pmatrix}
-K_{\text{off}} & K_{\text{on}} \\
K_{\text{off}} & -K_{\text{on}}
\end{pmatrix}
\]  \hspace{1cm} (9)

\[
\hat{r} = \begin{pmatrix}
r_1 \\
r_2
\end{pmatrix}
\]  \hspace{1cm} (10)

Figure S5.21: Analysis of protein production bursting using the change in production rate. We define a production burst (ON state) as a period in which the change in YFP production is positive, i.e. the second derivative of the YFP
fluorescence in time. We then quantify the on-rate as the number of OFF to ON transitions divided by the total time in the OFF state. Similarly we quantify the off-rate as the number of ON to OFF transitions divided by the total time in the ON state. Shown is the measured off and on-rate (from microscope data) versus the mean production rate, for either changing the polyT length or the binding site affinity. We find that when changing the polyT length the on-rate changes more than the off-rate. When changing the binding affinity we find that the off-rate changes more than the on-rate. Measured rates are in min⁻¹.

Figure S5.22: Burst size and burst length as a function of changing the binding site affinity. We define a “burst” as a period of relative high (YFP) protein production. Shown is the average burst length and average burst size, measured in microscope, for strains with differing binding site affinity. We quantified the length, in minutes, of the periods of activity (a) and the size in terms of average YFP production (b).
Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise

5.8.4 Supplemental References


Chapter 6 Conclusions

This dissertation contributes to our understanding of gene regulation, both in terms of individual interactions as well as from a topological perspective, and how gene expression noise plays a crucial role in any gene regulatory system. Our contributions are both in the design and application of computational models that generate biological hypotheses as well as in testing them using specific wet-lab experiments.

The first contribution of this dissertation is in understanding how gene regulation is encoded in the wiring of the gene regulatory and protein interaction networks during HIV infection of the human host. We quantify regulatory structure using network topology and show that large-scale network properties and small-scale network patterns can explain specific biological behavior, such as human immune response and virus immune system evasion.

The second contribution of this dissertation is in understanding how individual gene regulatory interactions are encoded in DNA sequence in yeast. Specifically, we make use of the stochastic nature of gene expression to infer mechanisms of gene regulation from steady state protein distributions and temporal gene activity. We use computer simulations as well as wet-lab measurements to study the mapping between promoter DNA sequence, gene expression and gene expression noise, by measuring single-cell activity of a set of native, mutant and synthetically designed promoters in vivo and in silico.

We found that there are multiple ways to achieve the same change in the mean expression level of a gene, but with different effects on the cell-to-cell variability of gene expression in yeast. Our computer model suggests that this is the result of different, albeit DNA encoded, mechanisms of gene regulation. Notably, we found that different native promoters can have the same mean expression level, but vastly different amounts of noise by having different DNA sequence. Moreover, we found that the way in which noise changes with expression, when a gene is induced, is gene-specific and highly dependent on promoter DNA sequence.

We found that a combination of activation and repression can result in a unique transcriptional state of low expression and low noise, where usually low expression is accompanied with high noise. In addition we found that further noise reduction can be achieved by protecting against extrinsic noise from the regulator. Our computer simulations suggested that coupling between activator and repressor, for example when they are the same molecule, would reduce the sensitivity to fluctuations in the activity of the transcription factor. Measurements in the lab confirmed this prediction. To the best of our knowledge we are the first to identify such a mechanism.
In addition, we investigated the effect on the temporal dynamics of gene expression and steady state gene expression noise of two different DNA encoded mechanisms of gene expression change in yeast. In specific, we measured, using time-lapse microscopy and flow-cytometry, the gene expression activity of a set of synthetically designed promoters in which gene expression was increased through increasing the binding site affinity of the activating TF or by decreasing the binding affinity of nucleosomes that compete for binding with the TF. We hypothesized that the first mechanism would increase the time the gene was active and that the second mechanism would increase the frequency of activity. Computational models suggested that the first mechanism would increase the noise while the second mechanism would decrease the noise, where both mechanisms would cause an increase in mean expression level. Measurements of temporal gene expression in microscopy and steady state protein distributions in flow-cytometry confirmed our hypotheses.

The work presented in this dissertation has contributed to the understanding of how DNA sequences encode for gene expression, gene expression noise, and the interaction between genes, and in turn how gene interactions give rise to complex network structures. It will be very interesting to use this knowledge to develop a quantitative model that predicts, from DNA sequence, expression and noise in a gene network in which multiple genes regulate each other.


