



University of Tennessee, Knoxville
**TRACE: Tennessee Research and Creative
Exchange**

Doctoral Dissertations


Graduate School

5-2011

Adaptation and Stochasticity of Natural Complex Systems

Roy David Dar
darr@utk.edu

Follow this and additional works at: https://trace.tennessee.edu/utk_graddiss

 Part of the [Biological and Chemical Physics Commons](#), [Biophysics Commons](#), and the [Systems Biology Commons](#)

Recommended Citation

Dar, Roy David, "Adaptation and Stochasticity of Natural Complex Systems. " PhD diss., University of Tennessee, 2011.
https://trace.tennessee.edu/utk_graddiss/959

This Dissertation is brought to you for free and open access by the Graduate School at TRACE: Tennessee Research and Creative Exchange. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of TRACE: Tennessee Research and Creative Exchange. For more information, please contact trace@utk.edu.

To the Graduate Council:

I am submitting herewith a dissertation written by Roy David Dar entitled "Adaptation and Stochasticity of Natural Complex Systems." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Physics.

Michael L. Simpson, Major Professor

We have read this dissertation and recommend its acceptance:

Robert N. Compton, Hanno H. Weitering, Zhenyu Zhang, John F. Cooke, Michael L. Simpson,
Chris D. Cox

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a dissertation written by Roy David Dar entitled "Adaptation and Stochasticity of Natural Complex Systems." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Physics.

Michael L. Simpson, Major Professor

We have read this dissertation
and recommend its acceptance:

Robert N. Compton

Hanno H. Weitering

Zhenyu Zhang

John F. Cooke

Chris D. Cox

Accepted for the Council:

Carolyn R. Hodges, Vice Provost and
Dean of the Graduate School

(Original signatures are on file with official student records.)

Adaptation and Stochasticity of Natural Complex Systems

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Roy David Dar

May 2011

Copyright © 2011 by Roy David Dar
All rights reserved.

DEDICATION

To my loving family,
Daniel, Hili, Gali, Dorit, Hanna, and Akiva,
whose continuous support and cheer helped me complete this work.

ACKNOWLEDGEMENTS

There is a noteworthy constellation of people and resources whose support has been unwavering throughout my doctoral process and scientific achievements. First and foremost is my doctoral committee whose insightful comments and guidance over the past 2.5 years helped shape the final hypotheses and content of this dissertation. These include Drs. Compton, Cooke, Cox, Simpson, Weitering, and Zhang. Second, and no less important, is my advisor Dr. Michael Simpson who constantly believed in me and my abilities. Mike provided endless support in developing my scientific skills and through mentorship has raised me from an entry level science intern to a highly motivated young investigator. Mike gave me perspective and ‘behind-the-scenes’ insights of a life in science including the keys to success. The Center for Nanophase Materials Sciences (CNMS) at Oak Ridge National Laboratory (ORNL) provided me with a world-class scientific work environment and financial support was from the DOE Office of Basic Energy Sciences.

Our systems and noise biology group has provided me with an exciting interdisciplinary research environment and dialog. The members over the years include Derek Austin, Mike McCollum, Mike Allen, David Karig, Chris Cox, and John Cooke all of whom were open to diverse research topics of high scientific value. Experimentally, extensive efforts and training on fluorescence microscopy and sample preparation are noted from Derek Austin, Mike Allen, John Wilgus, and David Karig. Mike McCollum and Tommy Trimeloni provided well-needed support with advanced stochastic simulation tasks.

I’d like to thank Leor Weinberger for the successful development of an exciting and fruitful collaboration. I’m grateful to have the opportunity to work with Leor and his lab members. Leor, Brandon Razooky, and Abhi Singh performed all of the T-cell sample preparation and microscopy experiments supporting this work.

There are several affiliates at the University of Tennessee, Knoxville (UTK) that supported me over the years including John Dunlap in the microscopy lab in SERF, Gary Saylor and the Center for Environmental Biotechnology (CEB), and the Physics

Department. The U.S. Department of Energy, Office of Science and the Oak Ridge Institute for Science and Education (ORISE) provided me with the Science Undergraduate Laboratory Internship (SULI) which laid the groundwork for pursuing graduate studies at the University of Tennessee, Knoxville (UTK).

My University of Tennessee and Oak Ridge National Laboratory classmates who provided me with friendship, support, and a positive graduate experience include Kate Klein, Ben Fletcher, Sarah Fletcher, Nicole Edwards, Piro Siuti, Laurene Tetard, and Amal Al-Wahish. Additional support and discussions came from the Nanofabrication Research Laboratory (NRL) at the CNMS and Mitch Doktycz's laboratory.

Finally I deeply appreciate my family and friends, whose dedicated support was an essential ingredient in making it through the roller-coaster ride of grad school and completion of this dissertation.

ABSTRACT

The methods that fueled the microscale revolution (top-down design/fabrication, combined with application of forces large enough to overpower stochasticity) constitute an approach that will not scale down to nanoscale systems. In contrast, in nanotechnology, we strive to embrace nature's quite different paradigms to create functional systems, such as self-assembly to create structures, *exploiting* stochasticity, rather than overwhelming it, in order to create deterministic, yet highly adaptable, behavior. Nature's approach, through billions of years of evolutionary development, has achieved self-assembling, self-duplicating, self-healing, adaptive systems. Compared to microprocessors, nature's approach has achieved eight orders of magnitude higher memory density and three orders of magnitude higher computing capacity while utilizing eight orders of magnitude less power. Perhaps the most complex of functions, homeostasis by a biological cell – i.e., the regulation of its internal environment to maintain stability and function – in a fluctuating and unpredictable environment, emerges from the interactions between perhaps 50M molecules of a few thousand different types. Many of these molecules (e.g. proteins, RNA) are produced in the stochastic processes of gene expression, and the resulting populations of these molecules are distributed across a range of values. So although homeostasis is maintained at the system (i.e. cell) level, there are considerable and unavoidable fluctuations at the component (protein, RNA) level. While on at least some level, we understand the variability in individual components, we have no understanding of how to integrate these fluctuating components together to achieve complex function at the system level. This thesis will explore the regulation and control of stochasticity in cells. In particular, the focus will be on (1) how genetic circuits use noise to generate more function in less space; (2) how stochastic and deterministic responses are co-regulated to enhance function at a system level; and (3) the development of high-throughput analytical techniques that enable a comprehensive view of the structure and distribution of noise on a whole organism level.

TABLE OF CONTENTS

CHAPTER 1: Introduction and Hypothesis	1
1.1 A bounded complex adaptive system	3
1.1.1 Key constraints in a bounded and finite complex system.....	5
1.2 Scope of Dissertation	9
1.3 Organization of Dissertation.....	11
CHAPTER 2: Fundamentals and Methodology	12
2.1 Gene Expression in a Nutshell.....	12
2.2 Stochasticity in Gene Expression.....	15
2.2.1 Sources of noise in gene expression	15
2.2.2 Gene expression noise structure.....	19
2.2.3 Analytical and computational methods for gene circuit noise analysis.....	20
2.3 Measuring Noise in Gene Circuits.....	22
2.3.1 The Green Fluorescent Protein (GFP)	22
2.3.1.1 Brief History	22
2.3.1.2 Basics of Fluorescence.....	23
2.3.1.3 Fluorescent Reporters for Detection of In Vivo Protein Levels	24
2.3.2 Flow Cytometry – A High-Throughput Fluorescence Screener	25
2.3.3 Correlation Spectroscopy Methods.....	28
2.3.4 Time-Lapse Single-Cell Fluorescence Microscopy.....	30
2.3.5 Image Processing of Single-Cell Gene Expression Experiments	32
2.3.5.1 Cell Segmentation.....	33
2.3.5.2 Cell Tracking	35
2.3.5.3 Fluorescent Intensity Quantification.....	36
2.3.5.4 Time-line of increase in single-cell experimental throughput	39
2.3.6 Fluorescent Intensity Signal Processing	40
2.3.6.1 Separating Stochastic Expression from Deterministic Expression.....	40
2.3.6.2 High Frequency (HF) Noise Processing	44
CHAPTER 3: The Coupling of Gene Circuit and Noise Structures	47

3.1 Gene circuit structure	47
3.2 Open loop constitutive gene circuit	48
3.2.1 Experimental investigation of constitutive gene expression.....	51
3.2.1.1 Single-cell noise frequency range distributions	54
3.2.1.2 Modulation of protein dilution and decay.....	54
3.3 Autoregulated Gene Circuits in Nature.....	56
3.3.1 Analysis of an Autoregulated Gene Circuit	58
3.3.2 Experimental investigation of negative autoregulation	60
3.3.2.1 Detection of various strengths of negative autoregulation	62
3.3.2.2 Modulation of extrinsic noise with a drug – control experiment.....	64
3.3.2.3 Effect of a non-regulated repressor – control experiment	66
3.3.2.4 Summary of negative autoregulation effects on noise	66
3.3.3 Experimental investigation of positive autoregulation	68
3.3.3.1 Correlation shifts in minimal positively autoregulated gene circuits	71
3.3.3.2 Reduction of the minimal circuit strength of positive autoregulation	74
3.3.3.3 Correlation shifts in wild type transactivated HIV-1	74
3.3.3.4 Reduction of strength of HIV-1 positive autoregulation	76
3.3.3.5 Summary of positive autoregulation effects on noise.....	78
3.4 Transcriptional Regulation.....	78
3.4.1 Two-state model of transcriptional bursting	79
CHAPTER 4: Noise Mapping.....	83
4.1 Noise Maps as a Gene Circuit Discovery Tool.....	83
4.1.1. Defining the noise regulatory vector and 3-D noise map space	84
4.1.2. Theoretical noise maps of main regulatory motifs	86
4.1.2.1 Theoretical noise map of transcriptional regulation	86
4.1.2.2 Theoretical noise map of negative autoregulation	88
4.1.3. Noise vector domains for various regulatory motifs.....	91
4.2 The experimental reality of noise maps	92
4.3 Noise maps to study HIV latency	94

4.3.1 Experimental investigation of genome-wide transcriptional bursting in humans	95
4.3.1.1 Methods	95
4.3.1.2 Creating an experimental noise map	101
4.3.1.3 Results	106
4.4 Deterministic implications of the two-state model of transcription	119
CHAPTER 5: The Coupling of Stochastic and Plastic Response	120
5.1 Transcriptional two-state model describes coupling between stochasticity and plasticity	120
5.2 Distribution and regulation of stochasticity and plasticity in <i>Saccharomyces cerevisiae</i>	123
5.2.1 System-wide noise	123
5.2.2 System-wide plasticity	123
5.2.3 Regulatory arrangements that control noise and plasticity	127
5.2.4 Results	129
5.2.4.1 Noise-Plasticity coupling is widespread across the genome	129
5.2.4.2 Noise-Plasticity coupling strength is dominated by regulatory arrangement	129
5.2.5 Discussion	134
5.3 Distribution and regulation of stochasticity and plasticity in <i>E. coli</i>	137
5.3.1 Introduction	137
5.3.2 System-wide noise	137
5.3.3 System-wide plasticity	140
5.3.4 Regulatory arrangements that control noise and plasticity	142
5.3.5 Results	142
5.3.6 Discussion	145
5.4 A Novel Unicellular Noise-Plasticity Scaling Law	146
5.5 Noise-plasticity coupling is wide-spread but not all genes are coupled	146
5.6 Noise-plasticity coupling of yeast and <i>E. coli</i> regulators	150

CHAPTER 6: Summary and Conclusions.....	154
LIST OF REFERENCES.....	159
APPENDIX.....	170
7.1 Fundamentals and Methodology.....	170
7.1.1 Biased versus Unbiased Autocorrelation.....	170
7.1.2 Automated tracking of an adhered slow-growing cell monolayer.....	171
7.1.3 Cellular Fluorescent Intensity and Fluorescent Protein Abundance are Correlated.....	174
7.1.4 Stochastic Simulation and Gillespie’s Algorithm [52].....	176
7.1.4.1 Basic concept.....	176
7.1.4.2 Considerations when simulating.....	177
7.1.4.3 Biospreadsheet: A User-Friendly Simulator.....	177
7.1.4.4 An Example of Stochastic Simulation.....	181
7.1.5 Manual Quality Control of Acquired Fluorescence Signals.....	183
7.2 The Coupling of Gene Circuit and Noise Structures.....	184
7.2.1 Half-Correlation Time Error Bar Estimation.....	184
7.3 Noise Mapping.....	186
7.3.1 Advantages and Disadvantages of Polyclonal Noise Mapping.....	186
7.3.2 Cell-cycle synchronization.....	187
7.3.3 Example of noise maps for 6 LTR-d2GFP Isoclonal experiments.....	190
7.3.4 Longer cell recovery has little effect on noise map signature.....	194
7.3.5 Noise map scatter dependence on experiment duration.....	196
7.3.6 Experimental Methods Summary.....	199
7.3.7 Resampling algorithm for converting polyclonal noise maps to O-k probability landscapes.....	200
7.3.8 Determination of the basal transcription level.....	212
7.3.9 Multiple General Trends for Polyclonal or 2-Reporter Experiments.....	213
7.3.10 Are noise map shifts due to extrinsic noise?.....	215
7.4 The Coupling of Stochastic and Plastic Response.....	216

7.4.1 Derivation of the relationship between excess noise and PL.....	216
7.4.2 Additional Representation of Noise-Plasticity Coupling in Yeast	219
7.4.3 Additional Representation of Noise-Plasticity Coupling in <i>E.coli</i>	220
VITA.....	221
PUBLICATIONS	222

LIST OF TABLES

Table 2.1 Summary of Correlation Spectroscopy Methods	29
Table 2.2 Summary of the noise processing algorithm	42
Table 3.1 Stochastic simulation model of ATc-ribosome inhibition	64
Table 3.2 Summary of Experimental Composite $\tau_{1/2}$ and Strength of Regulation (T).....	76
Table 4.1 Characterization of Transcriptional Burst Landscapes	108
Table 7.1 Simulation of a basic gene expression model	176
Table 7.2 2-pole stochastic simulation model.....	184
Table 7.3 Parameters for the 2-state model simulation library	210

LIST OF FIGURES

Figure 1.1 A complex adaptive information processing system.....	4
Figure 1.2 Distribution of limited system resources	6
Figure 1.3 A population uncertainty principle	8
Figure 1.4 Plastic versus Stochastic Response.....	10
Figure 2.1 Gene Expression in a Nutshell.....	14
Figure 2.2 Noise in molecular populations	17
Figure 2.3 Intrinsic and extrinsic noise sources in gene expression	18
Figure 2.4 Noise autocorrelation yields both variability and frequency content.....	19
Figure 2.5 Green fluorescent protein	23
Figure 2.6 Stimulated emission of a photon	23
Figure 2.7 Protein-GFP fusion library covering 4159 budding yeast genes	24
Figure 2.8 Flow cytometry – a high throughput fluorescence screener.....	26
Figure 2.9 Raw FC fluorescence distributions of 88 budding yeast protein-GFP fusion populations.....	27
Figure 2.10 Correlation spectroscopy methods.....	29
Figure 2.11 Schematic of confocal microscopy operation.....	30
Figure 2.12 Automated Olympus Spinning Disc Confocal Microscope	32
Figure 2.13 LOG image segmentation of bacterial cells	34
Figure 2.14 A fluorescent T-cell image and its segmented binary array	34
Figure 2.15 Schematic of a single lineage or trajectory in a growing <i>E.</i> <i>coli</i> colony.....	35
Figure 2.16 Equivalent autocorrelations for voxel and whole cell sampling.....	37
Figure 2.17 Limited sampling introduces white noise.....	38
Figure 2.18 Timeline of experimental methodology and single-cell throughput increase.....	39
Figure 2.19 Gene expression HF noise processing	43

Figure 2.20 Baseline expression shifts are indistinguishable from low frequency fluctuations.....	45
Figure 2.21 High frequency noise processing.....	45
Figure 2.22 HF-processing focuses on intrinsic and filters out extrinsic noise	46
Figure 3.1 Time-domain constants of constitutive gene expression.....	50
Figure 3.2 Diagram of <i>E. Coli</i> Experiment Setup.....	52
Figure 3.3 Noise frequency range detection with fluorescence microscopy	53
Figure 3.4 Effects of cell doubling time and protein half-life on noise frequency range.....	55
Figure 3.5 Positive autoregulation in well-known viruses.....	57
Figure 3.6 Overlaid positive and negative feedback loops in animal viruses	57
Figure 3.7 Negative autoregulation increases noise bandwidth.....	59
Figure 3.8 Effect of negative autoregulation on noise frequency range	61
Figure 3.9 Regulation strength modulation of noise frequency range.....	63
Figure 3.10 A model of ATc inhibition of translation	65
Figure 3.11 Non-regulated repressor control frequency range remains log-normal	67
Figure 3.12 Positive-feedback extends the lifetime of gene expression transients	70
Figure 3.13 Sample setup and fluorescent image of GFP expressing human T-cells.....	72
Figure 3.14 Measuring positive-feedback strength by exploiting inherent gene expression noise	73
Figure 3.15 Positive-feedback strength drives an extended Tat expression transient in both minimal Tat circuits and full-length HIV-1	75
Figure 3.16 SirT1 over-expression, in full-length HIV-1 decreases positive-feedback strength and increases the probability of latency	77

Figure 3.17 Transcriptional regulation and bursting.....	80
Figure 3.18 The 2-state transcription model	80
Figure 3.19 Diagram of operator state and 2-state transcription model.....	81
Figure 4.1 Noise mapping as a gene circuit discovery tool	84
Figure 4.2 The noise regulatory vector and its relationship to the 3D noise map.....	86
Figure 4.3 Noise regulatory vectors for slow gene activation kinetics	87
Figure 4.4 Negative autoregulation and autoregulator-DNA binding	89
Figure 4.5 Noise regulatory vectors for negative autoregulation.....	90
Figure 4.6 Summary of noise vector domains for various regulatory motifs.....	91
Figure 4.7 Convergence of noise map spread with increasing experimental duration	93
Figure 4.8 Scheme for probing transcriptional bursting across the genome.....	96
Figure 4.9 Bursty gene expression dominates across the human genome	99
Figure 4.10 Creating a Noise Map	102
Figure 4.11 HF-CV ² vs. average fluorescence level for LTR d2GFP monoclonal	104
Figure 4.12 Distributions of HF- $\tau_{1/2}$ s measured for LTR d2GFP monoclonal	104
Figure 4.13 Combined noise map for monoclonal C32 and D36.....	105
Figure 4.14 Equivalent polyclonal correlation time distributions.....	107
Figure 4.15 Equivalent normalized composite autocorrelations for the	107
Figure 4.16 A model of two-state transcriptional bursting	108
Figure 4.17 Modulations of genomic transcriptional bursting landscape by integration site, promoter type, and signaling molecules	109
Figure 4.18 LTR promoter transcriptional start is delayed by stalling of RNA polymerase.....	112
Figure 4.19 A robust genome-wide frequency-band of transcription	114
Figure 4.20 Modulations of burst transition rates with TNF α	116

Figure 4.21 Detailed representation of the LTR promoter.....	118
Figure 4.22 EF-1A expression slightly decreases with TNF α addition.....	118
Figure 5.1 The yeast Environmental Stress Response (ESR).	126
Figure 5.2 Distinct regulatory features of stress and growth related genes.	128
Figure 5.3 Primary nucleosome occupancy patterns	128
Figure 5.4 Widespread genome-wide noise-plasticity coupling in Yeast.....	130
Figure 5.5 Excess noise and plasticity are related and strongly dependent on gene regulatory architecture.....	132
Figure 5.6 Excess noise and plasticity are positively correlated.....	133
Figure 5.7 Expected inverse relationship between variability and deterministic expression.....	135
Figure 5.8 Noise in 1000 <i>E.coli</i> protein-YFP fusion strains	138
Figure 5.9 System-wide noise measurements in <i>E.coli</i> and Yeast.....	139
Figure 5.10 Comparison of excess noise in <i>E. coli</i> and Yeast	139
Figure 5.11 A constructed <i>E.coli</i> stress microarray compendium	141
Figure 5.12 Comparison of plasticity range between <i>E.coli</i> and yeast	141
Figure 5.13 Widespread genome-wide noise-plasticity coupling in <i>E. coli</i>	143
Figure 5.14 Noise-plasticity coupling among sigma-factor regulators.....	144
Figure 5.15 A Novel Unicellular Noise-Plasticity Scaling Law	147
Figure 5.16 The two-state model can couple and uncouple noise and plasticity.....	149
Figure 5.17 Noise-plasticity coupling of yeast regulators.	151
Figure 5.18 Yeast regulators are important stochastic exploiters	152
Figure 5.19 Non –AR <i>E. coli</i> TFs are noise-plasticity coupled	153
Figure 6.1 Problems with resource rich driven synthetic design	156
Figure 7.1 “Segmentator” program for segmentation and tracking of adhered T-cells.....	172
Figure 7.2 Clustering of oversampled cell seeds	173

Figure 7.3 Fluorescence intensity and fluorescent protein abundance are correlated.....	175
Figure 7.4 Information tab of BioSpreadsheet Simulator	178
Figure 7.5 Species tab of BioSpreadsheet Simulator	179
Figure 7.6 Reactions tab of BioSpreadsheet Simulator	179
Figure 7.7 Simulation settings tab in BioSpreadsheet	180
Figure 7.8 Stochastic simulation of single cell gene expression.....	182
Figure 7.9 Simulated error in 12 hour HF-T50.....	185
Figure 7.10 NPD and difference map for synchronized and unsynchronized cells	188
Figure 7.11 HF-T50 distribution comparison for synchronized and unsynchronized cells.....	189
Figure 7.12 Correlation time is independent of cell cycle state.....	189
Figure 7.13 Noise magnitude and correlation for Ld2G monoclonal.....	191
Figure 7.14 Individual noise map signatures for 6 Ld2G monoclonal	192
Figure 7.15 Monoclonal and polyclonal HF-T50 distributions are similar.....	193
Figure 7.16 Plateau of Ld2G monoclonal general intensity trends.....	194
Figure 7.17 Longer sample recovery does not change the experimental results	195
Figure 7.18 Noise map limited duration correlation cutoff.	197
Figure 7.19 EF-1 α d2G and LTR d2G polyclonal fluorescence intensity distributions.....	198
Figure 7.20 Examples of two-state noise map simulations.....	201
Figure 7.21 Probability point spread functions of known O-k simulations	205
Figure 7.22 Resolution in the burst probability landscape.....	207
Figure 7.23 Best simulation match to experiments.....	208
Figure 7.24 Samples of simulated NPD maps	209
Figure 7.25 Composite noise map for an array of O and k values.....	212

Figure 7.26 Multiple deterministic trends for analyzing polyclonal experiments	214
Figure 7.27 Simulated HF extrinsic noise cannot explain reported HF-NPD map shifts	215
Figure 7.28 Response dependent noise-plasticity coupling in yeast.....	219
Figure 7.29 Response dependent noise-plasticity coupling in <i>E. coli</i>	220

LIST OF SYMBOLS AND VARIABLES

$\alpha, k_m, \text{ or } k_r$	transcription rate
k_p	translation rate
γ_m	mRNA half-life
γ_p	protein half-life
b	translational burst rate
k	transcriptional burst kinetic rate
O	average “on time” or “on fraction”
f_B	burst frequency

LIST OF ABBREVIATIONS

ACF	autocorrelation function
CAC	composite ACF
NCAC	normalized composite ACF
T50 or $\tau_{1/2}$	half-correlation time
CV	coefficient of variation
HF	high frequency
DNA	deoxyribonucleic acid
RNA	ribonucleic acid
mRNA	messenger RNA
P	protein
RNAP	RNA polymerase
GFP	green fluorescent protein
T	loop transmission or strength of feedback
voxel	volumetric pixel
Tet	tetracycline
ATc	anhydrotetracycline
Tat	Trans-Activator of Transcription
TNF α	Tumor Necrosis Factor-alpha
HIV-1	human immunodeficiency virus type 1

CHAPTER 1: Introduction and Hypothesis

Perhaps the *Holy Grail* of nanoscience is discovering the “rules of composition” that could provide the ability to mimic, manipulate, and engineer both natural and synthetic devices exhibiting the advanced functionality, efficiency, and robustness which already exist in natural complex nano-scale systems (i.e. living organisms). Currently there is no general theory to guide the organization of complex networks of interacting elements into highly functional systems. As a result, much scientific activity concentrates on two distant scientific realms of a complexity continuum. At the high end of this continuum, top-down observation of the organization of natural complex networks of nanoscale elements provides some clues about the nature of the rules of composition. At the low end of the continuum, work focuses on trying to construct synthetic systems that mimic some limited portion of the function of the natural complex systems. The ultimate goal is to connect these two approaches such that the modeling of natural and synthetic genetic networks, observation of network organizational principles, and the discovery of novel structure-function relationships are funneled down to the bottom-up synthesis of complex nanomaterials and integration of advanced synthetic devices. This reverse engineering of biological complexity has drawn much attention from the scientific community[1]. Nano-biotechnology serves as an intermediate between the top-down and bottom-up fields in which developed nanomaterials and novel tools are used to interface and characterize biology on the small scale. It is only through matching the functional density and scale of natural systems that one can begin to aspire and mimic its complexity. These sorts of nano-enabled synthetic and systems biology efforts may provide the very first bridge between these two distant worlds [2].

Just how much more efficient is Nature from man-made design? We can gain some insight by contrasting a modern microprocessor with a bacterium. *E. coli* has a cross-sectional area of $\sim 2\mu\text{m}^2$, 9.2 megabit memory (based on DNA base pairs) and the equivalent of $\sim 1,000$ logic gates (i.e., $\sim 5 \text{ Mbit}/\mu\text{m}^2$ and $\sim 500 \text{ logic gates}/\mu\text{m}^2$)[3]; it

solves complex information extraction problems (e.g. chemotaxis) on a time scale of minutes with power consumption of 10^{-15}W , or a power density of $5 \times 10^{-16}\text{W}/\mu\text{m}^2$. A state-of-the-art Intel chip (e.g., i5-600) has a cross-section of $\sim 1000\text{mm}^2$ (or $10^9\mu\text{m}^2$), contains 4MB of memory and ~ 500 million logic gates (or $\sim 3 \times 10^{-8}\text{Mbit}/\mu\text{m}^2$ and $0.5\text{logic gates}/\mu\text{m}^2$) and has a power consumption of $\sim 100\text{W}$ (or $\sim 10^{-7}\text{W}/\mu\text{m}^2$). Thus, through billions of years of evolutionary development, nature has developed a self-assembling, self-duplicating, self-healing, adaptive processing unit that has 8 orders of magnitude higher memory density, 3 orders of magnitude higher computing capacity while utilizing 8 orders of magnitude less power.

Understanding the “function” of living organisms presents a complex, multi-layered problem to which various disciplines take diverse approaches. On the genomics level, advancement in sequencing technologies along with efficient algorithms have established a computational thrust towards identifying and characterizing genes and their evolution. On the molecular level, much research concentrates on characterizing protein constituents, surface residues, protein-protein interactions, and the forces involved. Questions of charge, folding structure and states, signaling pathways, and multiple component machines are central and lie within a detailed identification and characterization of a protein’s function or pathway. New frontiers of biological physics and systems biology take an alternate approach by modeling gene circuit structure and function within a gene network framework. Transcription, translation, gene activation and repression, protein-protein interactions, and multiple-component molecular machines are all accounted for as the final objectives focus on deducing mathematical models that describe the circuit dynamics within the network.

Understanding the organizational principles of cells has spanned many scales of the problem. Initial efforts were aimed at quantifying the topology of genetic networks along with their implied system characteristics [4-6]. Later, re-occurring sub-components of genetic networks (coined “network motifs”) were identified and characterized for their function and dynamics [7, 8]. Finally much recent focus has been on the implications of stochasticity in gene circuits and networks and it is these implications which will be

studied in detail in this dissertation.

1.1 A bounded complex adaptive system

For the purposes of this work, a system is considered *complex* if it is made up of many highly interconnected components which promote a large and diverse range of adaptive functions (Fig. 1.1A). The adaptive system is continually affected by external forces such as fluctuations in temperature, pH, chemical environment, and physical stimuli. These forces dictate different distributions in system activities such as movement, energy processing and storage, information processing and storage, and more. There is a limit to the system's ability to perform many activities in parallel due to a limitation in resource. So the system needs to balance its various functions. In Figure 1.1B, the adaptive information processing system traverses in a state (phenotype) landscape. To do this, the system receives inputs in the form of fluctuating environmental conditions, has an intrinsic information processing framework which involves the highly interconnected system components, and finally produces a composite output solution based on many component processes working in parallel. E.g. in Fig. 1.1B, the adaptive system changes from "State A" to "State B" while conserving homeostasis – the property whereby a system regulates its internal environment to maintain a stable or constant condition. Here the system is in a quasi steady-state all the way from "State A" to "State B". "State B" can be an unfavorable state or simply an alternative healthy state.

This simplified adaptive system is bounded with a highly interconnected internal composition of many components capable of producing its many functions. The system has high configurability, an internal design, and memory of previous states to enable it to provide an accurate, highly diverse, and advanced set of functions or actions. Looking deeper into the system constraints yields an important conservation law and analytical relationship for resource distribution.

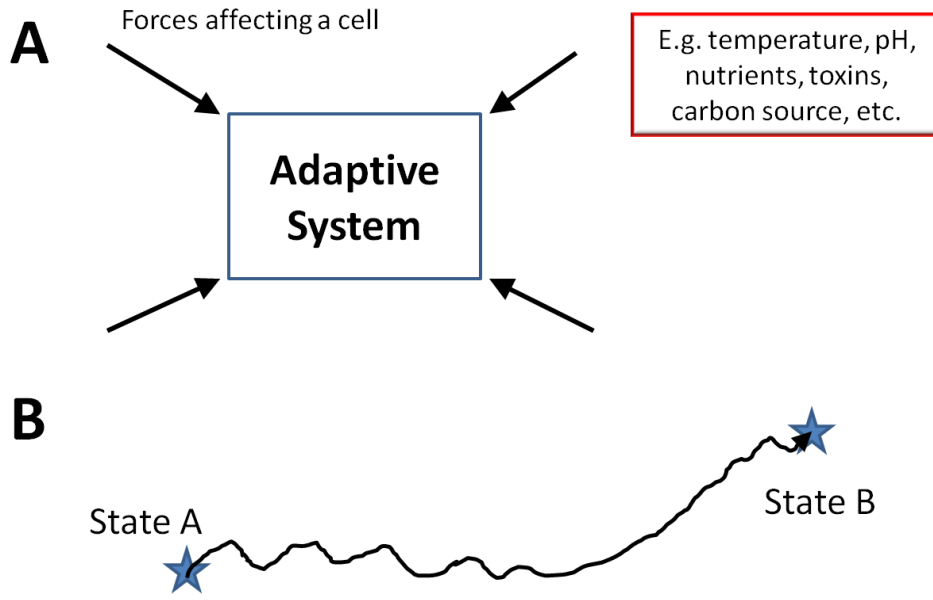


Figure 1.1 A complex adaptive information processing system. **A.** A physicist's simplified view of the complex system, with all of its composition and complexity, as a box. The box, or adaptive system, is continually under the effect of various forces. **B.** The system, influenced by many environmental inputs conserves homeostasis or a quasi system steady-state while traversing states from "A" to "B". The path traversed is dependent upon the adaptive information processing infrastructure and the final state ("State B") is the output of the system, a composite phenotypic state.

1.1.1 Key constraints in a bounded and finite complex system

The dynamic and complex adaptive system depends heavily on a resource pool of energy, space, and elements that drive the production and composition of the system constituents and infrastructure. An important consequence of the system being finite and bounded is that the pool of resources is limited. This results in a ***limitation and distribution of total resources*** and sharing results in a distribution in the populations (or concentrations) of various protein species and various components of the system (Fig. 1.2). The population distributions often result in small protein populations where noise or ***stochasticity*** becomes dominant compared to the population mean. Owing to the random timing and discrete nature of biochemical interactions (single molecules at a time) for production or decay of a particular protein, these lower abundance sub-populations are intrinsically noisy. This intrinsic and unavoidable noise source is a natural byproduct of the protein production and decay processes and is termed the ‘shot-noise’ of the system [9, 10].

The shot noise can be described by the square of its coefficient of variation,

$$CV_{shot}^2 = \frac{\sigma_P^2}{\langle P \rangle^2} \quad (1.1)$$

which is equal to the protein population variance over its mean squared where the coefficients of variation (CV; standard deviation/average) for typical protein populations in the system would range from 1-100% (i.e. from negligible to dominant). This shot noise term is a low-noise limit as it is the basal noise produced by the system. Additional stochasticity on top of the shot noise is possible and will be explained in more detail later.

This excess noise can be shown to have the form [10]:

$$\Delta CV^2 = CV^2 - CV_{shot}^2 \propto \frac{(1-O)}{O}, \quad O \in [0,1] \quad (1.2)$$

where O is a quantity related to the amount of resources allocated to produce a gene

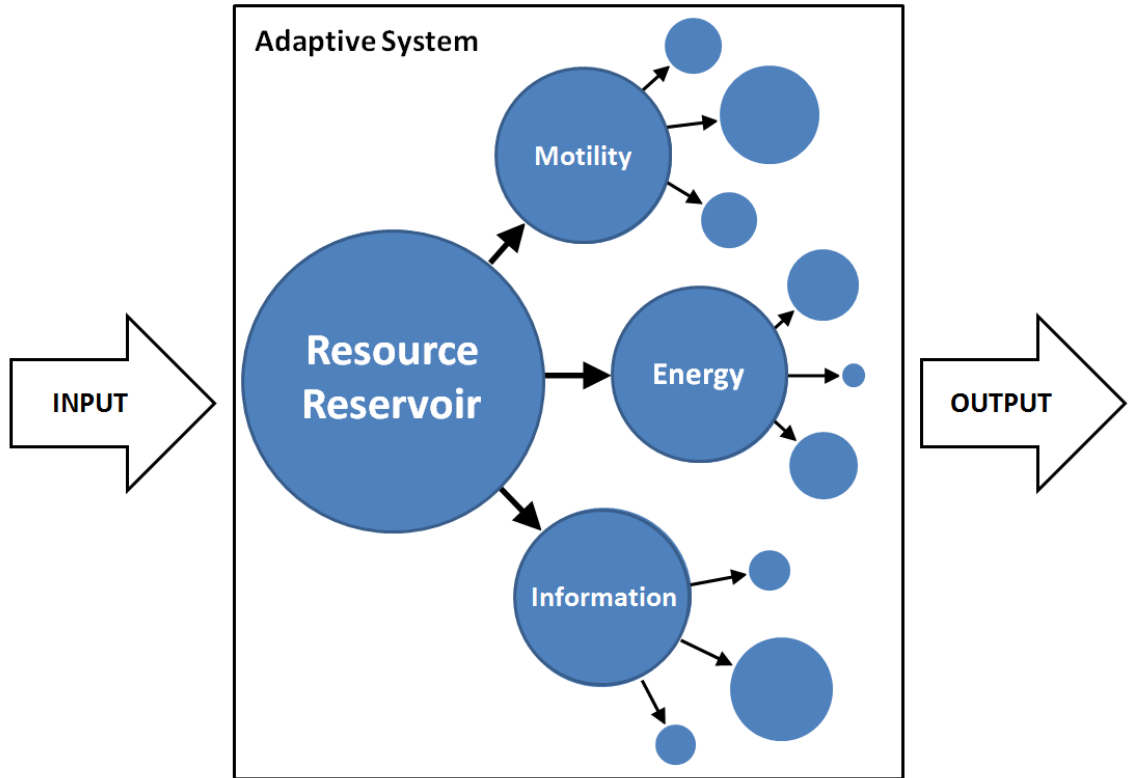


Figure 1.2 Distribution of limited system resources. The schematic depicts a simplification of the limited resource distribution problem. After an input enters the adaptive system, the resource processing and sharing among system sub-components (here generalized to three main system functions) results in a composite system output. The resource allocation to each system subcomponent is dynamic and depends on the system input.

product (or protein) and can have values between 0 and 1. O equal to 1 means that the production of protein receives all of the resources it is capable of using. A consequence of the limited resources is the conservation of total O integrated over all the genes in the biological system,

$$\sum_{genes} O_i = Const. \quad (1.3)$$

which results in a **conservation and distribution of total excess noise**:

$$\sum_{genes} \Delta CV_i^2 = \sum_{genes} (CV_i^2 - CV_{shot_i}^2) \propto \sum_{genes} \frac{(1 - O_i)}{O_i} \approx Const. \quad (1.4)$$

So as a consequence of the first limited resource constraint, there exists a ***conservation and distribution of (excess) stochasticity***, and little is known about how this excess stochasticity should be distributed across different functions or classes of function in complex systems. In addition, there is the possibility that natural selection has resulted in non-Poisson processes or mechanisms in certain protein populations where CVs are either much over or much under the shot noise. This would indicate protein production events that were not independent of one another, and a consequence of certain production architectures/mechanisms. Overall the resource limitation constraints leads to two distinct and coupled laws -- limitation and active distribution of total resources (and production capacity) which results in a conservation and distribution of total system excess stochasticity.

The stochastic re-distribution described above can be thought of as a higher dimensional molecular species population analog to the Heisenberg Uncertainty Principle (Fig. 1.3). There is a limit in how much is known about the actual population level of a certain protein species (e.g. P_1), this uncertainty is due to variability and stochasticity in its production, and the amount of uncertainty allocated to this protein affects the uncertainty allocation and population level knowledge of a different protein (e.g. P_2 , P_3 , etc.).

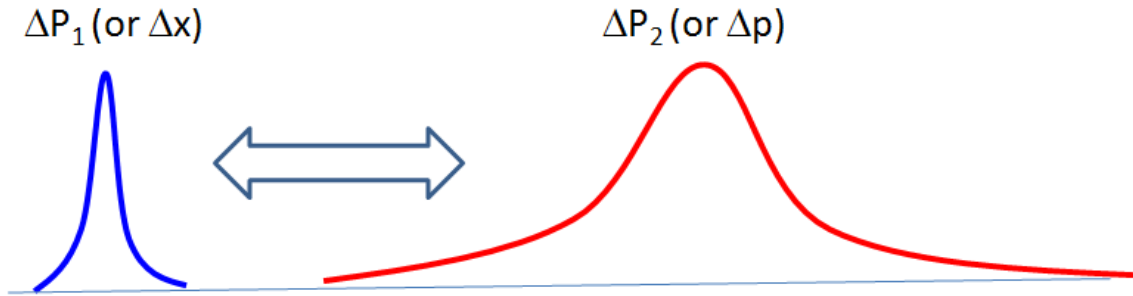


Figure 1.3 A population uncertainty principle. Similar to Heisenberg's Uncertainty Principle where the precision in measuring the momentum (p) and the location (x) of a particle are coupled one to another and limited by Planck's constant, similarly there is an uncertainty in knowing the precise population level for any two information carrier species in the system. Their variability is coupled as a consequence of resource limitation. In the figure only two information carrier populations are shown when in fact the variability distribution and coupling may occur across thousands of information carrier species and depends on the bound system volume and total resources available.

1.2 Scope of Dissertation

The work in this thesis will mostly address two main hypotheses. The first is that **stochastic and plastic responses are actively co-regulated and controlled to achieve functional objectives**. Plasticity is defined as the ability of a complex system, cells in this work, to change its state in response to changes in the environment. A plastic response may be thought of as pre-programmed response with deterministic and optimized output levels proportional to the strength of a perturbation (Fig. 1.4). Conversely, a stochastic response may occur independent of the strength of the external stimulus (Fig. 1.4). It is ‘un-programmed’ and therefore not an optimized response to a stimulus, but it may be exploited to create contrarian responses that hedge against sudden changes in the environment. As a consequence of the conservation of stochasticity imposed at the nanoscale, every component in the system responds plastically and stochastically to some degree (Fig. 1.4) but the relationship between the deterministic and stochastic response components has not been thoroughly explored. It is a central goal of this thesis to explore this relationship in some detail.

The second hypothesis of this thesis is that **stochasticity can be used as a functional component in complex nanoscale systems, and thereby generate more function in less space**. Along this line of inquiry, this thesis will focus on viral gene circuits, and in particular retroviruses are of interest as they perform complex tasks with a very limited set of components – i.e. these are ideal model systems for understanding how fluctuations may be used to get more function in less space. This work will consider the HIV-1 circuit, which is a genetic decision circuit subject to the contrarian effects enabled by noise[[11-13](#)], that mediates the decision between active infection and latency, and this circuit is known to have high noise[[14](#)] that is further enhanced by positive feedback[[15](#)].

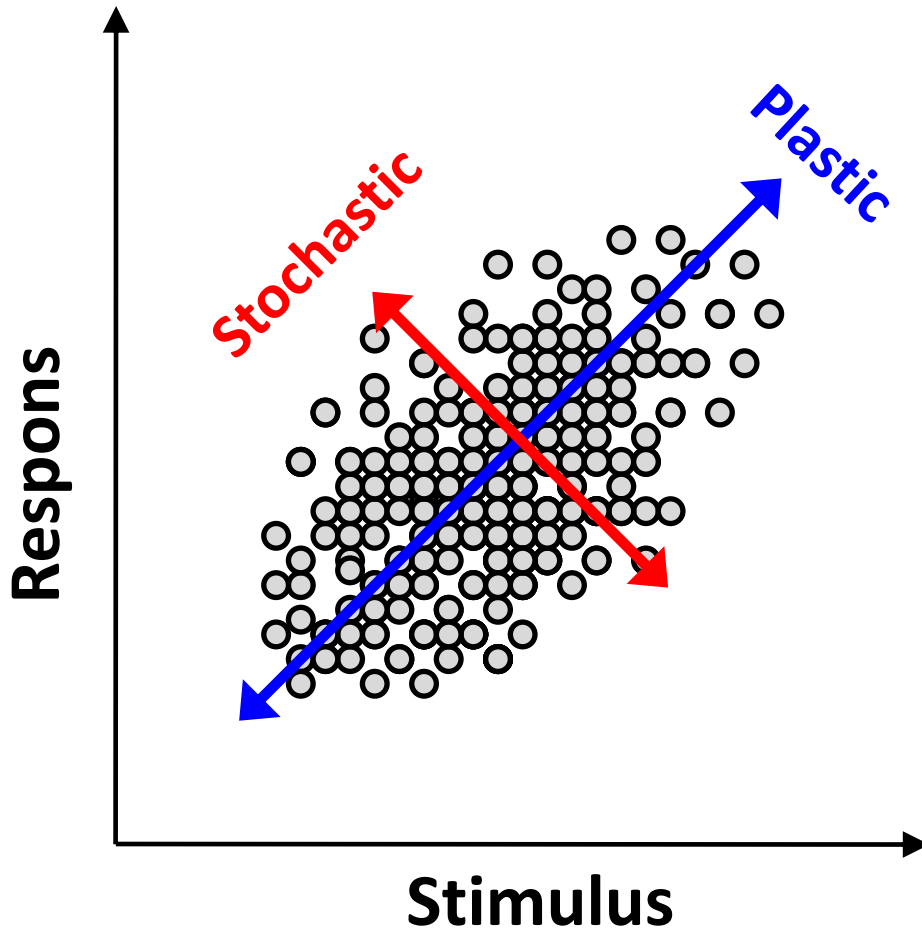


Figure 1.4 Plastic versus Stochastic Response. In response to an external stimulus an information carrier species can have a range of responses with varied strengths of plasticity or stochasticity depending on how correlated (or anti-correlated) the output level of production response is to the input level of external stimulus received.

1.3 Organization of Dissertation

Chapter 2 provides a basic biological background sufficient to understand the subsequent chapters. In addition, this chapter describes how stochastic fluctuations are analyzed, modeled, simulated, and measured. The focus of chapter 3 is to show the coupling between gene circuit and noise structures. This chapter will show the mechanisms that gene circuits can employ to regulate noise, and will show experimental results measured for three important gene circuit motifs: constitutive gene expression; negative autoregulation; and positive autoregulation. This chapter begins the consideration of the functional use of noise in gene circuits with a close look at the regulation of noise in the HIV-1 circuit. The final section of this chapter describes transcriptional bursting, an apparently ubiquitous gene expression motif that may be the central player in establishing a coupling between plastic and stochastic responses. Chapter 4 presents a novel noise analysis technique – noise mapping – that is used to experimentally explore transcriptional bursting, and in particular, transcriptional bursting that generates the noise used in the HIV-1 circuit. Building on the results of chapters 2-4, chapter 5 explores the relationship between stochasticity and plasticity in two model organisms: *Saccharomyces cerevisiae* (budding yeast) and *Escherichia coli* (bacteria). Finally, chapter 6 discusses the implications of the findings in this thesis and provides some thoughts on future work.

CHAPTER 2: Fundamentals and Methodology

The main findings described in this dissertation are all driven by theoretical hypotheses, biological phenomena, and direct experimental investigation. This chapter provides a tutorial of the biology fundamentals and experimental methods needed to understand the research presented here. The tutorial includes a description of gene expression, stochasticity in gene expression, and the green fluorescent protein (GFP) and its applications. After the basic concepts and terminology have been established two primary methods, flow cytometry and time-lapse fluorescence microscopy, will be presented concluding with detailed fluorescence microscopy image and signal processing protocols that are used in the following chapters.

2.1 Gene Expression in a Nutshell

The following section includes an overview of the most basic concepts of gene expression to provide the necessary biological background for the non-specialist. For additional reading there are several excellent books that address these issues in more detail and are accessible to a general audience [[16-18](#)].

There are three major molecules that are essential for all known forms of life: **deoxyribonucleic acid (DNA)**; **ribonucleic acid (RNA)**; and **proteins**. The instructions needed to produce all of the machinery and structures of the cell are encoded in **DNA**. DNA has a double-helix sugar and phosphate group backbone within which there are arrangements of four bases (nucleotides) called adenine (**A**), cytosine (**C**), guanine (**G**), and thymine (**T**). These four bases are pair-wise complementary such that A binds only with T and C binds only with G (Fig. 2.1 left). A gene is a segment of DNA which holds the information needed to produce a molecule of **mRNA**. The mRNA is a single-stranded nucleic acid chain with structure and chemical composition similar to DNA except that the nucleotide base thymine (T) is replaced by uracil (U). In many cases the genetic message encoded in mRNA has the instructions to produce a functional **protein**. So in general, in all organisms and all types of cells, the basic dogma of gene expression is **DNA → mRNA → protein**.

In general, expression of a gene occurs in two main steps consisting of **transcription** of the genetic sequence encoded in DNA into a single-stranded **mRNA**, followed by **translation** of mRNA into a three-dimensional protein (Fig. 2.1 right). The multi-component protein machine (**enzyme**) that transcribes DNA into mRNA is called **RNA polymerase (RNAP)**. To transcribe a gene, RNAP needs to unwrap the DNA double helix to access and duplicate a single strand of mRNA that is complementary to the DNA sequence of the gene. As the RNAP progresses along the length of the DNA it adds additional nucleotides (bases) to a growing mRNA chain.

For prokaryotes (cells lacking a nucleus) once sufficient mRNA has been transcribed, translation of the protein may commence in parallel, before completion of the whole mRNA. In eukaryotes (cells with a nucleus) transcription and translation are uncoupled and occur in different sub-compartments of the cell. During translation a different multi-component machine called the **ribosome** binds and translates mRNA into a three-dimensional **protein** complex which is made up of different combinations of 21 **amino acids (aa)**. Typical proteins are made up of a few hundred aa, but there are also some that are much smaller or much larger.

During transcription the RNA polymerase identifies its DNA binding and start site by a gene sequence called the **promoter**. Often found within the promoter sequence are regions called **operators** where **regulatory proteins** bind and can either activate (increase) or repress (decrease) the rate of gene expression. Information is stored and processed in the cell by DNA, mRNA, proteins, transcription, translation, and the cells **regulatory system**. It is the collection of diverse gene expression programs from its **gene network**, chemical signaling, and various modes of regulation that provide the cell with a broad functional (phenotypic) range. Figure 2.1, right, is a simplified schematic depiction of gene expression and summarizes the basic concepts described above.