Summary

Genomes of even the simplest of organisms are comprised of many genes, the units of genetic information that are stored in the DNA code. One of the biggest challenges in the post-genomic era is to infer biological function from DNA sequence. A genome is like a musical composition. It is comprised of many parts, some of which are played at the same time while others must be played at different times. Therefore, the correct execution of a symphony requires an orchestra to play in harmony - the score dictates ‘who’ plays ‘what’ and most importantly ‘when’. Likewise, the observed complexity of biological functioning lies in the way genes are directed and organized. The functioning of a single gene only makes sense when we take into account ‘when’ the gene is active and ‘whom’ it interacts with.

In order to advance our understanding of gene regulation (when?) and gene interaction (with whom?) we have studied two different biological systems: Infection of the human host with HIV-1 and transcriptional regulation in baker’s yeast.

For HIV-1 infection we have studied the interaction network that arises from the multitude of cellular interactions that take place between viral and human genes and proteins during infection. We have found that the topology of this network, in terms of global structure and local re-occurring patterns can be used to study HIV-1 infection. In specific we found that the network structure shows how the cellular machinery of the human immune system tries to inhibit virus infection and replication, and that the virus in turn tries to evade the host’s immune response. Taken together our results shed light on how HIV-1 infection occurs by hijacking the host cellular machinery. In addition, using protein interaction data and a bioinformatics algorithm we predicted novel cell surface proteins in human lymphocytes that potentially interact with HIV. This set constitutes a well-founded starting point for experimental testing of cell and tissue susceptibility to different HIV strains.

To understand individual gene interactions and how control of gene expression occurs we studied transcriptional regulation in baker’s yeast by measuring the activity of genes in individual cells. We focused on the promoter region of the DNA to unravel the ‘language’ of transcriptional control. Assisted by a quantitative model that we developed, we found that phenotypic cell-to-cell variability, or ‘noise’, in gene activity between cells that have the same genotype is a result of specific transcriptional mechanisms that are encoded in the promoter DNA sequence. We showed that genes have a unique relationship between their transcriptional activity and their noise.
Next, using a micro-fluidic microscopy platform measuring single-cell expression of a set of synthetic designed promoters over time we investigated the effect of two different DNA sequence changes on temporal gene expression. We found that both adding a binding site of a transcriptional activator and adding a sequence that increases the accessibility of the DNA to activators increase transcriptional activity, however with opposing effects on the cell-to-cell variability. Evolution can therefore operate on several mechanisms for changing gene expression that result in systems with different stochastic properties. Taken together we found that even very small changes in the promoter DNA sequence can give rise to large differences in the noise of gene expression.

Hence, the work presented in this dissertation has contributed to the understanding of how DNA sequence encodes for gene activity and its variability between cells, as well as how interactions of genes give rise to complex network structures.
Zelfs de meest simpele organismen zijn opgebouwd uit vele genen, de eenheden van genetische informatie die opgeslagen zijn in het DNA. Een van de grootste uitdagingen in het post-genome tijdperk is het vertalen van DNA sequentie in biologisch functioneren. Een genoom is als een muzikale compositie: het bestaat uit meerdere delen, waarvan sommige tegelijkertijd gespeeld worden en anderen op verschillende tijden. Voor de correcte uitvoering van een symfonie moet een orkest in harmonie spelen; de partituur dicteert ‘wie’ speelt, ‘wat’ en vooral ‘wanneer’. Evenzo ligt de geobserveerde complexiteit van de biologie in de manier waarop genen gedirigeerd worden en georganiseerd zijn. Het functioneren van een enkel gen is alleen zinvol wanneer we rekening houden met ‘wanneer’ het gen actief is en met ‘wie’ het interacties aangaat.

Om ons begrip van genregulatie (wanneer?) en geninteractie (met wie?) te verbeteren hebben wij twee verschillende biologische systemen bestudeerd: HIV-1 infectie in de menselijke gastheer en transcriptie regulatie in bakkersgist.