Some typical gene expression timescales for the Bacterial *E. coli* cell, the Single-Celled Eukaryote *Saccharomyces cerevisiae* (Yeast), and a mammalian Cell (human fibroblast) are (taken from [19]):

1. Time to transcribe a gene = ~1 min (80 bp/sec in E.coli/yeast),

bp = base-pair

2. Time to translate a protein = ~2 min (40aa/sec in *E.coli*/yeast), aa = amino acids
3. Typical mRNA lifetime = ~2-10 minutes
4. Cell generation (doubling) time = 30 min - few hours (*E.coli*), 2 - 6 hours (yeast), 20 hrs or more (mammalian)
5. Equilibrium binding of small molecule to a protein = ~1msec – 1 sec
6. Mutation rate = $\sim 10^{-9}$ bp/generation (*E. coli*), 10^{-10} bp/gen (yeast), 10^{-8} bp/year (human)

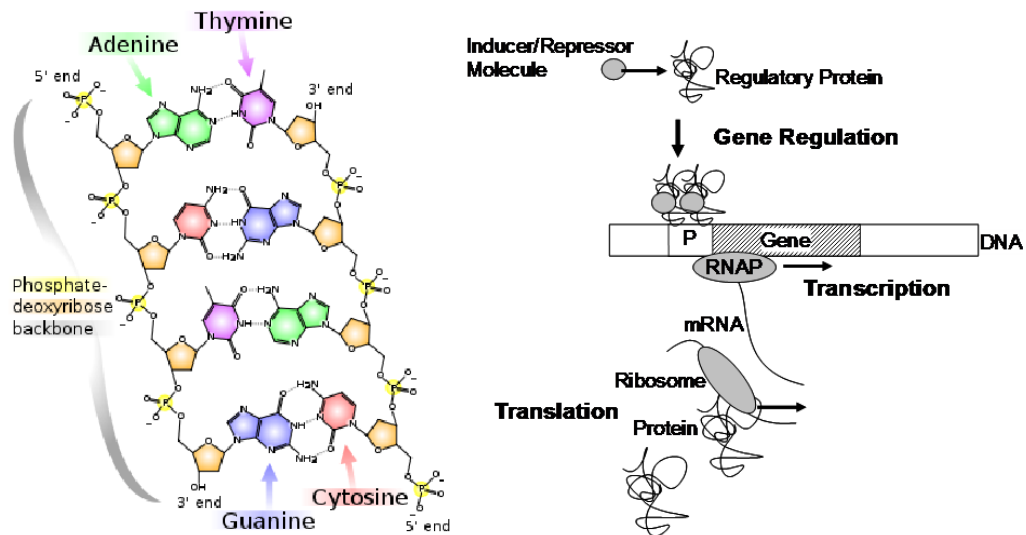


Figure 2.1 Gene Expression in a Nutshell. (Left) Double stranded DNA molecule with sugar and phosphate group backbone and A-T, C-G, nucleotide base pairing. (Right) Schematic representation of the basic elements of gene circuit function including transcription, translation, and gene regulation. [Left figure adapted from Wikipedia (M. P. Ball) and right figure from Simpson et al, (2004) [20]]

2.2 Stochasticity in Gene Expression

Initial studies of stochasticity in biological systems were reported by the physicist Max Delbruck back in 1940 who made the connection between small enzymatic populations, fluctuations in their biochemical reactions, and potential significant impacts on cell physiology [21]. He also studied variability in the number of produced viruses in lysis (cell death) of an infected bacterium [22]. From Delbruck until present day the study of stochasticity in gene expression, or in short “noise biology” has come a long way. Initial studies concentrated on the implications of fluctuations in gene regulation. Prokaryotic examples include regulation of *lac* expression at low levels of induction[23], the lysis-lysogeny decision in phage- λ [24, 25], and the swimming and tumbling periods of bacteria during chemotaxis[26]. Later studies concentrated on identifying, quantifying, and modeling the primary sources of noise in gene expression [27-29]. Most recently noise biology has concentrated on biological systems with stochastically-driven phenotype variability [30, 31] or the measurement and analysis of large-scale genome-wide stochasticity in *E.coli* [32], yeast in healthy[33] and stressful [34] conditions, and human cancer cells in response to a drug[35]. Finally, a single-cell view of clonal populations reveals that stochasticity may be used as a bet-hedging strategy. This ensures that a few cells remain poised to exploit changing environmental conditions [36, 37] and results in an improvement of cellular fitness [38]. For additional reading and the review of important developments in this fast-pace and high visibility branch of research see: Kaern et al, (2005) [39], Longo and Hasty, (2006) [40], Kaufmann and van Oudenaarden, (2007) [41], Shahrezaei and Swain, (2008)[42], Larson et al, (2009) [43], and Simpson et al, (2009) [44].

2.2.1 Sources of noise in gene expression

To understand the sources of noise in genetic circuits and networks constitutive gene expression as a simple birth-death process is examined (Fig. 2.2a). The time evolution of the population of the produced molecule, $P(t)$ may be modeled using (Fig. 2.2b):

$$\begin{aligned}\frac{dP(t)}{dt} &= \alpha - \gamma P(t) \\ P(t) &= \frac{\alpha}{\gamma} (1 - e^{-\gamma t})\end{aligned}\tag{2.1}$$

where α is the average rate of production and γ is the rate constant for decay of molecule P (rate of decay = $\gamma P(t)$). However, this continuous representation neglects the discrete nature (integral numbers of molecules) and random timing of molecular transitions (Fig. 2.2b), both sources of noise. An actual time evolution could follow many different possible trajectories (Fig. 2.2c). The noise component of any individual trajectory may be isolated by subtracting that trajectory from the average of all possible trajectories in the population (Fig. 2.2c).

A more accurate representation of gene expression includes two coupled ordinary differential equations describing transcription and translation (Fig. 2.2d):

$$\begin{aligned}\frac{dr}{dt} &= \alpha_R(t) - (\gamma_R + \delta)r \\ \frac{dp}{dt} &= k_P r - (\gamma_P + \delta)p\end{aligned}\tag{2.2}$$

Here r and p refer to mRNA and protein concentrations respectively; γ_r and γ_p are decay rates for mRNA and protein ; δ is the rate of dilution due to cell growth (i.e. volume expansion); and α_R and k_P are the production rates for transcription and translation. From these equations the mRNA steady state is $\alpha_R/(\gamma_R+\delta)$ and the protein steady state is $\alpha_R k_P / ((\gamma_R+\delta)(\gamma_P+\delta))$. Noise sources exist at each step of production (transcription and translation) and degradation (of mRNA and Protein) (Fig. 2.2d and Fig. 2.3).

In general, noise sources fall into two main categories (Fig. 2.3): Intrinsic noise, as described above, is attributed to the random timing and discrete nature of molecular interactions occurring during transcription, translation, and degradation processes affecting a single gene. Extrinsic noise is attributed to fluctuations in global resources shared by gene expression off all promoters in the system (e.g. RNAP, ribosomes, amino acids, etc.). In general intrinsic noise is higher in frequency than extrinsic noise [45].

Both intrinsic and extrinsic noise sources are modulated by genetic architecture and regulation (E.g. slow gene activation or autoregulation which will be discussed in more detail in chapters 3 and 4).

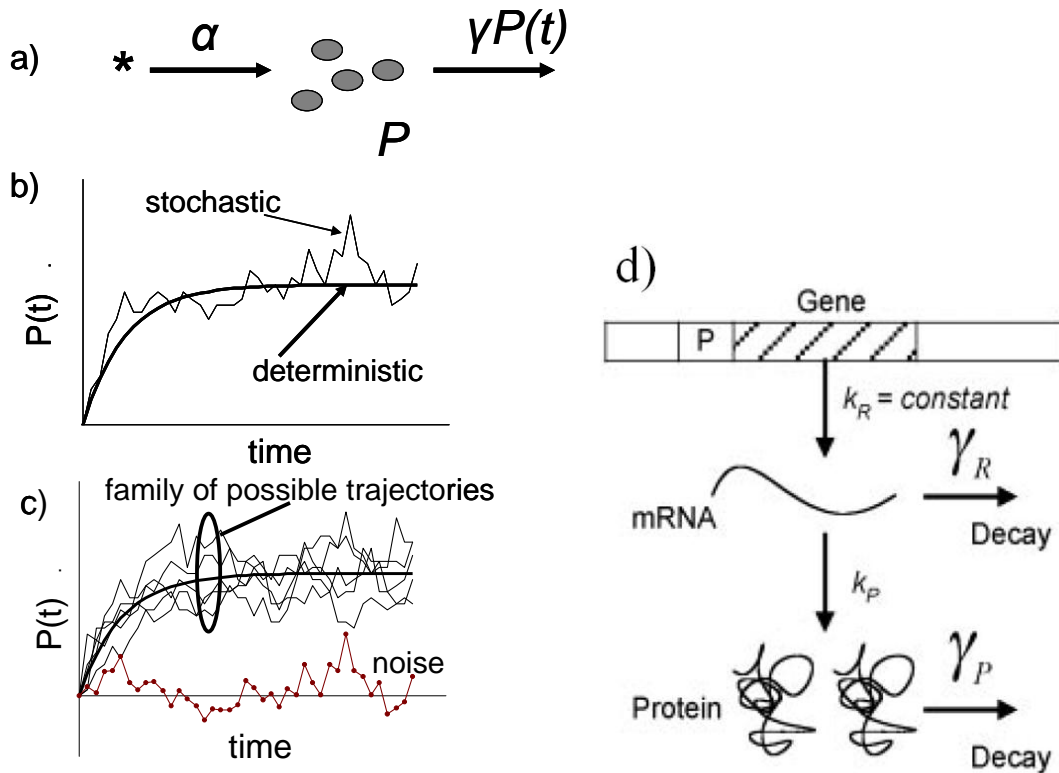


Figure 2.2 Noise in molecular populations. (a) A simple birth-death molecular process and (b) the deterministic and stochastic rise to steady-state molecular population level. (c) A family of possible stochastic trajectories for the birth-death process. The smooth curve represents the average of all possible stochastic trajectories. The noise for any of the possible trajectories is found from the difference between the trajectory and the average of all trajectories. (d) Transcription and translation of mRNA and Protein. Every production and decay step has an intrinsic noise source associated with it. [Figures adapted from [9, 44]].

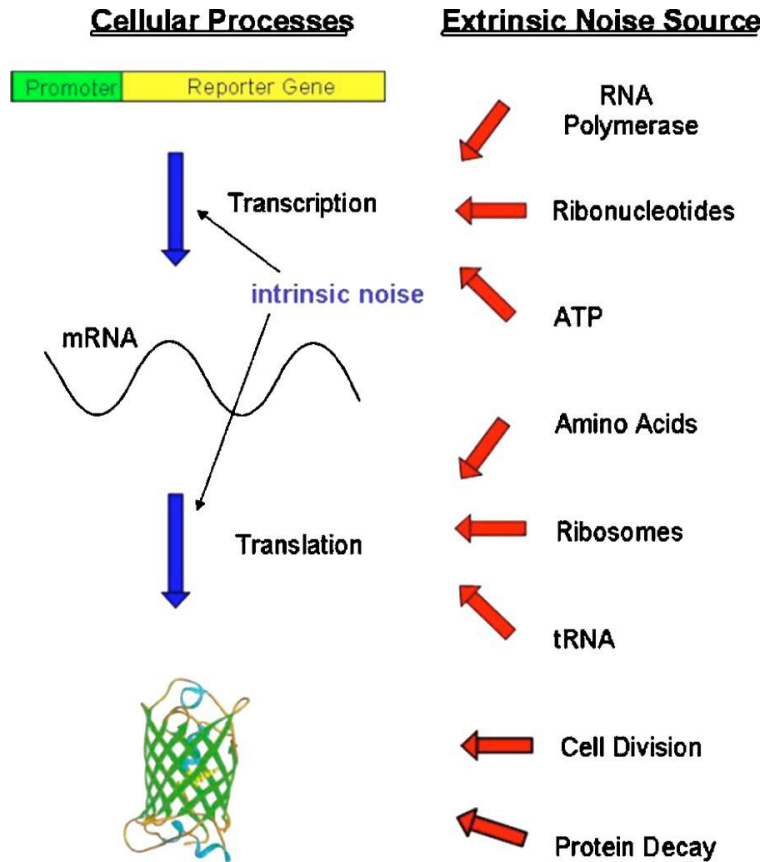


Figure 2.3 Intrinsic and extrinsic noise sources in gene expression. Intrinsic noise is higher frequency noise associated with the random timing and discrete nature of molecular interactions from expression of a single gene. Extrinsic noise is shared by all genes in the cell and is attributed to lower frequency fluctuations in the populations of global resources. Figure adapted from Cox et al, *Chaos* (2006) [46].

2.2.2 Gene expression noise structure

In general, gene expression noise can be characterized by two components: (1) noise magnitude describing variability and the size of fluctuation deviations from the mean protein level, and (2) noise correlation, or frequency content, which describes the characteristic time lag or duration that a fluctuation deviates above/below the mean protein level (Fig. 2.4). The autocorrelation function forms a Fourier pair with the power spectral density and thus has within it frequency information (Fig. 2.4). The general form of the autocorrelation function (ACF) is

$$\Phi(\tau) = E[X(t) \cdot X(t + \tau)] = \int_{t=-\infty}^{+\infty} X(t) \cdot X(t + \tau) dt \quad (2.3)$$

where τ is known as the lag time, $X(t)$ is the noise in gene expression at time t , and E is the expected value. By definition the zero-lag value of the ACF is the variance of the noise signal (Fig. 2.4). Autocorrelation functions for analyzing finite duration noise signals are discussed in the Appendix. It is often convenient to characterize noise correlation by the half-correlation time ($\tau_{1/2}$), which is the value of τ where the ACF has dropped to half of its zero-lag value (Fig. 2.4).

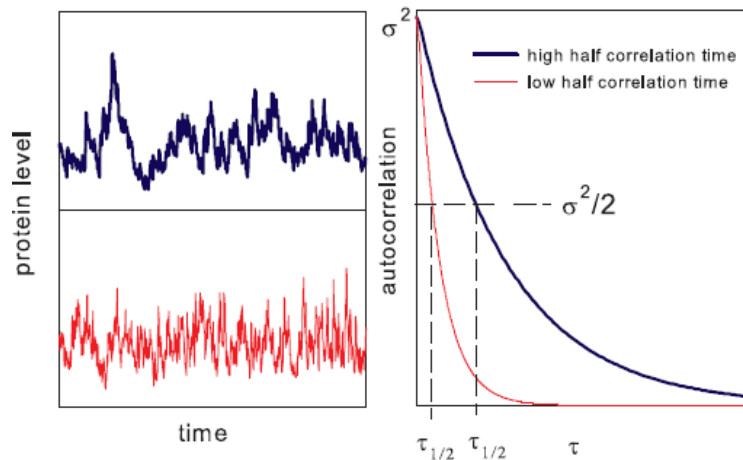


Figure 2.4 Noise autocorrelation yields both variability and frequency content. Time series and autocorrelation functions for two stochastic protein populations characterized by identical $\langle p \rangle$ and variance but different half correlation times, $\tau_{1/2}$. [Figure adapted from Cox et al., 2008[47]]

2.2.3 Analytical and computational methods for gene circuit noise analysis

Chemical Master Equation

The ODEs presented in equations (2.1) and (2.2) model gene expression as a continuous process with no noise. Dealing with a more complete picture of gene expression generally starts with a chemical master equation approach. The chemical master equation (CME) is a fundamental treatment of a model of a set of chemical species (S_1, \dots, S_N) with list of reactions (R_1, \dots, R_N). The state of the system is defined by the state vector $\mathbf{X}(t) = (X_1, \dots, X_N)$, where X_i denotes the number of molecules of species i at time t . The system transitions to a new state when one of the chemical reactions occurs. The propensity that a specific reaction occurs is $a_j(\mathbf{X}(t))$, which is defined as the probability that one reaction R_j , will occur in the system in the time interval $[t, t+dt]$. The reaction propensity is related to the rate constant of the reaction (k_j) and the current state of the system. A system's trajectory comprises a series of transitions from one state to another with the likelihood and frequency of the transitions dictated by the reaction propensities.

The CME describes fundamental properties of the system by knowledge of the probability $P(\mathbf{x}, t)$ that the system evolves into a state $\mathbf{X}(t) = (\mathbf{x}, t)$ at time t according to [48]:

$$\frac{d}{dt} P(\vec{x}, t) = \sum_{j=1}^M \left[a_j(\vec{x} - \nu_j) P(\vec{x} - \nu_j) - a_j(\vec{x}) P(\vec{x}) \right] \quad (2.4)$$

The CME is the most rigorous and accurate mathematical representation for calculating the discrete stochastic time evolution of a reacting system, but its biggest limitation is that it can only be solved for simple systems and becomes impractical for more complex circuits and networks. At present, the CME can address only extremely simple genetic circuits.

Chemical Langevin Equation and Fokker-Planck

Two mathematical simplifications of the CME which have been applied to larger systems are the chemical Langevin equation (CLE)[49] and Fokker-Planck (FP) approaches [49]. The Fokker-Planck equation (FPE) is the time evolution of the continuous system probability density function [49]. In the CLE the state vector $\mathbf{X}(t)$ is treated as a continuous variable as opposed to discrete in the CME, and noise is bundled in a random variable.

The following is a time-dependent chemical Langevin equation describing the production and decay process for a molecule species (M):

$$\frac{dM}{dt} = -(\gamma_M + \delta)M + \alpha_M + \eta_M(t) \quad (2.5)$$

where M is the molecule concentration, α_M is the birth or production rate of M, γ_M its decay rate constant, δ is the dilution rate, and $\eta_M(t)$ is a random variable that represents the noise of the synthesis, decay, and dilution of molecular species M. Gillespie has rigorously examined the CLE and found conditions that allow $\eta_M(t)$ to be approximated as wideband white noise. However, aside from having large molecular populations it is difficult to ascertain if these conditions are met within a particular system [49]. An analogy to shot noise in electronic systems was used to demonstrate that certain linear genetic circuit processes, such as mRNA and protein synthesis, exhibit wideband white noise even at low molecular populations [50]. Cox *et al.* show that this holds true even for some non-linear processes [51]. At present, most investigators either make this white noise approximation for $\eta_M(t)$, or resort to computation methods based on the CME.

Stochastic Simulation

For cases of complex gene circuits with strong nonlinear behavior or the interaction of several genes, the analytical approaches become impractical. In such cases, exact stochastic simulation has been used to provide individual trajectories of possible time evolutions of the system. Extensive simulations can yield statistics such as distributions, variances, and autocorrelation functions. A widely used simulation

approach proposed by Gillespie [52] is equivalent to Monte Carlo simulation of the CME and considered to be exact. This Exact Stochastic Simulation (ESS) approach is demonstrated and described in more detail and in the Appendix.

2.3 Measuring Noise in Gene Circuits

2.3.1 The Green Fluorescent Protein (GFP)

Since its discovery the green fluorescent protein (GFP) has had a significant impact on biology. While biologists could previously only characterize with biochemical reporters and assays, the fluorescent protein allowed the visualization of gene expression and cellular structures and components. Its wide range of applications and heavy impact on molecular biology research earned Martin Chalfie, Osamu Shimomura, and Roger Tsien the 2008 Nobel Prize in chemistry for their discovery and development of GFP.

2.3.1.1 Brief History

In the 1960's and 70's Osamu Shimomura isolated GFP from *Aequorea Victoria*, a jellyfish species that fluoresces green (509 nm) when exposed to blue light (395 nm). In *A. victoria* aequorin (a photoprotein) interacts with Ca²⁺ ions inducing blue light which is partially absorbed by GFP which in turn emits green light (Fig. 2.5). In 1992 Douglas Prasher reported the cloning and nucleotide sequence of wild type wt-GFP [53]. Later in 1994 Martin Chalfie's lab expressed the coding sequence of fluorescent GFP in heterologous cells of *E. coli* and *C. elegans* [54]. Remarkably, the GFP molecule folded and was fluorescent at room temperature, without the need for exogenous cofactors specific to the jellyfish. Although this wt-GFP was fluorescent, it had several drawbacks, including dual peaked excitation spectra, poor photostability and poor folding at 37°C. In 1996 Remington's group reported the first crystal structure of a GFP S65T mutant [55]. Also in 1996 Phillips group independently reported the wild type GFP structure [56]. These crystal protein structures were vital for understanding protein formation and residue (amino acid) arrangement.

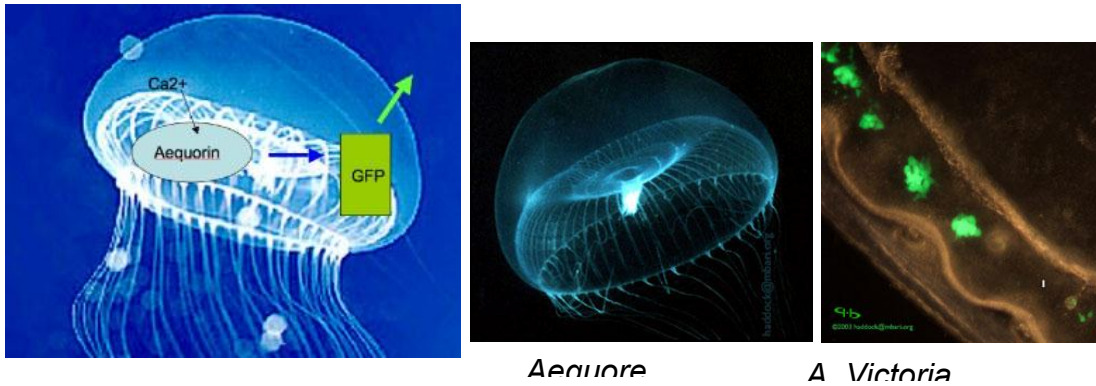


Figure 2.5 Green fluorescent protein originates in a jellyfish species. *Aequorea Victoria* has blue light emitted from a Ca^{2+} and Aequorin interaction (**left**) which excites GFP emission from the jellyfish photo-organs (**right**) (Photocredit: Steve Haddock and the Monterey Bay Aquarium Research Institute).

2.3.1.2 Basics of Fluorescence

Fluorescence is a luminescence in which the molecular absorption of a photon triggers the emission of another photon with a longer wavelength (Fig. 2.6). The energy difference between the absorbed and emitted photons determines the final electron energy level and dissipates as molecular vibrations or heat.

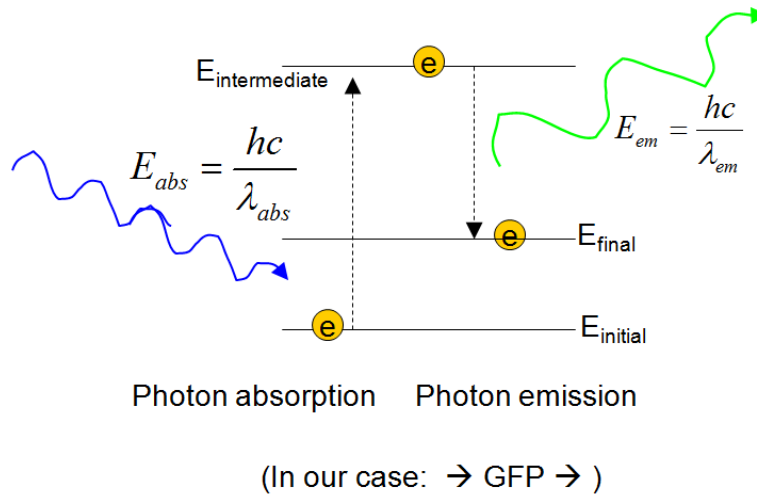


Figure 2.6 Stimulated emission of a photon. When imaging GFP under a fluorescence microscope the incident excitation photon is 488 nm (Mercury lamp source) and emitted light is green in the 500-550 nm range.

2.3.1.3 Fluorescent Reporters for Detection of *In Vivo* Protein Levels

Fluorescent reporter expression in live cells (*in vivo*) is used to monitor gene expression and protein levels. Reporters may be either transcriptionally or translationally fused. Transcriptionally fused reporters are co-transcribed with the target of interest, but translated into an individual and separately functioning protein. Translationally fused reporters are co-transcribed and co-translated with the target of interest into a single protein, where (hopefully) the function of both the target of interest and the reporter maintain their function. It has been shown that even with the fused fluorescent protein most target protein functions are conserved. Reporters that fluoresce at many different wavelengths, including blue, yellow, and red, were created by mutating the fluorophore core of GFP [56-58]. There are three fluorescent reporter libraries of interest for genome-wide investigation:

1. In 2003, Huh *et al*, reported the creation of a protein-GFP fusion library for about 2/3rds (~4k genes) of the budding yeast genome using homologous recombination downstream of every gene's open reading frame (ORF) (Fig. 2.7)[59]. They then used this library for a global protein localization study in yeast [59].
2. In 2008, Cohen *et al*, reported the construction and imaging of 1000 endogenously yellow fluorescent protein (eYFP) tagged proteins in human cancer cells [35].
3. In 2010, Taniguchi *et al*, reported the creation and characterization of an *E. coli* library consisting of over 1000 chromosomal YFP-protein fusions [32].

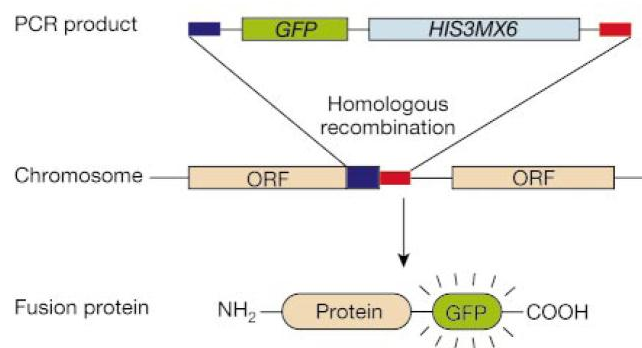


Figure 2.7 Protein-GFP fusion library covering 4159 budding yeast genes. (Image adapted from Huh *et al*, (2003) [59])

2.3.2 Flow Cytometry – A High-Throughput Fluorescence Screener

The Flow Cytometer (FC) has been extensively used for fluorescence based sorting of cellular sub-populations and to characterize the fluorescence distribution of cellular colonies by flowing large numbers of single cells within a streamline of single-celled width across an excitation laser source (Fig. 2.8) [33, 34, 60, 61]. This measurement enables the collection of large samples of single cell fluorescence distributions. E.g. ~50k cells can be flowed in ~1 min (Figures 2.8 and 2.9).

A typical FC is shown in figure 2.8. This model (an Influx Fluorescence Activated Cell Sorter (FACS) by Cytopia (2006), now BD) contains various laser sources and an array of detector channels for different ranges of emission wavelengths. The ability to excite and detect different ranges of wavelengths enables experimental measurements involving the expression of more than one fluorescent reporter in each cell. The laser beam is directed through a cellular flow cell and both forward and side scattered light are collected. Forward scattered light yields information regarding cell size while side scattered fluorescent light is directed and filtered to an array of photon multiplier tube (PMT) detectors. Side scattered light also yields information about granularity and membrane integrity for an additional characterization of overall cellular health. Examples of raw fluorescence intensity data from the FC in figure 2.8 using selected targets from the yeast protein-GFP fusion library described in the previous section are shown in figure 2.9. These distributions allow the quantification of the moments of the population, such as the mean, the variance, and the coefficient of variation ($CV = \text{standard deviation of fluorescence} / \langle \text{fluorescence} \rangle$).

Although of great value for its efficient and high throughput single-cell fluorescence measurements figure 2.8 (lower) illustrates why FC is unable to measure gene expression correlations. A FC collects sequential measurements taken from separate stochastic processes, and as a result all correlation information is lost. Correlation information is captured using time-lapse single cell fluorescence microscopy, which is described below.

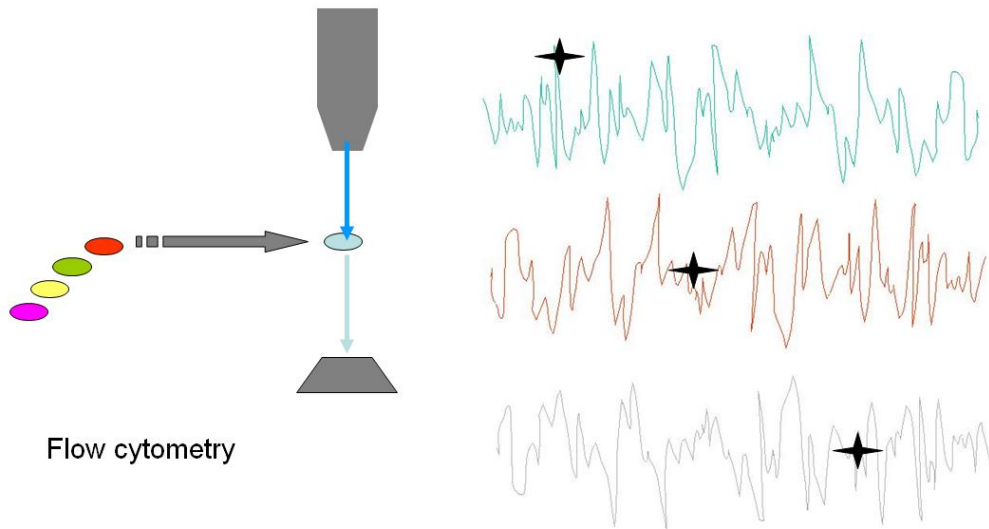
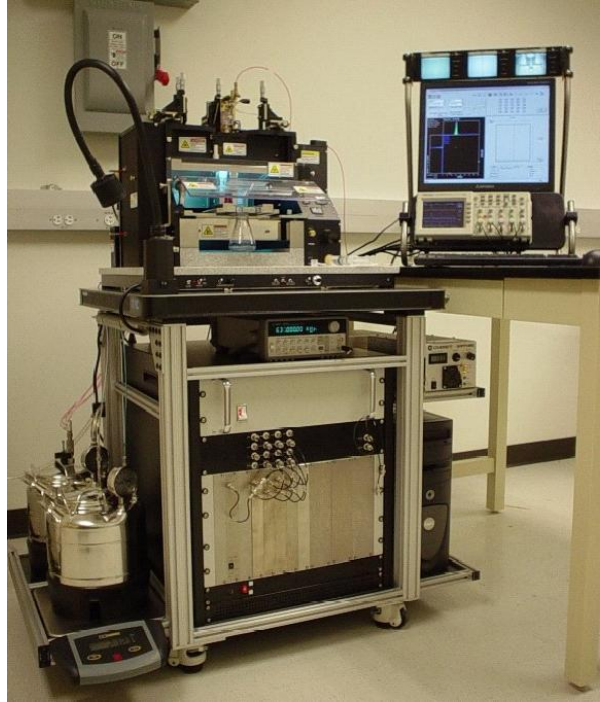


Figure 2.8 Flow cytometry – a high throughput fluorescence screener. (above) Influx Fluorescence Activated Cell Sorter (FACS), (bottom) Schematic of flow cytometry operation. A single-file stream of cells passes in front of a laser source and detector (bottom-left). FC collects sequential measurements taken from separate stochastic processes from within each individual cell, and as a result all correlation information is lost (bottom-right).

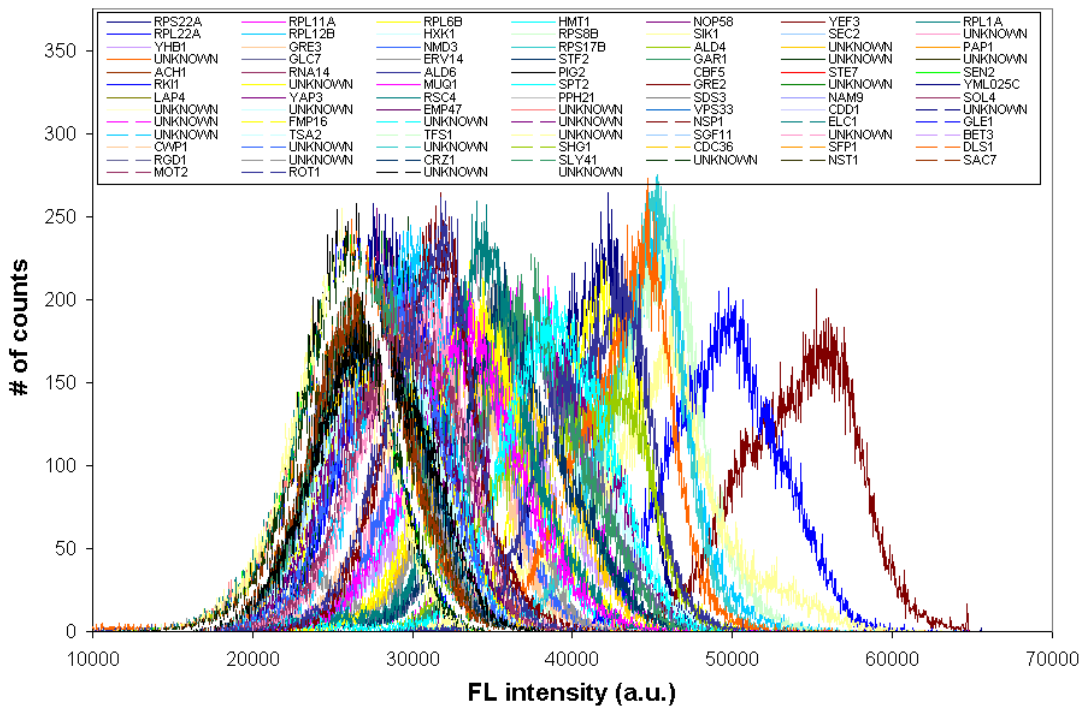


Figure 2.9 Raw FC fluorescence distributions of 88 budding yeast protein-GFP fusion populations. (Unpublished experiments recorded with D. Karig and S. Allman at ORNL).

2.3.3 Correlation Spectroscopy Methods

Fluorescence correlation spectroscopy (FCS) was founded in the early 1970s by Magde, Elson, and Webb [62-64]. In FCS a pulsed laser beam excites fluorescent proteins in a detection volume of limited size (Fig. 2.10). On short time-scales (msec and sec) the fluctuating emission signal provides quantitative information about processes such as diffusion coefficients, hydrodynamic radii, average protein concentrations, and kinetic chemical reaction rates (Fig. 2.10, red signal) [65]. The fluctuating signal frequency content is analyzed using autocorrelation or power spectral density analysis (and will be described in more detail later). In 1998-2001, Wiseman *et al.* developed an image correlation spectroscopy (ICS) approach which utilizes the excitation and detection in the pixels of a time-lapse or spatial fluorescence microscopy image [66-68]. Rather than utilizing a pulsed laser source such as in FCS, ICS capitalizes on the scanning laser source imaging in a confocal fluorescence microscope. This extends the spatial range and dimension of the correlation spectroscopy as the pulsed laser is confined to a precise detection volume while scanning lasers cover a larger region of a biological sample (See application differences between FCS and ICS in Table 2.1). Analyzing gene expression fluctuations with an autocorrelation analysis is a natural long-time scale extension to FCS (or ICS) methodology (Fig. 2.10, blue signal). Here, longer experiments (hours to days) and imaging intervals (minutes) enable the detection and correlation analysis of gene expression with longer duration time constants such as protein production and degradation, protein dilution, gene regulation, and their role in larger-scale cellular function. For comparison, this long time-duration FCS method is termed “Gene Expression Noise Correlation Spectroscopy” or GENCS (see Fig. 2.10 and Table 2.1).

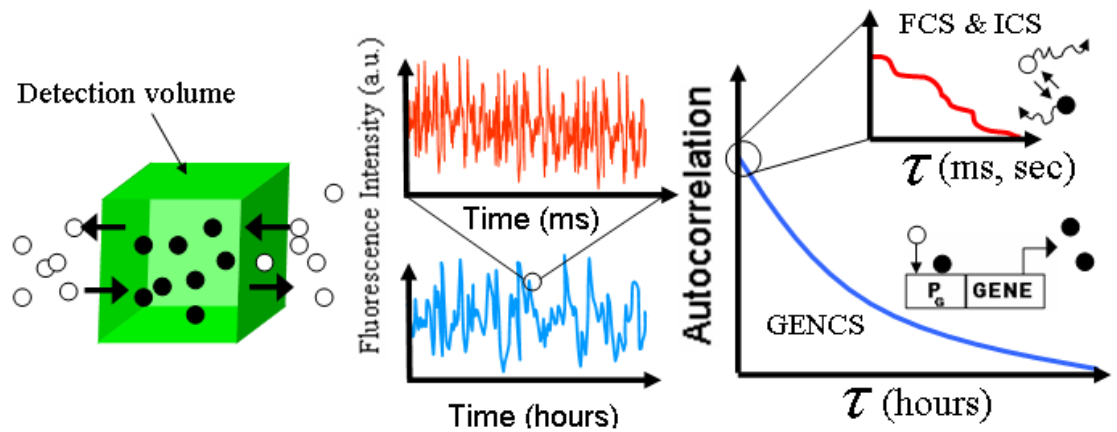


Figure 2.10 Correlation spectroscopy methods. (middle and right) Conventional FCS and ICS sample a detection volume (left) with msec time scale fluctuations (red signal and correlation function). These measurements are too short to quantify the longer time constants underlying gene expression which are on the order of several to many hours (blue signal and correlation function).

Table 2.1 Summary of Correlation Spectroscopy Methods

Method	FCS	ICS	GENCS
Application	Triplet state and photobleaching Diffusional mobility in cells Reporter concentration Protein transport Protein-Protein binding kinetics GFP properties	Characterization of membrane protein cluster densities and sizes Cell structure studies	Hill kinetics Gene regulation Protein production and degradation Protein dilution Gene Activation Kinetics
Source of fluctuation in the detection volume	Diffusivity Reporter interaction Chemical binding affinities	2D and 3D diffusivity of aggregates and structure bound reporters	Gene activation/repression dynamics Random nature of transcription and translation Variations in shared global resources to gene expression Chromatin Remodeling
Timescale of fluctuation source	1usec - 100msec	msec	sec, minutes, hours
Time sampling Temporal resolution	100 nsec - 10 usec <100nsec	usec	5-10 min
Typical experiment duration	1 sec - 10 min	1 sec - 10 min	4 - 48 hours

FCS – Fluorescence Correlation Spectroscopy (Magde, Elson, and Webb, ~1972) [62-64]

ICS – Image Correlation Spectroscopy (Wiseman et al, ~2000) [66, 68]

GENCS – Gene Expression Noise Correlation Spectroscopy (~2005-2006) [45, 69, 70]

2.3.4 Time-Lapse Single-Cell Fluorescence Microscopy

Time-lapse single-cell fluorescence microscopy provides the single cell tracking needed for gene expression noise correlation analysis. Although providing information that flow cytometry cannot, fluorescence microscopy has lower cell throughput, which usually demands multiple experiments to achieve suitable statistics. Figure 2.11 is a schematic representation of a confocal microscope. A laser excitation source passes through an aperture, filter, and objective before illuminating a focal plane of the biological specimen. Emitted fluorescence light passes through a dichroic mirror, filter, and aperture before detection by a photon multiplier tube (PMT). The focal plane can be adjusted to collect a 3-dimensional sample image with the use of mirrors and optics. Similar to the flow cytometer, the confocal microscope offers multiple laser excitation sources and detectors for diverse applications and function in a wide-range of excitation and emission wavelengths.

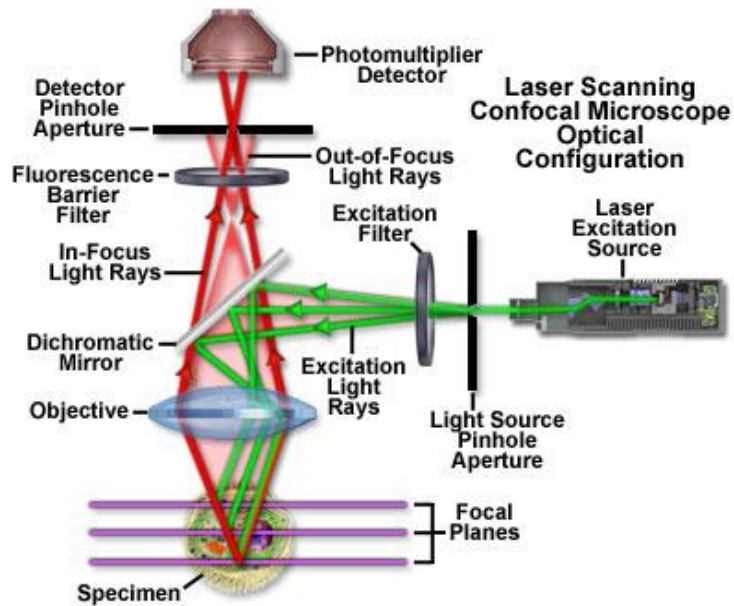


Figure 2.11 Schematic of confocal microscopy operation. A variety of apertures, filters, mirrors, and optics are precisely controlled to excite and acquire fluorescence emission from a defined focal plane in the biological specimen. [Image adapted from <http://www.olympusmicro.com/>].

The three microscope resources used for long duration imaging experiments in this dissertation include:

1. Upright Leica Laser Scanning Confocal Microscope
2. Inverted Leica Laser Scanning Confocal Microscope.
3. Inverted Olympus Spinning Disk Confocal Microscope.

Details on the Leica laser scanning microscope and Olympus Disc Scanning Unit (DSU) can be found at:

1. <http://www.leica-microsystems.com/products/confocal-microscopes/details/product/leica-tcs-sp2/>
2. http://www.olympusamerica.com/seg_section/product.asp?product=1009

Figure 2.12 shows the Olympus Disc Scanning Unit (DSU) used in the Weinberger Laboratory (University of California, San Diego) for all human cell experiments presented in this dissertation. Here the confocal scan method is a disc rotation method and the excitation wavelengths are 350nm-600nm. The environmental chamber allows the control of temperature, humidity, and CO₂, which are all needed for maintenance of mammalian cell cultures. Fluorescent light is collected by a cooled CCD camera.

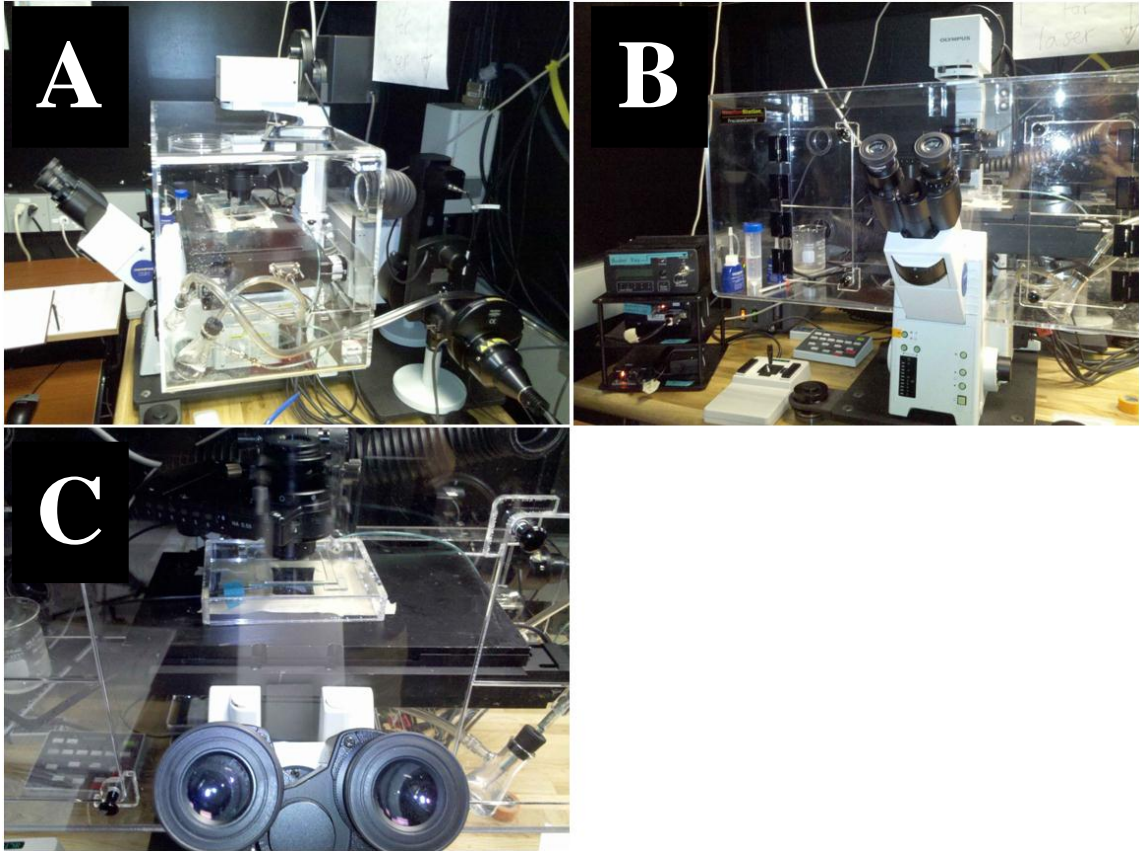


Figure 2.12 Automated Olympus Spinning Disc Confocal Microscope. **A.** side view of microscope, sample stage, and environmental chamber, **B.** front view of the inverted microscope. **C.** close-up of the motorized sample stage.

2.3.5 Image Processing of Single-Cell Gene Expression Experiments

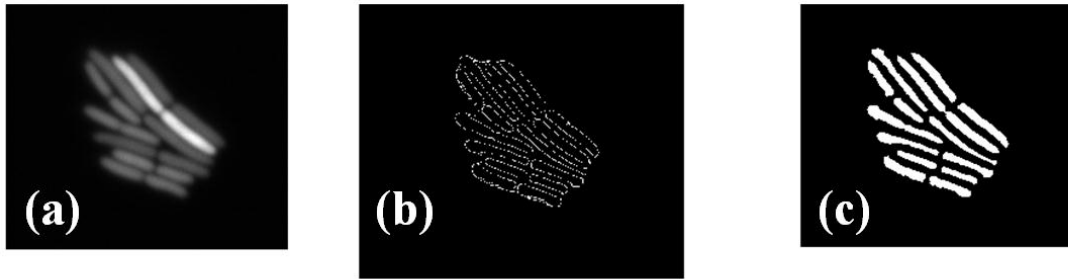
An important step in characterizing time histories of single-cell protein levels is the image processing of the time-lapse fluorescence microscopy image stacks. Typical image processing and quantification requires three main steps: (1) single-cell image *segmentation* to isolate and identify the pixel region within the image belonging to a specific cell; (2) single-cell *tracking* (i.e. following the same cell from image to image); and finally (3) *quantification* of the level of fluorescence at each pixel in the image.

To date many quantitative single-cell studies have utilized custom programmed image processing algorithms from within their labs with few standardized image

processing programs available. Approaches vary depending on many variables: (1) fluorescence microscopy system, camera quality, and image type; (2) experimental sample, cell type, reporter type(s), and gene circuit or expression dynamics being observed; and (3) image processing algorithm including a palette of segmentation, tracking, and quantification approaches. For a review on time-lapse single-cell fluorescence imaging see Locke and Elowitz, (2009) [71]. This section covers the main processing methods used throughout the studies of this dissertation. It is possible that in the future, long time-lapse experiments will have standardized imaging equipment and processing protocols in the single-cell research community. This will enable easier collaboration through experiment repeatability and sharing of large single-cell resources in a public-domain database.

2.3.5.1 Cell Segmentation

Figure 2.13 depicts one of many image processing algorithms used from MatlabTM's image processing toolbox that have worked for segmenting single cells. Here edge detection uses the Laplacian of Gaussian (or 'LOG') filter of the grayscale intensity image to find a 2-dimensional zero-curvature border of the fluorescent cell with the dark image background. As shown in the figure, after retrieving the detected edge (Fig. 2.13b) a 'fill' operation can be used to identify a labeled pixel region for each identified cell (Fig. 2.13c). Each label region, or single cell can be colored with a different color and superimposed on the original image to manually check segmentation and tracking quality of the custom program (Fig. 2.13d). Sometimes additional image manipulation techniques are needed such as pixel dilation and erosion to help fill in and bridge the edge being detected (these are operations that use nearest neighbor logic to either add or subtract labeled pixels). Figure 2.14 shows an example of a raw intensity and segmented image of human T-cells.



Edge detection based on 'log' filter of fluorescent images
Laplacian of Gaussian

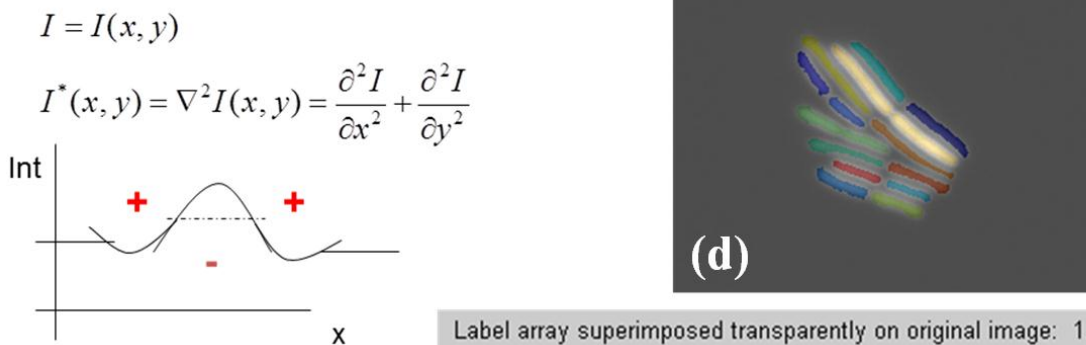


Figure 2.13 LOG image segmentation of bacterial cells. (a) Raw grayscale fluorescence intensity image. (b) Application of a 'log' filter to the image in (a), where (lower-left) the 'log' filter looks for zero-crossings in the second spatial derivative of the intensity image in (a). (c) After applying additional morphological operations the spaces and breaks in the 'log' filter result are bridged and filled for individual cell regions. (d) Image segmentation is tested for quality by superimposing the segmented pixels of each cell transparently on the original image from (a).

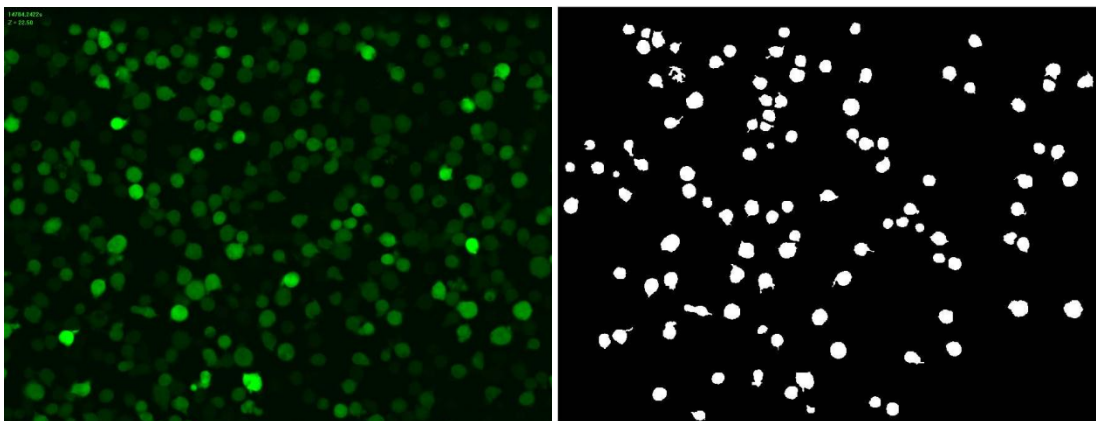


Figure 2.14 A fluorescent T-cell image and its segmented binary array. Small contaminants or very large connected cell regions are filtered out by testing for cell area sizes in the total number of pixels.

2.3.5.2 Cell Tracking

Automated tracking of a cellular colony growing in a monolayer

A straightforward strategy for automated cell tracking of a growing colony of *E. coli*, budding yeast, or human T-cell is to use a nearest neighbor approach. This assumes that the cell in the next image that is closest to the cell's previous image position is in fact the same cell. Cells are segmented from the last image and tracked backwards in time to the first image, connecting daughter cells to their parent cell lineage (Fig. 2.15). Assuming that the imaging interval is short enough that the cell has not moved too far, there should be some spatial overlap between the cell location in consecutive images. It is convenient to use the cell centroid and 'connect' image label regions (a single-cell's pixel region) by projecting the cell centroid of the n^{th} image to the cell's corresponding pixel region in the $(n-1)^{\text{th}}$ image. High quality images are needed to assure proper segmentation and tracking. In addition a cell must grow in a monolayer for this tracking strategy to work.