Voor HIV-1 infectie hebben wij het interactie-netwerk bestudeerd dat ontstaat uit de vele cellulaire interacties die plaatsvinden tussen virus en menselijke genen en eiwitten tijdens infectie. We hebben ontdekt dat de topologie van dit netwerk, in termen van globale structuur en locale terugkerende patronen, gebruikt kan worden om HIV-1 infectie te bestuderen. Specifiek hebben wij ontdekt dat de netwerkstructuur laat zien hoe het cellulaire apparaat van het menselijke immuunsysteem het virus probeert te onderdrukken, en dat het virus op zijn beurt probeert de menselijke immuunreactie te ontwijken.

Alles bij elkaar genomen werpen onze resultaten licht op hoe HIV-1 infectie plaatsvindt door het cellulaire apparaat van de gastheer te kappen. Daarnaast, door gebruik te maken van een door ons ontwikkeld bio-informatica algoritme, hebben wij nieuwe cellmembraan eiwitten voorspeld die potentieel interacties aangaan met HIV. Deze set eiwitten vormt een goed startpunt voor het experimenteel testen van de vatbaarheid van specifieke cel-types en weefsels voor HIV infectie.

Om individuele geninteracties en de besturing van genexpressie beter te begrijpen hebben wij transcriptieregulatie in bakkersgist bestudeerd door de activiteit te meten van verschillende genen in individuele cellen. Wij hebben ons gericht op de promotor regio van het DNA om de taal van transcriptie regulatie te ontrafelen. Geholpen door een door ons ontwikkeld kwantitatief model, hebben wij ontdekt dat fenotypische cel-tot-cel variabiliteit, of ‘ruis’ in genactiviteit tussen cellen van hetzelfde genotype, een resultaat is van specifieke transcriptie mechanismen die gecodeerd zijn in de promotor DNA sequentie. Wij hebben aangetoond dat er voor elk gen een unieke relatie bestaat tussen zijn activiteit en zijn ruis.
Vervolgens, door gebruik te maken van een zogenaamd ‘micro-fluidic’ microscoop platform om genexpressie te meten in individuele cellen over de tijd, hebben wij het effect onderzocht van twee verschillende DNA sequentie veranderingen op temporele gen expressie. Wij hebben ontdekt dat zowel het toevoegen van een transcriptie activator als het toevoegen van een specifieke sequentie die het DNA meer toegankelijk maakt voor activatie, de transcriptie activiteit verhoogt, maar met tegengesteld effect op de ruis. Evolutie heeft dus de mogelijkheid om de activiteit van een gen te veranderen met veel of juist weinig ruis. Samengenomen hebben wij gevonden dat zelfs kleine veranderingen in de DNA sequentie van de promotor kunnen zorgen voor grote veranderingen in de activiteit van een gen.

Ter conclusie, het werk dat wij hebben gepresenteerd in dit proefschrift heeft bijgedragen aan ons begrip van hoe genactiviteit en ruis is gecodeerd in DNA sequentie, en hoe interacties van genen leiden tot complexe netwerk structuren.
Acknowledgements

First and foremost I’d like to thank prof. Peter Sloat for his supervision of my PhD. Peter has given me structure and direction in times that this was necessary but also given me a lot of freedom for pursuing collaborations, both in the Netherlands and abroad. He has taught me that great ambition and scientific independence come with great responsibilities.

In addition I’d like to thank my co-supervisor prof. Jaap Kaandorp. He was able to light the spark for my interest in biology and prepare me to be a true systems biologist.

Furthermore, I’d like to thank prof. Eran Segal who supervised me during my stay at the Weizmann Institute of Science. He was able to offer me a research environment of the highest level and a place where I could get experience in a wet-lab.

I am grateful to all the colleagues in the Computational Science group in Amsterdam: Emiliano, Gokhan, Rick, Narges, Chinh, Breandann, Alfons, and the colleagues of the Segal lab of computational biology in the Weizmann Institute: Lucas, Ohad, Eilon, Maya, Michal, Tali, Noam, Shai, Leeat.

I would like to thank Emiliano Mancini, Pierre Rezus, Ruben van Loon and Jordy Klaver for helping me during my defense and for being great friends.

Countless hours of scientific discussion were spent with Marten Postma, whom I’d like to thank for inspiring me in achieving the highest level of science.

I especially would like to thank Lucas Carey for intense collaboration at the highest level and for teaching me to be an experimental biologist. In addition I’d like to thank Gokhan Ertaylan for a wonderful and fruitful collaboration.

Finally, I’d like to thank my sister, Diana, and my parents, Mira and Hein, for their unconditional love, support, and inspiration.
List of publications


* Equal contribution