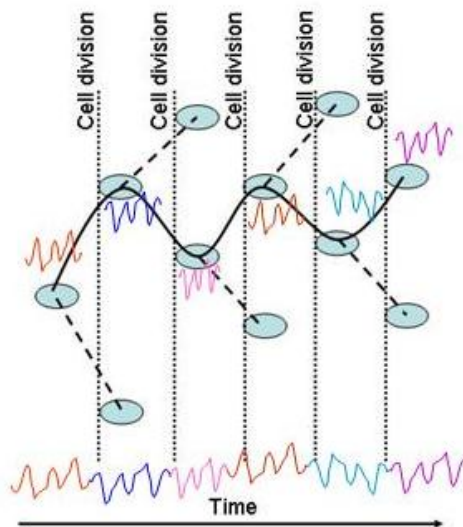


Figure 2.15 Schematic of a single lineage or trajectory in a growing *E. coli* colony. The solid curve shows a single trajectory through five generations of cell growth and the dashed lines show alternate routes that produce other trajectories. A representative noise trace is shown next to each cell in the trajectory. The noise trace of the complete trajectory (shown at the bottom) is constructed by sequentially combining the noise traces of each cell in the trajectory.

2.3.5.3 Fluorescent Intensity Quantification

The third and final stage of image processing is quantifying the fluorescence intensity for each cell in each image. A whole cell time-dependent fluorescence intensity concentration can be calculated using the following equation:

$$[I_i(nT_s)] = \frac{\sum_{j=1}^{N_i(nT_s)} I_j(x, y, nT_s)}{N_i(nT_s)} \quad (2.5)$$

where $[I_i(nT_s)]$ is the single cell intensity concentration of cell i at time nT_s where $n=1, 2, 3, \dots$ is the image number, and T_s time between samples, $I(x, y, nT_s)$ is the intensity of pixel (x, y) in the image array at time nT_s , $j = 1, \dots, N_i(t)$ are the pixels identified as belonging to the i^{th} cell at time nT_s , and $N_i(nT_s)$ are the total number of pixels belonging to cell i at time nT_s .

The Voxel Method

Low cell intensities or complex cell morphologies often limited the ability to automatically segment cell borders with image processing algorithms, which often resulted in heavily work intensive manual tracing of individual cells in each image interval. To solve this problem a new method (the voxel method (from VOlumatic piXEL)) was developed and relieved the bottleneck by avoiding the troublesome cellular boundaries. The voxel method uses a limited sampling region of the fluorescent cell interior. The z (out of plane) dimension of the voxel box comes from the thickness of the image slice provided by the confocal microscope. Voxel tracking of 4-8 single *E. coli* cells over time yielded the same autocorrelation functions as their whole cell tracking counterparts (Figures 2.16 and 2.17). A voxel of 9x9 pixels was found sufficient on a 512x512 pixel image of a 40x40 μm sample region (Fig. 2.17). The method must be carefully applied if fluorescence is localized in particular regions of the cell, and works best where fluorescence is uniformly distributed across the entire cell region.

This processing modification significantly increased experiment throughput (Fig. 2.18) and led to a larger sample size for the analysis of single cell autocorrelation distributions and composite population level expression dynamics (see chapter 3). The voxel method has since been applied to both yeast and T-cells and has important advantages over whole cell segmentation methods including:

1. Easily tracks cells with difficult morphologies
2. Can track sub-cellular localizations (e.g. nucleus localized proteins)
3. Insensitivity to fluctuations in the area of the cell bound to the surfaces.

Fluctuations in the area of the cell membrane adhered to the surface results in two types of problems: (1) membrane fluctuations may bridge and promote contact between neighboring cells impeding the whole cell segmentation; and (2) adhered cell area may be a non-fluorescing cell component (e.g. flaps of membrane), and add a fluctuation component due to cell adhesion between true and perturbed fluorescence concentrations.

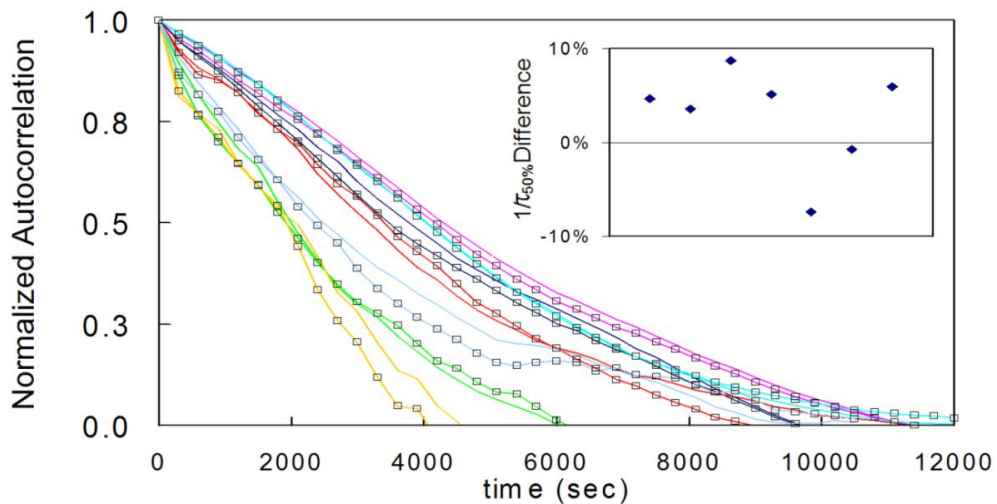


Figure 2.16 Equivalent autocorrelations for voxel and whole cell sampling. In the figure each *E. coli* experiment autocorrelation is depicted by a different color. The lines with no boxes result when fluorescence is measured from whole cells, while the lines with the open boxes result when fluorescence is measured using the voxel method. The inset shows that $1/\tau_{1/2}$ (noise frequency range (F_N)) measured using the voxel method are within 10% of those found using the whole cell method.

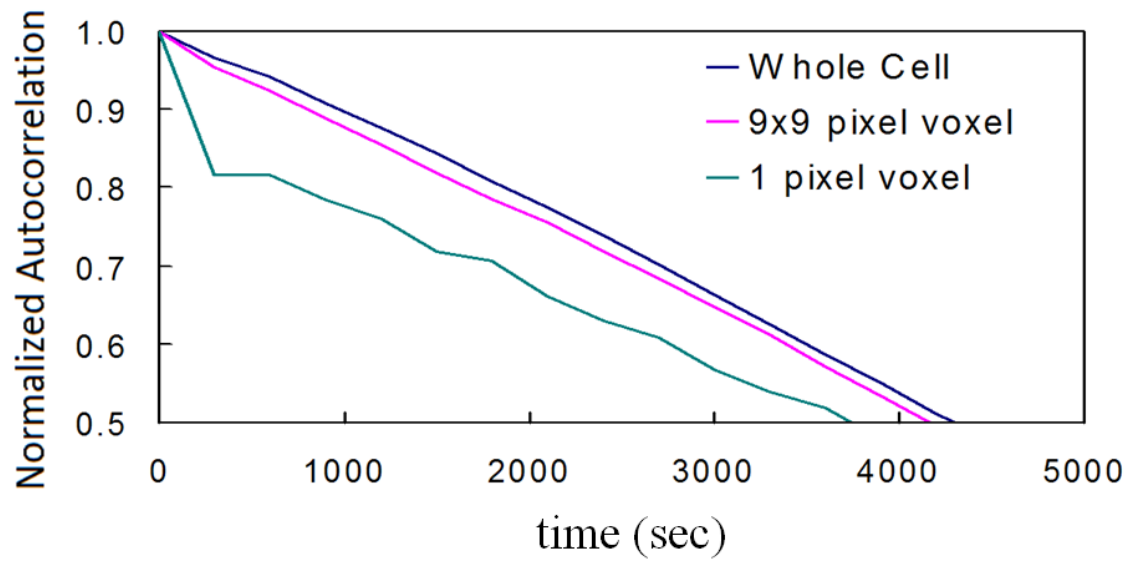
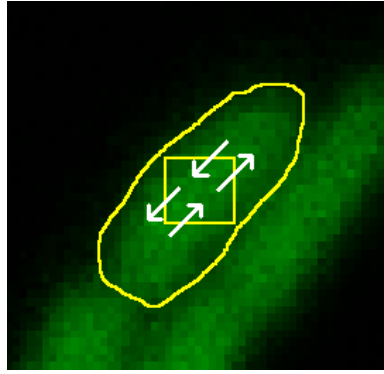


Figure 2.17 Limited sampling introduces white noise variance attributed to diffusivity of fluorescent particles in and out of the voxel region (**upper**) and is dependent on the size of the voxel. This diffusion white noise becomes negligible for voxels that are 9 pixels on a side or larger (**lower**).

2.3.5.4 Time-line of increase in single-cell experimental throughput

Figure 2.18 illustrates how advances in cell tracking has significantly increased experimental throughput for the work presented in this thesis. Initial *E. coli* experiment processing was completely manual and demanded tracing whole cells at each imaging interval. Introduction of a semi-automated voxel method in which voxel ‘seeds’ were manually planted using an efficient seed-planting program increased throughput about 2 orders of magnitude, enabling the determination of single cell autocorrelation distributions [70]. Later when working with monolayer T-cell samples, a fully automated single node voxel processing program was implemented [15] (see Appendix). Finally in recent years, using a motorized X-Y sample stage, automated voxel processing of multiple-node experiments increased throughput by almost another order of magnitude. If a total of 500 cells are collected from about 25 nodes then the throughput is about 20 cells per node and further experiment optimization continues the throughput increase. This throughput of 300-800 tracked cells is significant as many single-cell systems and synthetic biology published studies typically use a hundreds to thousands of single-cell measurements. On top of this, with the novel shotgun polyclonal noise mapping approach described in chapter 4, 500-1k cells per experiment enables a fair coverage of genome-wide behavior within 1-2 weeks of overnight imaging experiments.

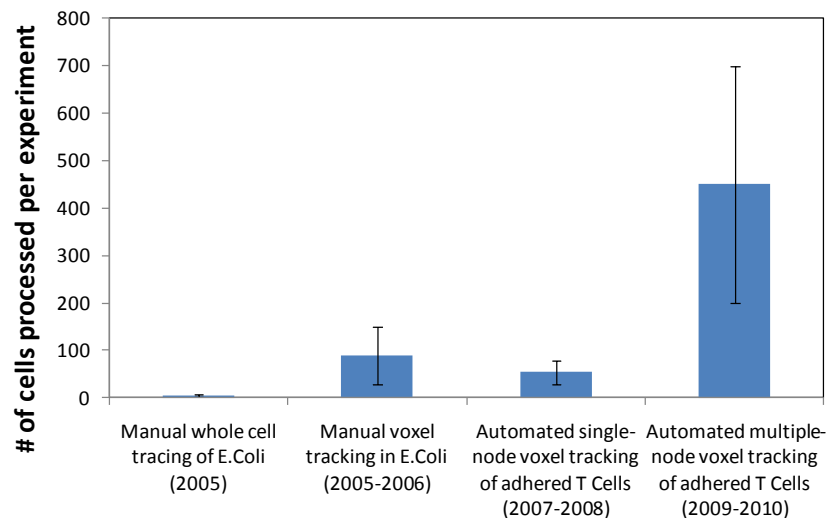


Figure 2.18 Timeline of experimental methodology and single-cell throughput increase.

2.3.6 Fluorescent Intensity Signal Processing

After preparing the cell sample, running the time-lapse microscopy experiment, processing the images, quantifying and quality controlling for successful single cell gene expression trajectories, signal post-processing may commence. The signal processing has two main requirements: (1) separating the stochastic from the deterministic components of gene expression; and (2) calculating the autocorrelation function for the stochastic components of gene expression. These two requirements are met using a multiple-step signal processing algorithm described below.

2.3.6.1 Separating Stochastic Expression from Deterministic Expression

To extract gene expression fluctuations from a deterministic process the following scaled noise definition is implemented:

$$N_i(nT_s) \equiv (I_i(nT_s) - g_i \cdot A(nT_s)) / A(nT_s) \quad (2.6)$$

where: $i = 1, \dots, M$, represent each of the M single cells tracked in an experiment; $N_i(nT_s)$ is the single-cell noise trajectory; $I_i(nT_s)$ the single-cell intensity trajectory; $A(nT_s)$ is the deterministic general intensity trend of an experiment defined by

$$A(nT_s) = \frac{\sum_{i=1}^M I_i(nT_s)}{M}; \quad (2.7)$$

T_s is the time between samples; n is the image number; and g_i is a gain factor that describes the extent to which the general trend couples into each individual noise trajectory.

The gain factors, g_i , allows for the general deterministic trend to couple into each cell gene expression with variable strengths. They are found from the zero-crossing of the zero lag cross-correlation between the gain dependent noise trajectory and the

deterministic trend

$$N_i(nT_s) \otimes \tilde{A}(nT_s) \quad (2.8)$$

or more explicitly the g_i which minimizes

$$\left| \sum_{n=0}^{\infty} N_i(nT_s) \cdot \tilde{A}(nT_s) \right| \quad (2.9)$$

where $\tilde{A}(nT_s)$ is the mean suppressed time dependent general intensity trend.

$$\text{i.e.} \quad g_i \equiv \left(g_i \left| N_i(nT_s, g_i) \otimes \tilde{A}(nT_s) = 0 \right. \right) = \left(g_i \left| \min \left(\sum_{n=0}^{\infty} N_i(nT_s, g_i) \cdot \tilde{A}(nT_s) \right) \right. \right) \quad (2.10)$$

Table 2.2 and figure 2.19 summarize the above noise processing algorithm steps and figure 2.19 shows that single cell gain factor histograms cluster around a value of 1. For experiments with frequent doubling events (e.g. bacteria), well stirred and homogeneous behavior can be assumed and the gain factor may be set to a value of 1 for all cells [70]. In such cases, it is reasonable to expect all cells to contribute equally to the population general trend. This is not the case for less mixed populations (e.g. slow-growing eukaryotes [15]). For these cases, individual cells may decouple from population behavior for a variety of reasons including, different local environmental conditions, different distributions of genetic circuit copy number (a consequence of cell duplication events and applicable to DNA plasmid experiments), and differences in basal gene expression levels.

Table 2.2 Summary of the noise processing algorithm

Noise Processing Component	Purpose
g_i	Decouples general intensity trend from single cell intensity on a single cell basis
g_i derived by minimizing cross correlation of $\tilde{A}(nT_s)$ with $N_i(nT_s)$	Removes any excitation driven deterministic component of the single cell trajectory. The mean-suppressed general trend ($\tilde{A}(nT_s)$) is used instead of the general trend ($A(nT_s)$) to determine the value of g_i and avoid over correction due to correlation of the baseline shift of the noise and the average value of $A(nT_s)$.
Scaling by $A(nT_s)$	Normalization of fluctuation with time dependent expression level so that fluctuations in reporter protein level are viewed in relation to the total protein population (i.e. fluctuation of 10 units in total population of 100 units is equivalent to a fluctuation of 1 unit in total population of 10 units).
Suppression of the baseline of individual noise trajectories	Removes the deterministic portion of the single cell trajectories that remain after the above corrections (e.g. baseline shifts due to differing basal gene expression levels). However, this also removes differences in baseline levels that are due to slow stochastic fluctuations (for more see the next section 2.3.6.2).

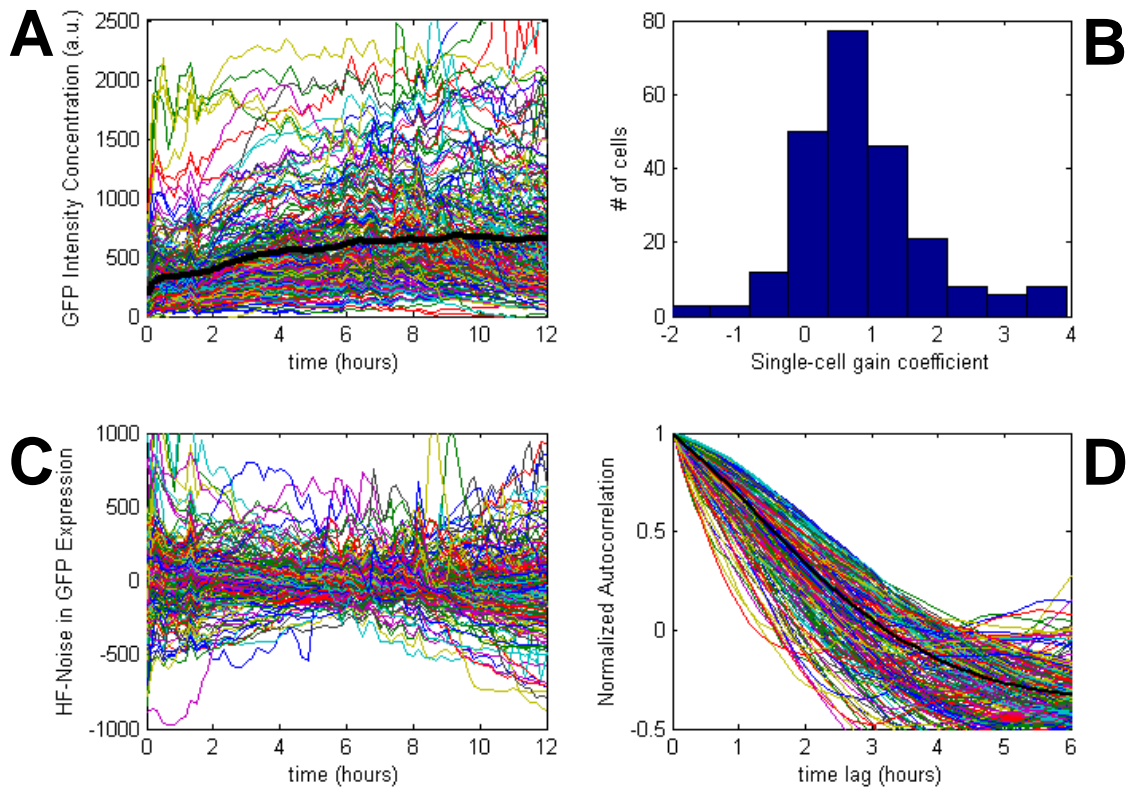


Figure 2.19 Gene expression HF noise processing steps of a typical fluorescence experiment. The above plots are a summary of the individual noise processing steps mentioned in the signal processing algorithm (Eqn 2.6) followed by high pass mean suppression. These plots are typical outputs collected when processing any individual experiment and allow manual inspection of each processing step. **A.** Raw GFP intensity concentration versus time for a population of cells, **B.** Histogram of individual single-cell gain coefficients, centered around $g_i=1$. **C.** HF-noise trajectories, individual mean suppression usually causes a “bow tie” where the center is pinched and the trajectory edges deviate more, **D.** Final normalized single-cell ACFs.

2.3.6.2 High Frequency (HF) Noise Processing

An unavoidable reality of long duration experimentation with slow gene expression processes, such as slow cellular growth, is that a low frequency component of a single cell gene expression trajectory is indistinguishable from a difference in basal gene expression levels (Figure 2.20). However, the noise processing described above will always calculate the long correlation time of the red trace in Figure 2.20, and thereby calculate erroneously long correlations for many cells. An alternate approach is to individually mean suppress each trajectory (Fig. 2.21, middle) before the calculation of its autocorrelation function. Mean suppression is high-pass filtering that removes lower-frequency fluctuations but preserves the higher-frequency fluctuations (see Appendix). Autocorrelations derived from these high-pass filtered (HF) noise traces are referred to as HF-ACFs. The overall steps for HF-noise processing are displayed in Figure 2.19.

Other than suppressing the ambiguous portion of the data in Figure 2.20 and 2.21 left, the mean suppression and high pass filtering essentially emphasizes or focuses on high frequency intrinsic noise and significantly attenuates the low frequency extrinsic noise (Fig. 2.22). To observe this effect, a simulation of constitutive gene expression using various levels of extrinsic noise using the noise simulation model described in Austin *et al*, (2006) [70] was implemented. Since the intrinsic noise is directly modulated by gene circuit structure and function, the high pass filtering enhances the quality of autocorrelation analysis for understanding the gene circuit function without an additional extrinsic noise background (such filtering is not possible through gating in flow cytometry and provides another advantage for fluorescence microscopy, i.e. extrinsic noise not related to cell-cycle and growth can get through flow cytometry cell gating). For example figure 2.22 shows that a process with 40% extrinsic noise is filtered down to 5% using the 12 hour HF processing, 55% down to 15% and so on.

Finally it is worth noting that although infinite duration analysis and measurements would be informative (if biology would not be so hard to observe over long periods of time) it is precisely these short-duration expression windows, which are being analyzed, that the individual cell ‘sees’ and it is over these expression windows that

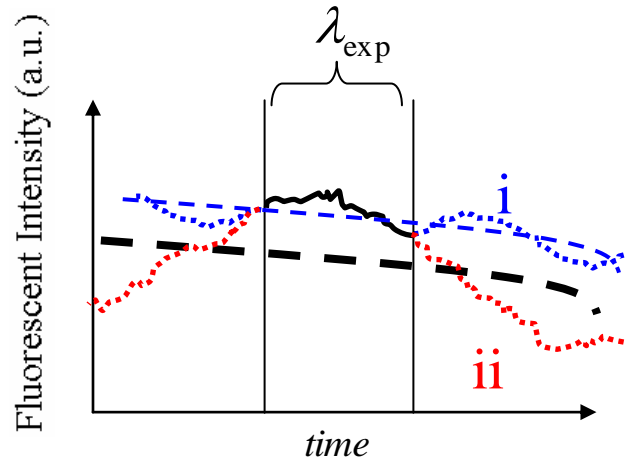


Figure 2.20 Baseline expression shifts are indistinguishable from low frequency fluctuations. Distinguishing between baseline expression level shifts and long correlation fluctuations in a limited imaging window becomes very difficult. A single cell expression level is measured over an experimental imaging window λ_{exp} (solid black) and found to exist above the deterministic population trend $A(t)$, (dashed black line). It is difficult to determine if the signal (dotted blue, (i)) is fluctuating quickly about a possible baseline shift of the deterministic trend (dashed blue line), or is a segment sampling of a low frequency fluctuation (dotted red line (ii)) fluctuating about the true deterministic trend. Both fluctuation trajectories (i and ii) are possible and cannot be discerned from the limited duration observation.

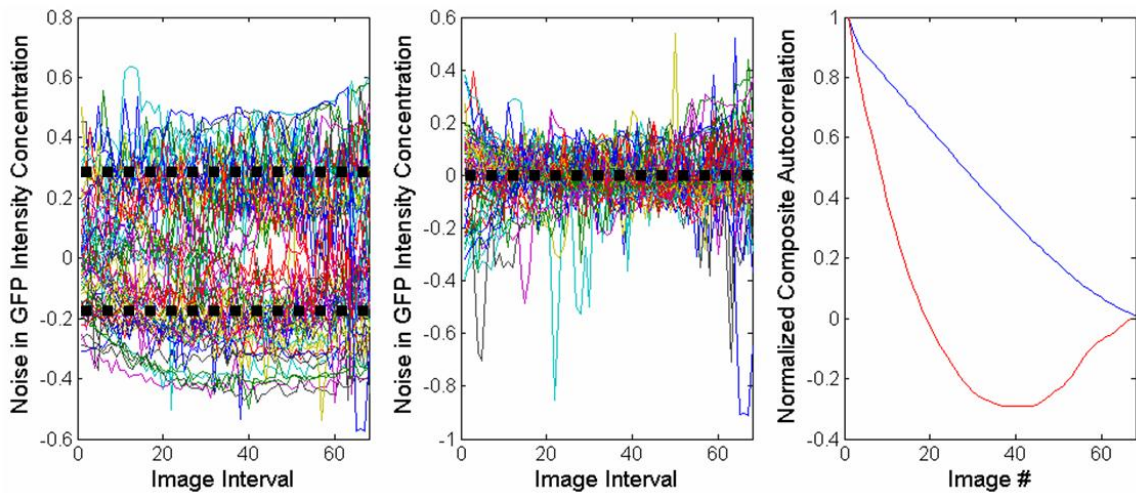


Figure 2.21 High frequency noise processing. Individual mean suppression of each noise trajectory high-pass filters the noise and results in HF-autocorrelation functions. Non-HF processing results in erroneously long correlations for non-mixed slow growing cell populations (left and blue ACF at right). HF-processing resolves this by focusing on HF-correlations and yields an ACF with features (middle and red ACF at right).

phenotypic dynamics and decisions take place. That is, the multiple-step noise processing algorithm provides an *in silico* tool with a biological view that is genuinely relevant *in vivo*.

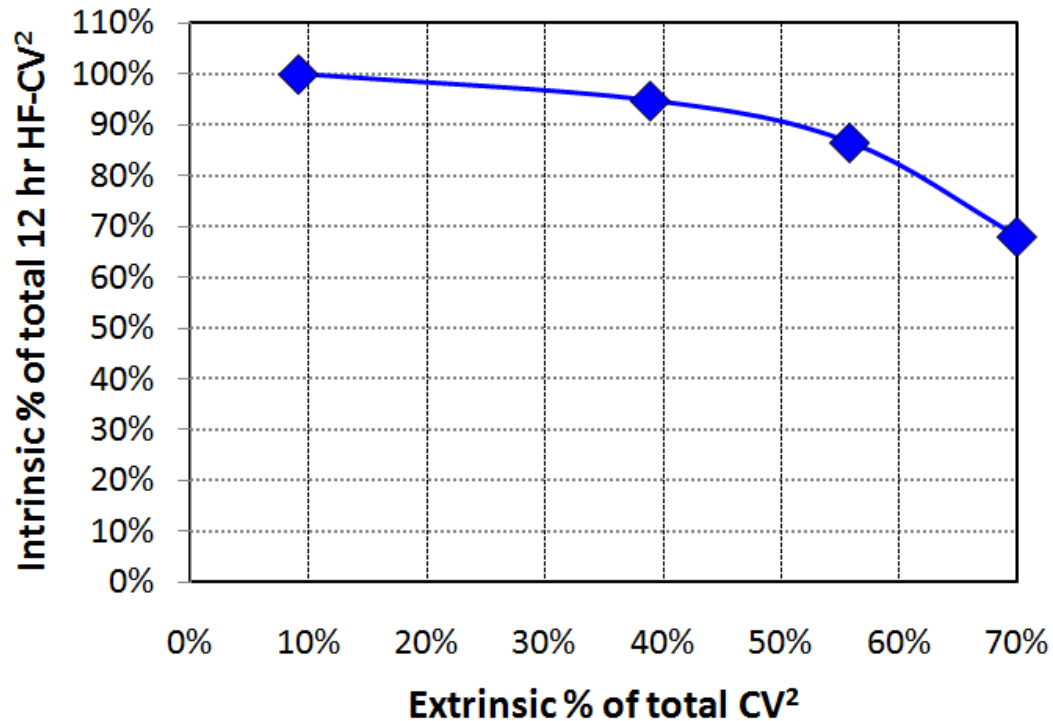


Figure 2.22 HF-processing focuses on intrinsic and filters out extrinsic noise. A simulation of intrinsic/extrinsic noise was implemented to estimate how much noise magnitude is filtered out or emphasized with a 12 hour HF-noise processing. For a large range of extrinsic noise contribution intrinsic noise contribution is enhanced ~1.1-2.3 times of the total noise while extrinsic noise is de-emphasized (filtered) down (e.g. 55% extrinsic of total noise filters down to ~15% of total noise, 40% down to 5%, etc.).

CHAPTER 3: The Coupling of Gene Circuit and Noise Structures

It is foundational to both hypotheses of this dissertation that noise in gene expression is not just a byproduct of limited resources, but can be a key functional component of the system. The functional components in a cell are actively regulated to achieve desired function, and the same should be true for noise. This chapter will explore how noise can be regulated in genetic circuits by demonstrating that noise structure is directly coupled to gene circuit structure.

3.1 Gene circuit structure

The previous chapter dealt with fundamentals and presented noise sources, structure, and analysis in detail. Gene circuit structure refers to the details of the gene expression process as well as the architectures of importance. This includes structural features of the promoter, protein-DNA interactions, protein-protein interactions, regulatory relationships and kinetic parameters of the gene expression. When suitable, gene circuit structure is simplified as much as possible by aggregating processes and rate parameters for secondary processes that lightly affects the gene expression behavior (e.g. GFP protein maturation has many relatively fast steps that are often ignored or lumped with the translation rate).

If \vec{G}_s is a vector that represents the gene circuit architecture and \vec{k}_s is a vector that represents the kinetic rate parameters of the gene circuit, the hypothesis of this chapter may be written as

$$\vec{N}_s = f(\vec{G}_s, \vec{k}_s) \quad (3.1)$$

where \vec{N}_s is a vector that represents the structure of the noise (here called the noise vector). The noise vector may be defined in many different ways (see next chapter), but for the purposes of this chapter, it will have two components: (1) noise magnitude as defined by variance, CV, or CV²; and (2) correlation as defined by half correlation time

($\tau_{1/2}$), where $\tau_{1/2}$ is the value of τ where the autocorrelation function, $\Phi(\tau)$, drops to half of its zero-lag value. The following sections will explore the relationship between gene circuit and noise structure for constitutive gene circuits[9, 46, 70], autoregulated gene circuits [9, 15, 70, 72], and transcriptional regulation [10].

3.2 Open loop constitutive gene circuit

The time dependent chemical Langevin equations representing a simple transcription-translation circuit are:

$$\begin{aligned}\frac{dr}{dt} &= -(\gamma_R + \delta)r + \alpha_R(t) + \eta_R(t) \\ \frac{dp}{dt} &= -(\gamma_P + \delta)p + k_P r + \eta_P(t)\end{aligned}\tag{3.2}$$

where r and p are mRNA and protein concentrations, γ_r and γ_p are their respective decay rate constants, and δ is the dilution rate. α_R is the transcription rate, k_p is the translation rate and η_R and η_P are random variables that represent the shot noise of the synthesis, decay, and dilution of mRNA and protein, respectively.

The frequency domain transfer functions from each of the two noise sources to the protein concentration are found by Fourier transform and solution of these equations to yield [9]:

$$\begin{aligned}H_R(f) &= \frac{b}{\gamma_P + \delta} \frac{1}{\left(1 + i \frac{f}{f_{mRNA}}\right) \left(1 + i \frac{f}{f_{protein}}\right)} \\ H_P(f) &= \left(\frac{1}{\gamma_P + \delta}\right)^2 \frac{1}{\left(1 + \left(\frac{f}{f_{protein}}\right)^2\right)}\end{aligned}\tag{3.3}$$

where the 2-pole frequencies are mRNA and protein decay/dilution ($f_{mRNA} = \frac{\gamma_R}{2\pi}$) and

($f_{protein} = \frac{\gamma_P + \delta}{2\pi}$), and b is the translational burst rate defined as the average number of

proteins produced from each mRNA transcript. Here it is assumed that mRNA decay is much faster than dilution, and dilution may be neglected in the mRNA pole. Furthermore, mRNA decay is often much faster than protein decay ($\gamma_R \gg \gamma_p$), allowing the mRNA pole to be neglected. In this case, the noise bandwidth approximation, where all the noise is assumed to be uniformly distributed in a frequency band between 0 and Δf_N , may be used and [9]:

$$\Delta f_N \approx \frac{\pi}{2} f_{protein} = \frac{\gamma_p + \delta}{4}. \quad (3.4)$$

This result shows that the dominant time constant found in the noise in constitutive, open-loop, gene expression is defined by the protein degradation and dilution rates (Fig. 3.1). Using these simplifications, the protein population noise variance is [9]:

$$\sigma_p^2 \approx (H_R^2(0)S_{RR} + H_P^2(0)S_{PP})\Delta f_N = \langle p \rangle (1+b) \quad (3.5)$$

where S_{RR} and S_{PP} are single-sided power spectral densities (PSDs) for mRNA and protein noise sources [9] and $\langle p \rangle$ is the mean of the protein population, and

$$CV^2 = \frac{\sigma_p^2}{\langle p \rangle^2} = \frac{(1+b)}{\langle p \rangle}. \quad (3.6)$$

Therefore, for constitutive gene expression, noise magnitude is controlled by the translation burst parameter b , and noise correlation is controlled by protein decay and dilution rates.

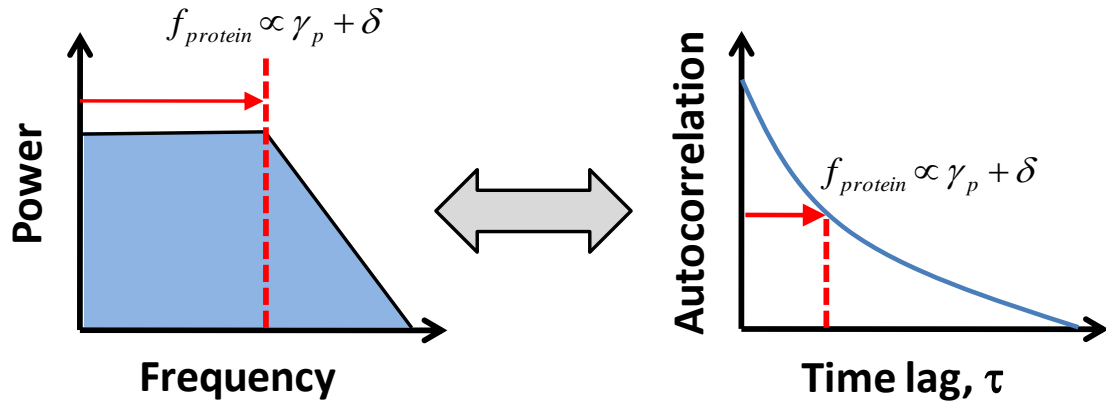


Figure 3.1 Time-domain constants of constitutive gene expression. (left) noise power spectral density (PSD) of a transcription-translation circuit is dominated by protein dilution and decay (assuming fast mRNA decay), and (right) the autocorrelation function (ACF) can be described by a single decaying exponent with a time-constant dominated by protein dilution and decay. Moving between the frequency and time domains occurs by using the Fourier transform.

3.2.1 Experimental investigation of constitutive gene expression

To test the predictions above, *Escherichia coli* (*E.coli*) is a convenient model system. *E.coli* is a prokaryote, which is a simple cell type that does not have defined internal sub-cellular structures such as a nucleus. Transcription and translation occur in parallel in the cell cytoplasm. In addition, *E.coli* is fast growing and the cell doubling time is easily modulated by environmental temperature. Finally, the experimental techniques for the genetic manipulations of these cells are well developed.

Using a high copy number plasmid (see Appendix), a gene circuit that constitutively expresses variants of GFP that possessed two different decay rates (pGFP_{asv} with a 110 min half-life and pGFP_{aaV} with a 60 min half-life (see Appendix)) was inserted into *E. coli* TOP10 cells, and the noise structure in the GFP populations in growing cultures of these cells was measured using the set-up shown in Figure 3.2 [70]. After the necessary cell sample preparations, a dose of cell culture was dispensed onto a layer of agarose gel suitable for growth. The agar was on a glass slide, the *E.coli* cells on top of the agar, followed by a glass cover slip, emersion oil, and the upright laser scanning confocal microscope (Leica) objective. To control cell growth rate, an external heating lamp was used with a thermocouple probe (Hanna Instruments) in the layer of agar. Due to the heat source and dry environment, the agar layer lost volume throughout a typical ~4-8 hour experiment. Therefore, the imaged and tracked cell population literally receded vertically and required continuous focal adjustments at each image acquisition.

In the experiments, the average GFP fluorescence, which corresponds to the concentration of mature GFP protein (see Appendix), was measured in individual cells in growing cultures for 4–8 h periods [27, 45] (Fig. 3.3b). Cells were healthy and were in a constant exponential growth phase throughout the experiment. Imaging (interval of 5 minutes) was performed through multiple generations of cell division (Fig. 3.3b), and the noise in gene expression was found as the difference between the fluorescence of individual cells and the population mean determined at each measurement time (Fig. 3.3c and Chapter 2). Individual noise traces (trajectories) were constructed by combining sequential noise traces of cells throughout lines of descent (Fig. 3.3b and Chapter 2). In

Figure 3.3 the population started with an initial ~4-8 parent cells and ended with over 150 daughter cells.

Normalized autocorrelation functions of noise in fluorescence were estimated for individual trajectories ($\Phi_m(\tau)$) and composites ($\Phi_c(\tau)$) of all tracked trajectories in each cell population for the duration of the experiment (Fig. 3.3d), where the composite autocorrelation (CAC) is the lag-dependent average of all individual ACFs before normalization. The composite autocorrelation functions was representative of the dynamics and underlying process of the whole population, while individual trajectory autocorrelation functions provided insights into gene circuit structure or function as described below. Figure 3.3 shows a typical experiment.

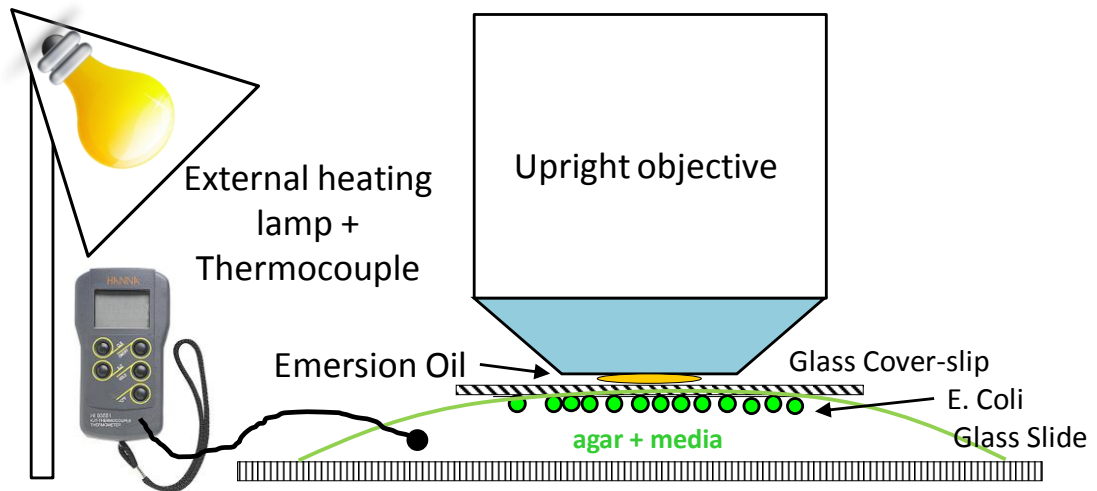


Figure 3.2 Diagram of *E. Coli* Experiment Setup. For upright fluorescence imaging with a confocal laser scanning microscope the above sample setup was used. *E. Coli* cell culture is dispensed and grown on an agar gel including the necessary culture medium. The agar layer is on top of a glass slide, the cells on the agar, covered by a glass cover-slip and finally imaged with an objective in emersion oil. To control environmental temperature and cell growth a heating lamp and thermocouple were used. Not depicted in this diagram are the laser scanning confocal microscope and computer console for microscopy control and fluorescence image storage.

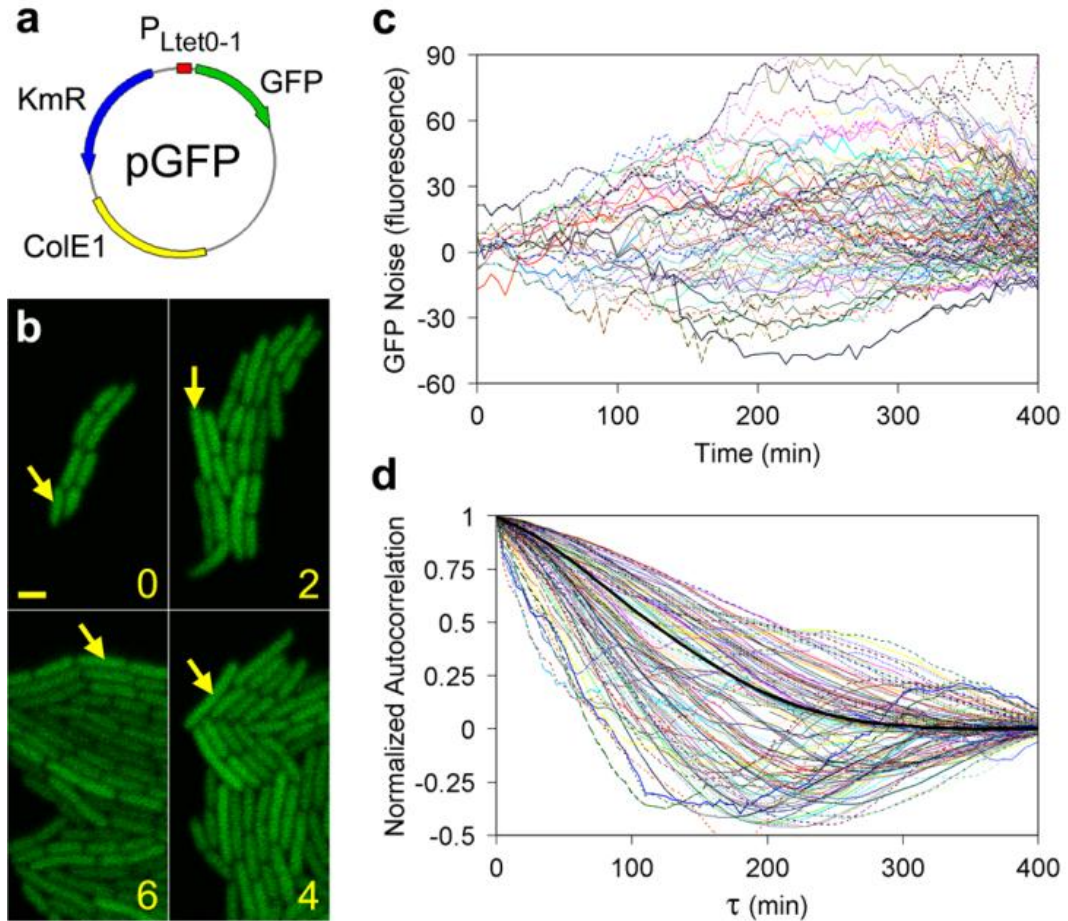


Figure 3.3 Noise frequency range detection with fluorescence microscopy. (a) a DNA plasmid containing pGFP_{asv} (110-min half-life) which is constitutively expressed. (b) 2 hour snapshots of a time-lapse experiment of a growing pGFP_{asv} cell colony. The yellow arrow depicts tracking of an individual cell through cell divisions and stages of growth. Scale = 2 μ m. (c) noise in GFP concentration, (d) normalized autocorrelation function of all noise trajectories in (c). The population composite autocorrelation is shown in black.

3.2.1.1 Single-cell noise frequency range distributions

In this work, correlation was characterized using the noise frequency range (F_N), which was defined as

$$F_N = \frac{1}{\tau_{1/2}} \quad (3.7)$$

Slower fluctuations remain correlated over longer periods of time and therefore have lower values of F_N .

Histograms of noise frequency ranges extracted from the individual trajectory autocorrelation functions are shown in Fig. 3.4a, b. These results were compared with noise frequency range distributions from exact stochastic simulation of constitutive gene expression [52, 73] accounting for intrinsic and extrinsic noise with the relevant protein dilution and decay rates. Simulations of 500 separate experiments with similar number of cells as those collected experimentally yielded an F_N distribution representing the probability of finding a given noise frequency range from a randomly selected cell trajectory in the cell colony.

3.2.1.2 Modulation of protein dilution and decay

Protein dilution and decay were varied to modulate their resulting noise frequency ranges. Protein dilution rate was varied by controlling the sample temperature with an external heating lamp. Figure 3.4a shows the frequency range shift observed for the circuit with 110 min GFP half-life when the cell doubling changed from 30 min to 90 min. The slower dilution rate shifts the noise frequency range to lower values.

Next, modulation of protein degradation rate was performed using the two different plasmids; pGFPasv produces a GFP variant that has a half-life of ~110 minutes and pGFPaav, which produces a GFP variant with a reduced half-life of ~60 minutes. Figure 3.4b shows a clear frequency range shift to higher values with the increased protein decay rate (pGFPaav).

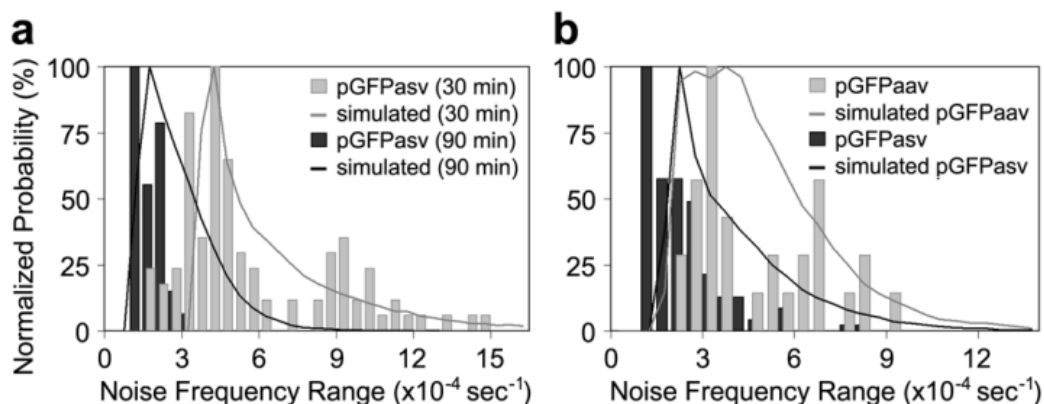


Figure 3.4 Effects of cell doubling time and protein half-life on noise frequency range. Measured distributions are shown as vertical bars and simulated distributions as solid lines. **a**, Shift in noise frequency range for the pGFPasv circuit as doubling time increased from ~30 min (100 trajectories; $T = 32\text{ }^{\circ}\text{C}$) to ~90 min (120 trajectories; $T = 22\text{ }^{\circ}\text{C}$). **b**, Shift in noise frequency range as protein decay time decreases from 110 min (pGFPasv; 59-min doubling time; 154 trajectories; $T = 26\text{ }^{\circ}\text{C}$) to 60 min (pGFPaav; 56-min doubling time; 33 trajectories; $T = 26\text{ }^{\circ}\text{C}$).

The experimental F_N distributions show a distinct clustering or bimodal signature at mid- to high frequency range values (grey distributions in figure 3.4). This signature was also observed in individually simulated experiments with similar numbers of cells (data not shown), but disappeared in the average distribution of all 500 simulated experiments (solid lines in figure 3.4). This suggests that this hint of bimodality in the experimental frequency distributions are simply an artifact of limited cell statistics.

As no significant variation in noise frequency range was found between the simulated model and the experimental results, additional non-constitutive noise modulating mechanisms such as transcriptional control via protein-DNA binding [10, 74], extrinsic noise [27, 45], protein dimerization [75], and GFP maturation were ruled out. All of these mechanisms should lower noise frequency range, so for example if the mRNA decay rate was lower or the oxidation step of protein maturation (with a literature range of 18-80 minutes) was long, these would have modulated the frequency range distributions. For this reason the oxidation step in the experiment was most likely on the lower side of the range and was too fast to significantly modulate the noise frequencies.

3.3 Autoregulated Gene Circuits in Nature

Autoregulation (AR) is a genetic architecture in which a protein controls the level of its own expression by activation (positive autoregulation, or +AR) or repression (negative autoregulation, or -AR) of its own promoter. Negative autoregulation is a very common motif found in gene circuits (e.g. ~40% of the roughly ~300 *E.coli* transcription factors are negatively autoregulated [8, 76]). In addition, a negatively autoregulated architecture is the core molecular mechanism behind circadian rhythms with a period of around 24 hours and found in nearly all living organisms ranging from cyanobacteria, plants and insects, to mammals [77, 78]. As such, modeling and understanding autoregulated system behavior may have important implications for global system-wide regulation, expression dynamics, and biochemical processing.

Other autoregulated gene circuits of interest are found in viruses. Some viruses, e.g. human immunodeficiency virus type 1 (HIV-1), are among the most remarkably compact and highly functional nanoscale systems in nature. Its 9 genes control all the viral stages, including infection, reverse transcription, integration, replication, and viral particle packaging. Interestingly, HIV is among a large group of viruses that encode a transactivation loop, or positive autoregulation architecture (Fig. 3.5) such as Herpes simplex virus 1 (HSV-1), Epstein-Barr virus (EBV), and Cytomegalovirus (CMV). In addition to this +AR architecture, animal viruses such as HIV and CMV are known to have an overlaid -AR loop that operates at a time after the +AR has completed its function (Fig. 3.6). So here again is an important biological nanoscale system whose transactivated architecture (+AR) plays a role in its function, and whose +AR structure-function relationship modeling may have great importance. In addition, the combined implications of isolated +/- AR systems may ultimately enable the understanding of overlaid +/- FB systems. The two studies reported in this chapter ([15, 70]) each have their own independent and unique findings. They also provide some of the preliminary foundations in the modeling and experimentation of +/-AR to pursue relevant and complicated layered autoregulatory motifs.

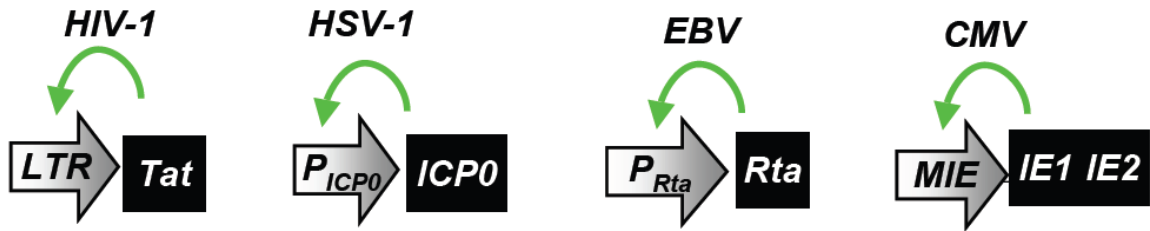


Figure 3.5 Positive autoregulation in well-known viruses. From left two right are the human immunodeficiency virus 1 (HIV-1), where Tat +AR the LTR promoter, herpes simplex virus 1 (HSV-1), where ICP0 +AR the ICP0 promoter, Epstein-Barr virus (EBV), where Rta +AR the Rta promoter, and finally Cytomegalovirus (CMV), where IE1 +AR the MIE promoter [Figure adapted from [79]].

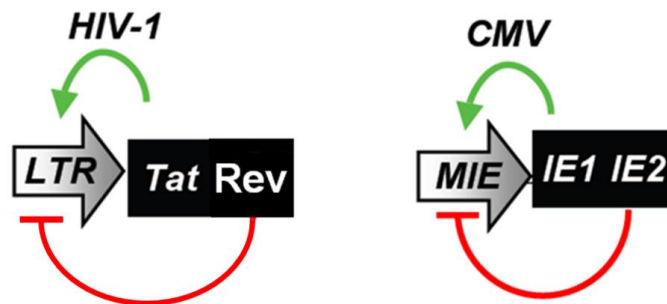


Figure 3.6 Overlaid positive and negative feedback loops in animal viruses. **(left)** In HIV-1 Tat activates viral transcription elongation (+AR) while Rev induces nuclear export of viral RNA's and effectively acts as -AR. **(right)** CMV is +AR by immediate-early protein IE1 and negatively by immediate-early protein IE2.

3.3.1 Analysis of an Autoregulated Gene Circuit

To derive the noise structure of an autoregulated system, the analysis applied to constitutive expression was extended by Simpson *et al.* in 2003 [9] and made use of concepts developed for electronic feedback amplifiers. In particular, the concept of loop transmission, T the transfer function around the loop, was applied to autoregulated gene circuits. T is the frequency dependent first derivative of the regulation strength and can be thought of as a measure of resistance of the circuit to deviation from a steady state.

Simpson *et al.*, (2003) [9] analyzed a $-AR$ system and the expression arrived at for the noise bandwidth was:

$$\Delta f_N \approx (1-T(0))f_{protein} = (1-T(0))\frac{\gamma_p + \delta}{4}. \quad (3.8)$$

Comparison of the results for negative autoregulation with the constitutive frequency domain results in an increase in noise bandwidth by $(1-T(0))$ (Fig. 3.7). (Note: $T(0)$ is negative for $-AR$ and positive for $+AR$). The increase in bandwidth occurs by shifting some of the noise to higher frequencies where it may subsequently be filtered out by downstream circuit elements [9].

In addition, with the assumption that autoregulator-promoter binding and unbinding is fast, the variance decreases as,

$$\sigma_{P_regulated}^2 \approx \frac{\sigma_{P_non-regulated}^2}{(1-T(0))} = \frac{\langle p \rangle (1+b)}{(1-T(0))} \quad (3.9)$$

and

$$\left(\frac{\sigma_{P_reg}^2}{\langle P_{reg} \rangle} \right) \approx \frac{\left(\frac{\sigma_{P_non-regulated}^2}{\langle p \rangle} \right)}{(1-T(0))} = \frac{(1+b)}{(1-T(0))}. \quad (3.10)$$

Here the noise magnitude is decreased by a factor of $1/(1-T(0))$. The above effects of negative autoregulation on the non-regulated noise power spectral density are depicted in figure 3.7.

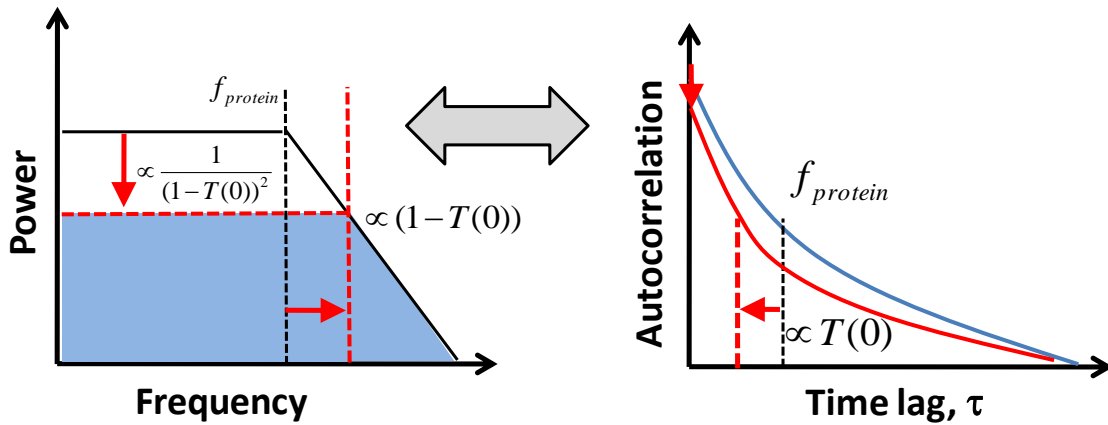


Figure 3.7 Negative autoregulation increases noise bandwidth. (left) PSD of a negatively autoregulated gene circuit is dominated by the strength of regulation ($T(0)$) and reduces noise magnitude by $(1+|T(0)|)^{-2}$ and increases noise frequency range by an amount of $(1+|T(0)|)$. (right) ACF half-correlation time is dominated by the strength of regulation ($T(0)$). Moving between the frequency and time domains occurs by using the Fourier transform.

3.3.2 Experimental investigation of negative autoregulation

To create a negatively autoregulated gene circuit, the gene for the protein TetR was inserted upstream of *gfp*, creating a transcriptional fusion (pTetR–GFP_{asv}; Fig. 3.8a). This circuit is negatively autoregulated, as its expression is repressed by TetR binding to operator sites in the promoter [80]. TetR binding in the promoter region inhibits transcription by blocking the binding of RNA polymerase. TetR binding to the promoter is relieved by Anhydrotetracycline (ATc), which may be added to the growth medium to modify repression and change feedback strength (Fig. 3.8b). By binding reversibly to TetR, ATc titrates out free TetR (Fig. 3.9b) and modulates the strength of regulation. Without any ATc, repression is so strong in the cell that GFP intensity does not exceed cellular autofluorescence levels (Fig. 3.8b).

Using the experimental set-up of Fig. 3.2, noise frequency range distributions and population composite noise frequency (based on the composite population autocorrelation function) were quantified for pTetR-GFP_{asv} grown in media with 100 ng ml⁻¹ of ATc. Composite F_N of the –AR circuit exceeded F_N values of the constitutive pGFP_{asv} + 100 ng ml⁻¹ of ATc by ~2-3 fold (Figs. 3.8d and 3.9a). Negative autoregulation had a distinct signature both on the composite and distribution of noise frequency range. Negative autoregulation-mediated noise remodeling increased the noise frequency range and modified the single cell noise frequency range distribution such that they had a more Gaussian profile (Fig. 3.8d). Autoregulation frequency response is limited by protein decay and dilution, and therefore has a larger effect on slower fluctuations than on faster fluctuations. Noise trajectories that would have clustered at the lower end of the frequency range distribution in unregulated cells were pushed to higher values by negative autoregulation, while those in the higher frequency tail of the distribution were only weakly affected (Fig. 3.8e). This results in noise frequency range distributions closer to normal distributions (Fig. 3.8d). The frequency shift and the change in distribution shape are indicative of the presence of negative autoregulation and presents a –AR single cell distribution signature that was not previously reported.

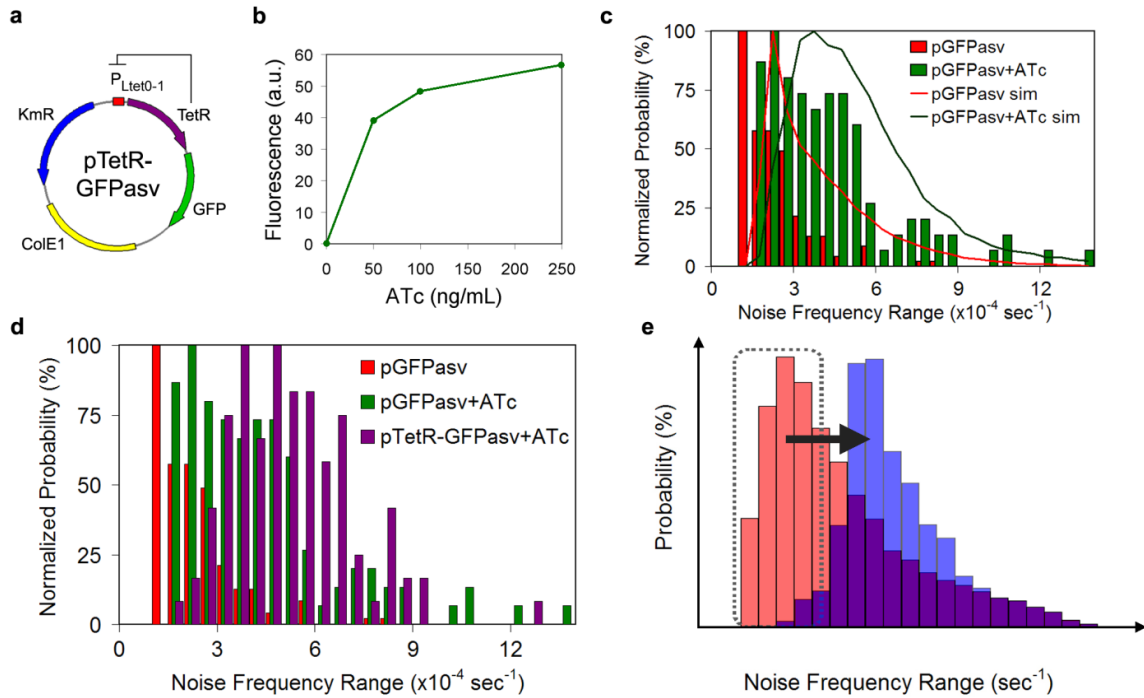


Figure 3.8 Effect of negative autoregulation on noise frequency range. **a**, pTetR–GFPasv negatively autoregulated gene circuit with **b**, repression strength modulated by ATc. **c**, Effect of ATc on the noise frequency range of the unregulated pGFPasv circuit (doubling time ~60 min; 154 trajectories without ATc; 114 trajectories with ATc). sim., simulated. **d**, Negative autoregulation-mediated shift of noise frequency range (doubling time ~60 min; pGFPasv: 154 trajectories without ATc, 114 trajectories with ATc; pTetR–GFPasv: 114 trajectories). **e**, Model of the shift of frequency range distribution shape due to negative feedback. The red bars represent an unregulated circuit distribution; blue bars represent distribution for the circuit with negative autoregulation. The dashed box and arrow show the shift of the low-frequency trajectories to the center of the distribution while the higher frequency trajectories are unaffected. Fluorescence in **b** is given in arbitrary units (a.u.).

3.3.2.1 Detection of various strengths of negative autoregulation

The magnitude of –AR frequency deviation from the constitutively expressed colonies was an indication of the strength of regulation and was a strong function of cell doubling time (Fig. 3.9a). Theoretical analysis predicts that negative autoregulation increases the noise frequency range such that[9]:

$$F_{N_autoreg} = (1 + |T|)F_{N_unreg} \quad (3.11)$$

The measured noise frequency range can be used to determine the strength of regulation using the above equation. Lower feedback strengths were detected at short and long cell doubling times, with an increased strength at intermediate doubling times (Fig. 3.9). The gene circuit model of Figure 3.9b describes the regulation strength as the product of (1) $d[\text{TetR}_2]/d\alpha$ (the change in free (not bound to ATc and capable of repression) TetR dimer concentration in response to changes in transcription rate (α)), and (2) $d\alpha/d[\text{TetR}_2]$ (the change of transcription rate in response to changes in free TetR dimer population). Imaged populations of cells with high growth rates consistently showed lower average fluorescence due to the reduction of GFP and TetR concentrations by rapid dilution.

At fast cellular growth rates, the strength of feedback was low because the small population of TetR molecules was mostly bound to ATc and unavailable for repression. At slow growth rates, there was an overabundance of TetR in the cell but regulation strength was low because the repression curve was saturated [10, 74] (Fig. 3.9). At intermediate cell growth rates, the population of free TetR dimer was high enough that it was not mostly bound to ATc, yet low enough that the repression curve had not saturated. As a result, strong negative feedback strength was found at the intermediate cell growth rates (Fig. 3.9c).

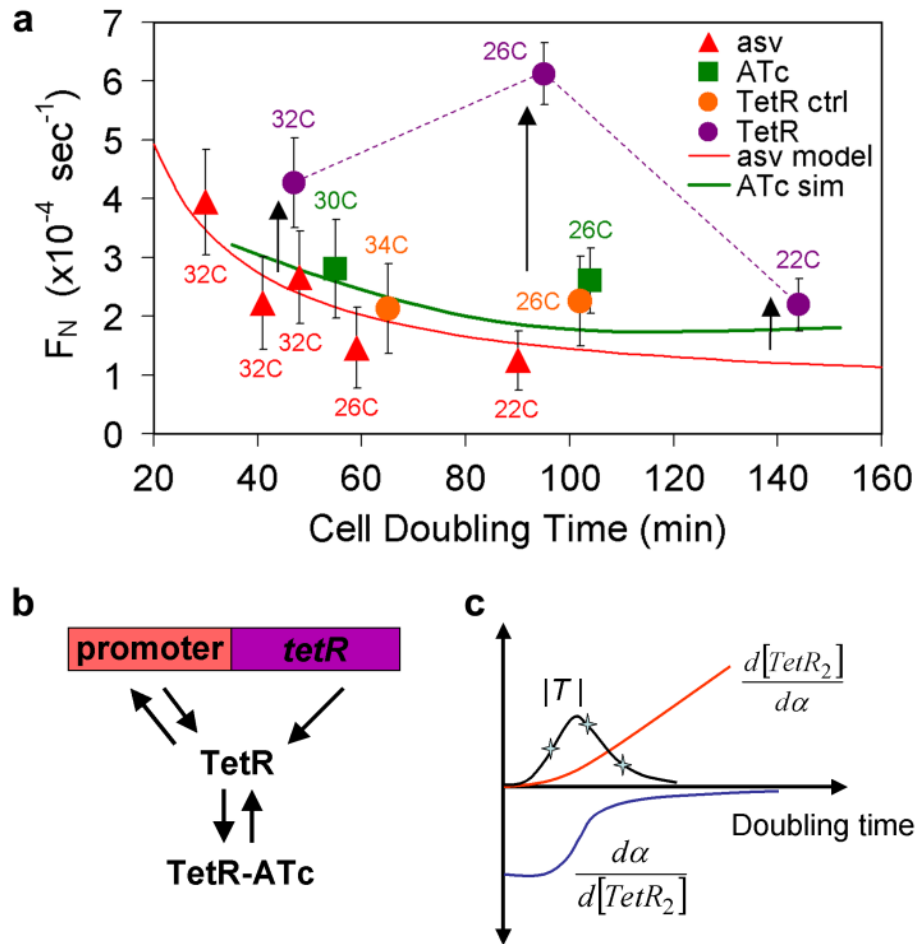


Figure 3.9 Regulation strength modulation of noise frequency range. **a**, Noise frequency range versus doubling time. Measured points are shown with $\pm 1\sigma$ error bars estimated from simulation. The red line is the analytical curve for the pGFP_{asv} circuit, and was found from the analytical expression for the autocorrelation function [70]. The green line is the simulated curve for pGFP_{asv} + 100 ng ml⁻¹ ATc and was found from the simulation of the pGFP_{asv} circuit with ATc–ribosome binding. Vertical black arrows represent regulation strength determined by the shift of the noise frequency range. The temperature (in °C) of each experiment is indicated by each data point. The TetR data points are for the circuit with autoregulated *tetR* expression, while the TetR ctrl data points are for the circuit with constitutive *tetR* expression. **b**, **c**, Regulation of the pTetR–GFP_{asv} circuit. The red curve shows the concentration of free TetR dimer ($[TetR_2]$) variation with transcription rate (α); the blue curve shows transcription rate variation with $[TetR_2]$; and the black curve shows net regulation strength. The stars illustrate points on the regulation curve similar to the TetR data in **a**.

3.3.2.2 Modulation of extrinsic noise with a drug – control experiment

As a control experiment, the constitutively expressed pGFPasv circuit was exposed to ATc to confirm that the results reported above were due to feedback rather than some ATc mediated mechanism. The noise frequency range of pGFPasv + 100 ng ml⁻¹ of ATc was measured, and ATc did produce a significant broadening of the distribution (Fig. 3.8c), which led to a small increase in the composite noise frequency range (Fig. 3.9a). This suggested that ATc either modified the processing of the noise or the nature of the noise sources. A stochastic simulation model of ATc inhibition of translation from ribosome-ATc heterodimer formation (a known effect of this drug [81-83]) yielded similar frequency range distributions and composite values (Fig. 3.8c and 3.9).

The stochastic model of GFP expression with ATc-ribosome inhibition is summarized in table 3.1.

Table 3.1 Stochastic simulation model of ATc-ribosome inhibition

Reaction	Rate
1. $R \rightarrow R + ribo$	k_1
2. $ribo \rightarrow ribo + GFP$	$b_{noise} * \delta$
3. $ribo \rightarrow *$	δ
4. $GFP \rightarrow *$	$\delta + \gamma$
5. $ribo \rightarrow ribo-ATc$	k_f
6. $ribo-ATc \rightarrow ribo$	k_r

Reactions 1 and 3 represent extrinsic noise that is filtered by the dilution rate. F_N is independent of the value of k_1 and the dilution rate δ was determined using the average

measured cell growth rates (doubling time). Reaction 2 represents the translation of mRNA whose stochastic variation is an intrinsic noise component that was modeled in the translation noise component. b_{noise} reflected a mechanism by which changes in temperature would affect the relative weighting of extrinsic and intrinsic noise and the best fit to the experimental measurements was achieved using a value of $b_{noise} \approx 4$ (Fig. 3.9, green curve). The mRNA decay rate was neglected as it is usually short compared to the dilution rate. Reaction 4 represents dilution and decay of GFP. Reactions 5 and 6 represent the effect of ribosome inactivation by ATc at a constant concentration. A diagram of this simulation is depicted in figure 3.10. Inhibition of translation in ATc experiments was modeled by assuming that the fraction of bound ribosomes was proportional to the fractional reduction in growth rate observed experimentally from the addition of ATc at constant temperature (Fig. 3.9a).

The GFP noise frequency range was whitened by the ribosomal extrinsic noise source. ATc reversibly binds ribosomes that are both on the mRNAs thereby hindering translation, as well as freely diffuse ribosomes in the cytoplasm. It is also worth noting that ATc exposed cells have reduced growth rates (Figure 3.9a) as the ribosome is one of the most abundant protein

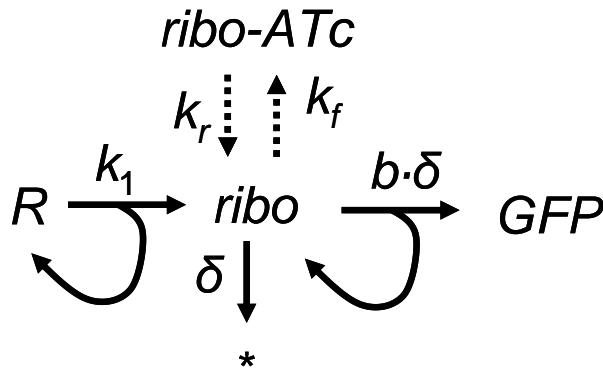


Figure 3.10 A model of ATc inhibition of translation. In this stochastic model of the gene circuit, all extrinsic noise is modeled by the stochastic production of the ribo species, while all intrinsic noise is modeled by the stochastic production of GFP. The proper weighting between extrinsic and intrinsic noise is achieved by varying b (or b_{noise} in the table and text). The effect of ATc on the extrinsic noise term is modeled by the reversible reaction denoted with the dashed arrows.

complexes in the cell and stalling of system-wide translation reduces growth rate of the cell. As seen in Figure 3.8c, the simulated modulation of translational extrinsic noise exhibited a distribution which showed the noted peak shift and broadening.

In addition to the reversible binding of ribosomes, several alternative mechanisms to explain the increase in F_N upon addition of ATc to pGFP_{asv} cells were considered and excluded. The mechanisms include (1) an increase in the protein decay/dilution rate, and (2) negative autoregulation, which would both yield a high frequency range shift observed with the addition of ATc (Figures 3.8 and 3.9), but no literature supported evidence was found for such mechanisms of ATc.

3.3.2.3 Effect of a non-regulated repressor – control experiment

A second control experiment was performed to ensure that the shift was due to TetR mediated –AR, and not just to the presence of TetR. To do this, a non-regulated plasmid, pGFP_{asv} from Figure 3.3a, was inserted into an *E. coli* cell line that had a chromosomal copy of *tetR* constitutively expressed from the strong P_{N25} promoter. This circuit does not have –AR as TetR has no effect on the P_{N25} promoter, and did not exhibit an increase in F_N (Fig. 3.9a, orange circles versus purple circles). The single cell frequency range distribution showed that the control circuit did not exhibit the characteristic shift in the shape of the noise frequency range distribution observed earlier with –AR (Fig. 3.11).

3.3.2.4 Summary of negative autoregulation effects on noise

Compared to constitutive expression, –AR both reduces the magnitude of the noise and shifts the remaining noise to higher frequencies. The high-frequency noise shift may have biological significance, as the faster fluctuations are more easily filtered by downstream gene circuits (e.g. in a genetic cascade) [9]. Transcription factors are the information carriers of cell, and it may be important to maintain high fidelity in these signals. Accordingly, the prevalence of –AR control of the expression of these transcription factors in *E. coli* may be explained by the noise filtering effect of this circuit architecture. Finally, the above study may increase our understanding and engineering

efforts of circadian rhythms, biological oscillators, and the understanding of mixed and layered feedback systems.

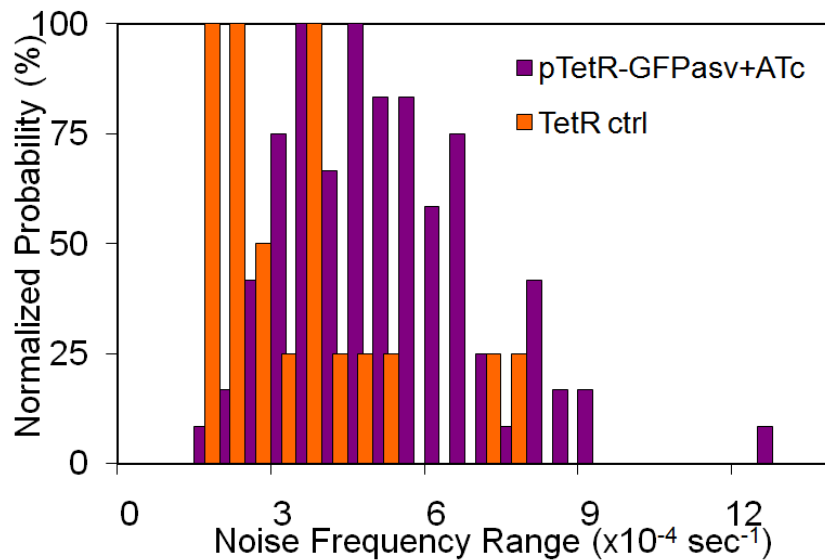


Figure 3.11 Non-regulated repressor control frequency range remains log-normal. The TetR ctrl uses a cell line that constitutively expresses TetR and the non-regulated pGFPasv plasmid. Here the TetR ctrl is not shifted in noise frequency range and doesn't have a normal distribution transition as in the purple negatively autoregulated case. These distributions are for ~60 minute cell doubling times and have added 100 ng/ml ATc.

3.3.3 Experimental investigation of positive autoregulation

Much like $-AR$, $+AR$ plays important roles in the function of many gene circuits [75, 84, 85]. The theoretical analysis of autoregulatory gene circuits described earlier may be applied to $+AR$ as long as the strength of $+AR$ is low enough that the circuit does not become unstable. Therefore, it cannot be applied (without modification) to $+AR$ circuits that latch or oscillate. However, for $+AR$ circuits with modest feedback strength, the only modification to the theory used for $-AR$ is that the loop transmission, T , changes sign from negative to positive. As a result, theory predicts that positive autoregulation increases noise magnitude and decreases noise frequency range into a more regulatory-relevant regime where it may play a role in the function of some genetic switching elements.

The analysis of gene expression noise in both minimal positive feedback circuits and full length (wild-type) transactivated ($+AR$) human immunodeficiency virus type 1 (HIV-1) in human T-cells was pursued in collaboration with the Weinberger Laboratory (University of California, San Diego) [15]. HIV-1 infected CD4+ T lymphocyte human cells can enter two different fates (Fig. 3.12a). Most infections lead to active replication where the cell is hijacked and its cellular resources exploited to produce hundreds of infectious viral pods which will continue to infect additional T-cells after lysing the host cell. In this active mode, the T-cell is destroyed and lysed in ~ 40 hours. The second possible cell fate is proviral latency, a long-lived quiescent state where viral gene expression is turned off [86, 87], but the HIV genetic code remains stably integrated in the host cell genome. Latency occurs at a very low probability compared to the active replication decision, yet it is the main culprit preventing effective HIV eradication from patients [88].

HIV-1 is a remarkable highly functional and compact nanoscale system which codes for all of its functions (i.e. reverse transcription, transport, integration, replication, packaging, etc.) in the expression of only 9 genes (Fig. 3.12a). Among these is a Trans-Activator of Transcription (Tat) protein which up-regulates the expression of all 9 genes, including itself, thereby establishing $+AR$. Tat protein has been shown as essential to

viral active replication and latent reactivation [89-91]. Tat activates the long terminal repeat (LTR) promoter of HIV by enhancing transcriptional elongation via RNA polymerase II hyper-phosphorylation [89, 92, 93]. The LTR promoter has a nucleosome (nuc1) right at its transcriptional start site (TSS) where RNA pol II stalls and waits. De-acetylated Tat releases the stalled RNA pol II. Tat positive feedback drives lytic replication by enhancing its own expression 50- to 100-fold above basal levels in addition to HIV Rev (the essential viral mRNA export factor) and Nef (a viral protein not essential for viral replication)[94].

While it is known that the Tat +AR circuit mediates the decision between active infection and latency, there has been some debate about the exact mechanism. One school of thought has held that the Tat +AR circuit establishes bistability, i.e. is capable of latching into one of two states. However, bistability requires relatively high positive feedback strength, while the Tat +AR circuit has been shown to have relatively low feedback strength [12, 79]. A second proposal holds that the weak positive feedback circuit drives expression pulses that would, in the absence of cell lysis, decay into a monostable latent or “OFF” state. In this model, the decision between active infection and latency is mediated by the duration of this expression pulse: if it is long enough, the active infection pathway is followed; otherwise, the circuit drops into latency and awaits the next stochastic expression burst, which again will lead to active infection if it persists long enough. In this model, the role of the +AR is to lengthen the duration of these noise expression pulses, and thereby increase the probability of active infection.

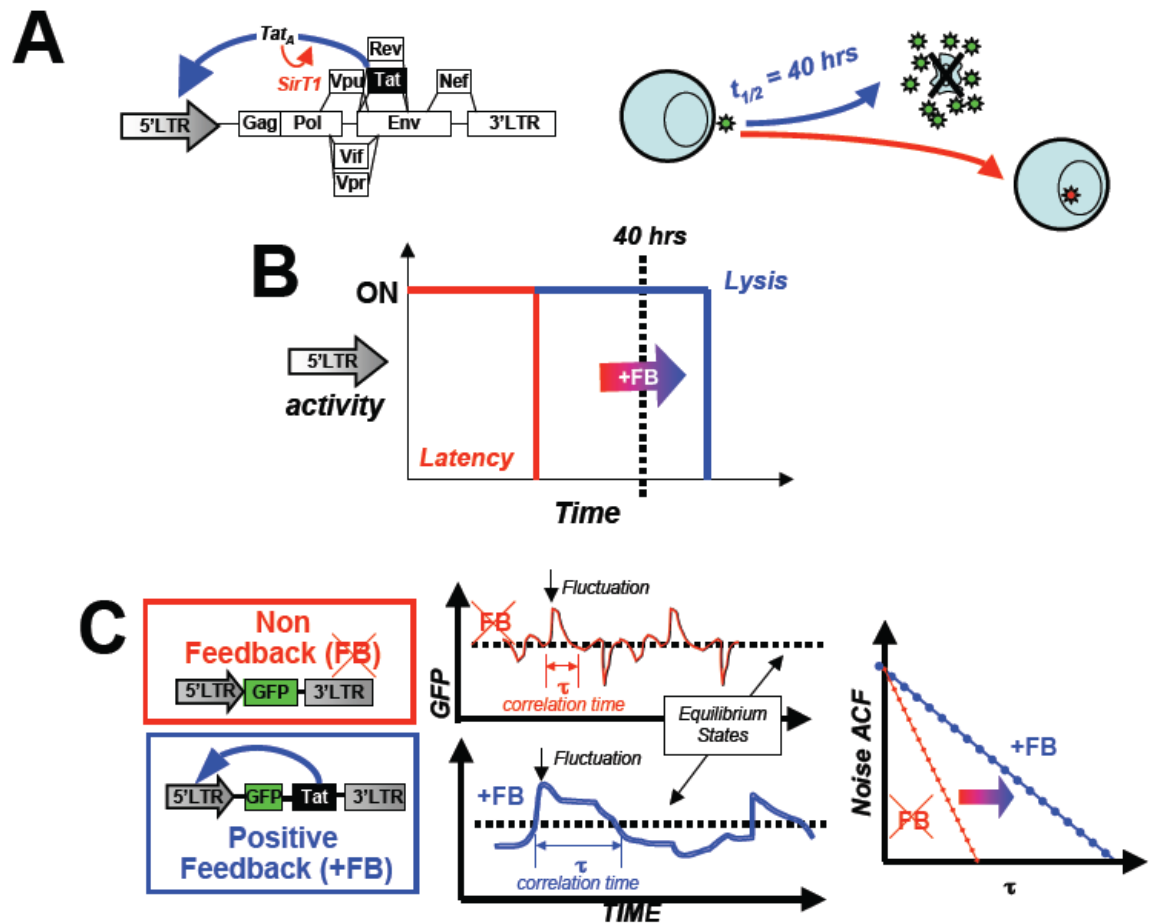


Figure 3.12 Positive-feedback extends the lifetime of gene expression transients. **a**, The HIV-1 genome encodes the Tat positive-feedback circuit. This circuit is comprised of HIV-1 Tat which in its short-lived acetylated form (Tat_A) transactivates the viral promoter within the LTR but is also rapidly deacetylated by human SirT1 [79, 95]. HIV-infected T-cells undergoing active viral replication (i.e. with active Tat positive-feedback) have an average lifetime of ~40hrs [96]. **b**, Expression transients without positive-feedback are short-lived and die out quickly leading to latency (red). But, positive-feedback (in direct proportion to its strength or loop transmission) can extend the duration of gene expression[9] transients thereby favoring lytic replication (blue). **c**. Positive-feedback strength can be directly measured in single-cells by examining fluctuations in gene expression (left and middle) and calculating a fluctuation auto-correlation function (ACF, right). Positive-feedback shifts the ACF decay by a magnitude that correlates directly to the strength of positive-feedback[70].

Using the noise autocorrelation analysis developed in the previous section for investigating negative autoregulation, the strength of Tat positive feedback was directly measured (Fig. 3.12) [9, 70]. Slight modifications in the processing of the noise were needed for analyzing gene expression of a slow growing human cell (e.g. single cell gain coefficients and high frequency noise processing, see Chapter 2). Positive feedback is predicted to extend expression pulse durations (the opposite of $-AR$ shift of figure 3.7). To test this prediction time-series gene expression experiments of minimal non-feedback (LTR-GFP, Long-Terminal Repeat HIV promoter driving GFP) or minimal positive feedback were compared (LTR-GFP-Tat, LTR driving GFP and Tat) (Fig. 3.12c, left). Positive feedback reinforces fluctuations away from the mean, which extends the duration of these fluctuations as compared to those from a non-feedback circuit (Fig. 3.12c, middle). Longer duration fluctuations produce an ACF that decays more slowly (Fig. 3.12c, right), making the ACF width an indicator of positive-feedback strength.

3.3.3.1 Correlation shifts in minimal positively autoregulated gene circuits

A simplified diagram of the experimental setup and a sample fluorescence image are shown in Figure 3.13. The experimental process with the human T-cells is very similar to the previous *E.coli* experiments (Figures 3.2, 3.3, 3.13, and 3.14a). Single-cells were imaged for 12-24 hours at a 10 minute imaging interval, tracked and quantified for their fluorescent intensity signals, and processed for their stochastic component. As the cell doubling time was very long (~ 14 hours), HF-noise processing was required to prevent the calculation of erroneously long autocorrelation times (see Chapter 2 for more details). Composite autocorrelation functions representing the underlying dynamics of the whole cell population were calculated by averaging single cell ACFs. Similar to the previous experiments, feedback strength was quantified by the half-correlation time values, and comparison to a non-FB system. The feedback strength (T) was estimated from the relationship $T \rightarrow 1 - \left(\tau_{1/2_nonFB} / \tau_{1/2_FB} \right)$ where the arrow \rightarrow represents an equality for true ACFs[70] and a mapping operator for high frequency ACFs (here experimental imaging durations were 12 hours). Negative values of T indicate negative

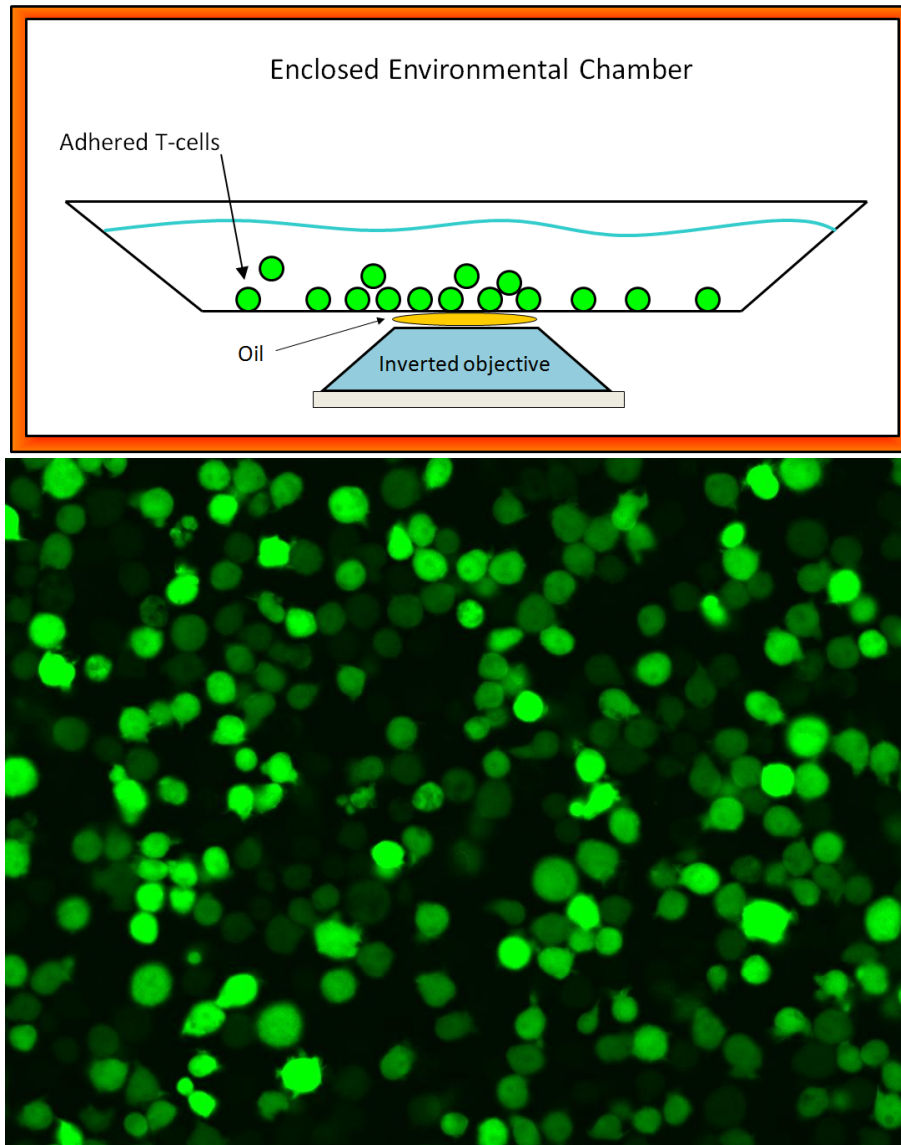


Figure 3.13 Sample setup and fluorescent image of GFP expressing human T-cells. (upper) Shows a schematic depiction of the sample and objective setup. Cells are adhered on a substrate with liquid media on top. They are imaged from below by the inverted microscope. Not shown are the microscope and computer acquisition components. (lower) Sample image of HIV-1 GFP expressing T-Cells. The variability in single cell intensity is obvious. Such an image from the Weinberger and Simpson labs made a recent 2010 cover of the Biophysical Journal [14].

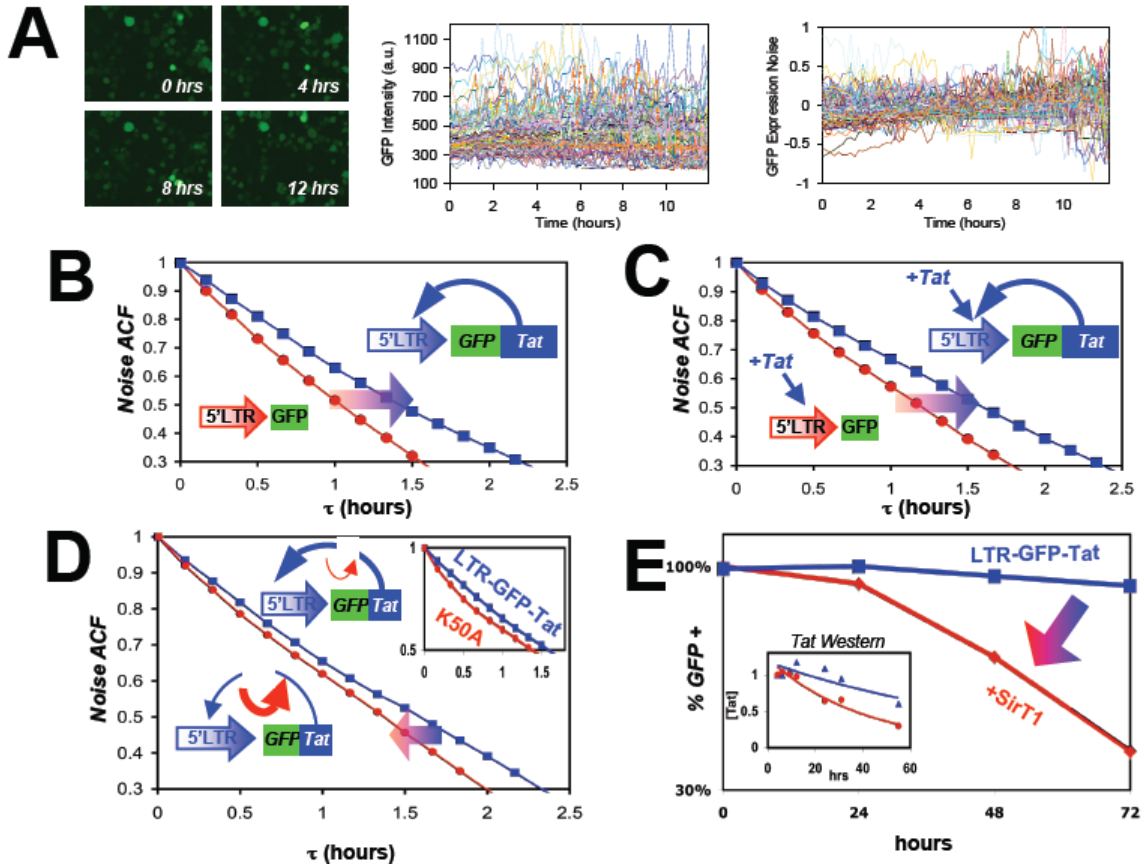


Figure 3.14 Measuring positive-feedback strength by exploiting inherent gene expression noise. **a**, Single-cell time-lapse microscopy images of LTR-GFP-IRES-Tat (heretofore termed LTR-GFP-Tat) isogenic Jurkat T-cells over 12hrs (left, images captured every 10 mins), single-cell intensity (middle) and processed noise trajectories (right) for determining high frequency noise autocorrelation functions (ACFs). **b**, Measured ACFs for LTR-GFP-Tat (ACF $\tau_{1/2} = 1.59 \pm 0.08$ hrs) and LTR-GFP control (ACF $\tau_{1/2} = 1.2 \pm 0.12$ hrs); positive-feedback shifts HF-ACFs to longer times. **c**, Measured ACFs after stimulation with exogenous Tat protein for LTR-GFP-Tat (ACF $\tau_{1/2} = 1.77 \pm 0.08$ hrs) and LTR-GFP control (ACF $\tau_{1/2} = 1.37 \pm 0.10$ hrs). **d**, Reducing feedback strength in LTR-GFP-Tat by over-expression of SirT1 (red circle) or via a mutant LTR-GFP-Tat-K50A circuit (inset) decreases ACF shift (ACF $\tau_{1/2} = 1.54 \pm 0.07$ hrs and 1.55 ± 0.08 hrs, respectively) compared to wild-type LTR-GFP-Tat circuit (blue diamond; ACF $\tau_{1/2} = 1.76 \pm 0.09$ hrs). Measurements performed after stimulation of positive-feedback with TNF- α . **e**, SirT1 over-expression in LTR-GFP-Tat cells (red) induces two- to sixfold quicker decay in LTR gene expression relative to wild-type LTR-GFP-Tat cells (blue), as measured by flow cytometry for GFP (105 cells sorted from the Tat transactivated state at time=0) or quantitative protein blot for Tat protein after 4 hr TNF α stimulation (inset), respectively.

feedback and positive values indicate positive feedback that increase the $\tau_{1/2}$ ($\tau_{1/2_FB} > \tau_{1/2_nonFB}$). Similarly, positive feedback also extends the duration of transient excursions by a factor of $1/(1-T)$ [9, 70].

Feedback strength of the minimal HIV LTR-GFP-Tat circuit [12] was quantified using time-lapse single-cell fluorescence microscopy (Fig. 3.14a). ACFs of the minimal feedback (LTR-GFP-Tat) and non-feedback (LTR-GFP) circuits were compared both with and without the presence of exogenous Tat protein stimulation (Fig. 3.14b,c). The expected extension of gene expression pulses by Tat positive feedback was observed, and resulted in an increase of pulse duration of at least 60% and as much as ten-fold.

3.3.3.2 Reduction of the minimal circuit strength of positive autoregulation

Tat positive feedback modulation was performed by overexpressing SirT1 (reducing the life-time of acetylated Tat needed for feedback in the system) or using a previously characterized Tat mutant [12, 79] (K→A substitution at amino acid 50). These two feedback modulations significantly reduced feedback strength, which was observed by a shift in the composite autocorrelation functions (CAC) (Fig. 3.14d). Figure 3.14e shows how this weakened feedback strength manifests in the decay of total levels of Tat and GFP in the system (either by quantitative protein blot of Tat or flow cytometry of LTR-GFP-Tat cells).

3.3.3.3 Correlation shifts in wild type transactivated HIV-1

The next objective was to measure positive feedback mediated correlation shifts from previously characterized full length HIV-1 [89, 92] containing GFP cloned in place of Nef (Fig. 3.15b). GFP is a direct reporter for the level of Tat in the system because Tat, Rev, and Nef are alternatively spliced from one mRNA [97]. Here experiments were induced with Tumor Necrosis Factor alpha, $TNF\alpha$, known to up-regulate promoters containing NF- $K\beta$ binding sites (e.g. the HIV LTR). The positive feedback shift in the full-length virus was significant and comparable to the minimal LTR-GFP-Tat circuit (Fig. 3.15a,b). To show that Tat feedback strongly biases the cell fate and drives sufficiently long gene expression transients, time-lapse microscopy and flow cytometry

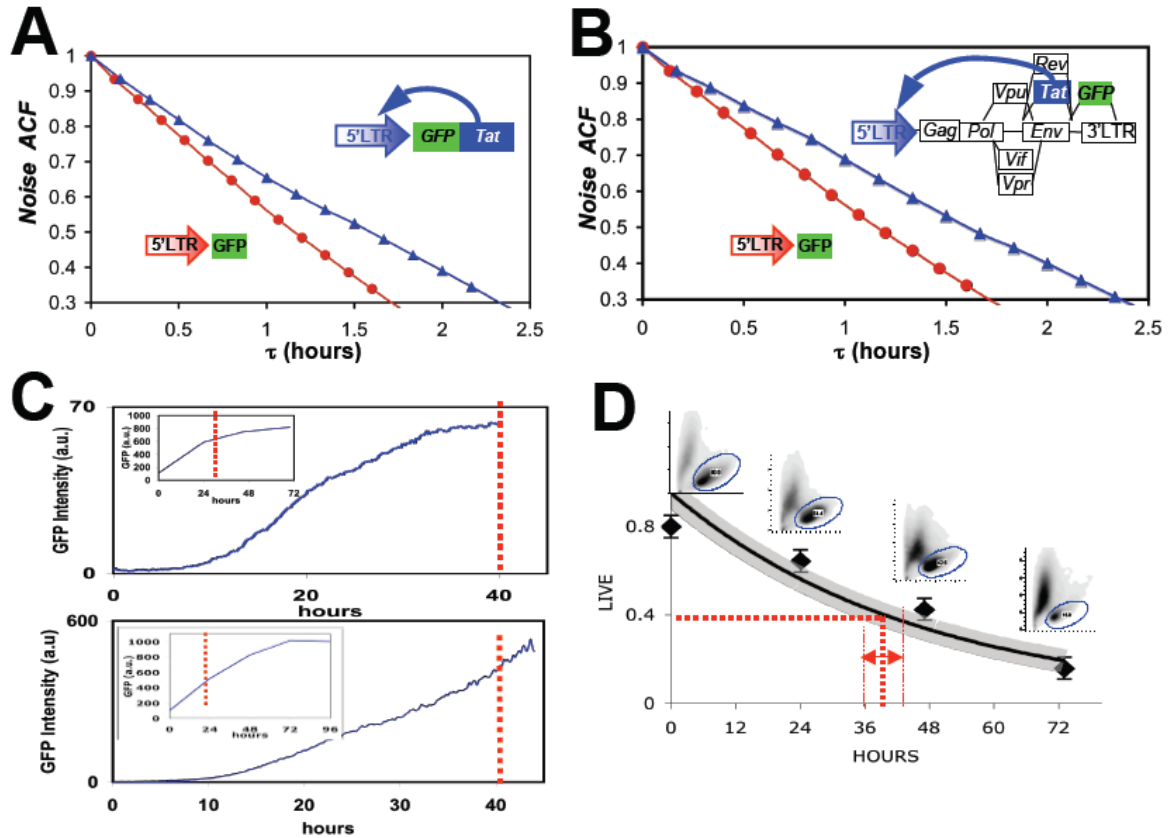


Figure 3.15 Positive-feedback strength drives an extended Tat expression transient in both minimal Tat circuits and full-length HIV-1. **a-b.** Noise ACF shift for LTR-GFP-Tat cells and full-length HIV-1 infected cells after TNF α induced reactivation. **c.** Time-lapse microscopy and flow cytometry (insets) for LTR-GFP-Tat (top) and full-length HIV-1 (bottom) after TNF activation show that expression continues to increase past 40hrs. **d.** Flow-cytometry live/dead analysis of full-length HIV-1 infected cells after activation by TNF α : half-life measured is 39.5 hrs \pm 5 hrs. Density plots shown above data points are forward-scatter (horizontal axis) vs. propidium iodide live/dead intensity (vertical axis). TNF α did not induce significant cell death over 72 hrs in LTR-GFP-Tat or LTR-GFP controls (data not shown here).

showed that the minimal circuit and full-length HIV have expression transients of more than 30 hours (Fig. 3.15c, full-length for >40hrs). This is sufficiently long as the reported and verified active and lytic T-cell half-life is $t_{1/2}=39.5\text{hrs}\pm 5\text{hrs}$ (Fig. 3.15d).

3.3.3.4 Reduction of strength of HIV-1 positive autoregulation

To test whether Tat positive-feedback acts as a probabilistic switch via modulations in feedback strength, similar to the minimal circuit case Tat positive-feedback strength was artificially weakened by overexpressing SirT1 in the full-length HIV-1 system (Fig. 3.16a). The over-expressed SirT1 ACFs showed a notable shift towards weakened feedback (Fig. 3.16a). Using flow cytometry overexpressing SirT1 cells showed an increased bias towards latency (Fig. 3.16b). This result supports the hypothesized model of a cell fate switch determined by Tat transcriptional pulses and modulated by variable strength Tat positive-feedback (for example, via SirT1 activity). A summary of all of the measured composite $\tau_{1/2}$ and strengths of regulation (T) are provided in table 3.2 [15].

Table 3.2 Summary of Experimental Composite $\tau_{1/2}$ and Strength of Regulation (T)

	Experiment	$\tau_{1/2}/\tau_{1/2\text{-NFB}}$	# of cells accounted for	Estimated T
1.	LTR-GFP No Drug	1	31	0
2.	LTR-GFP + Tat	1	43	0
3.	LTR-GFP + $\text{TNF}\alpha$	1	30	0
4.	LTR-GFP-Tat No Drug	1.33 (± 0.06)	77	0.9 (+0.1, -0.4)
5.	LTR-GFP-Tat + Tat	1.29 (± 0.06)	71	0.6 (± 0.4)
6.	LTR-GFP-Tat + $\text{TNF}\alpha$	1.29 (± 0.07)	57	0.6 (± 0.4)
7.	LTR-GFP-Tat – SirT1 overexpressed + $\text{TNF}\alpha$	1.19 (± 0.05)	94	0.22 (± 0.07)
8.	LTR-GFP-Tat k50A mutant + $\text{TNF}\alpha$	1.2 (± 0.06)	63	0.23 (± 0.08)
9.	Full length HIV-1 + $\text{TNF}\alpha$	1.38 (± 0.09)	41	0.97 (+0.03, -0.37)

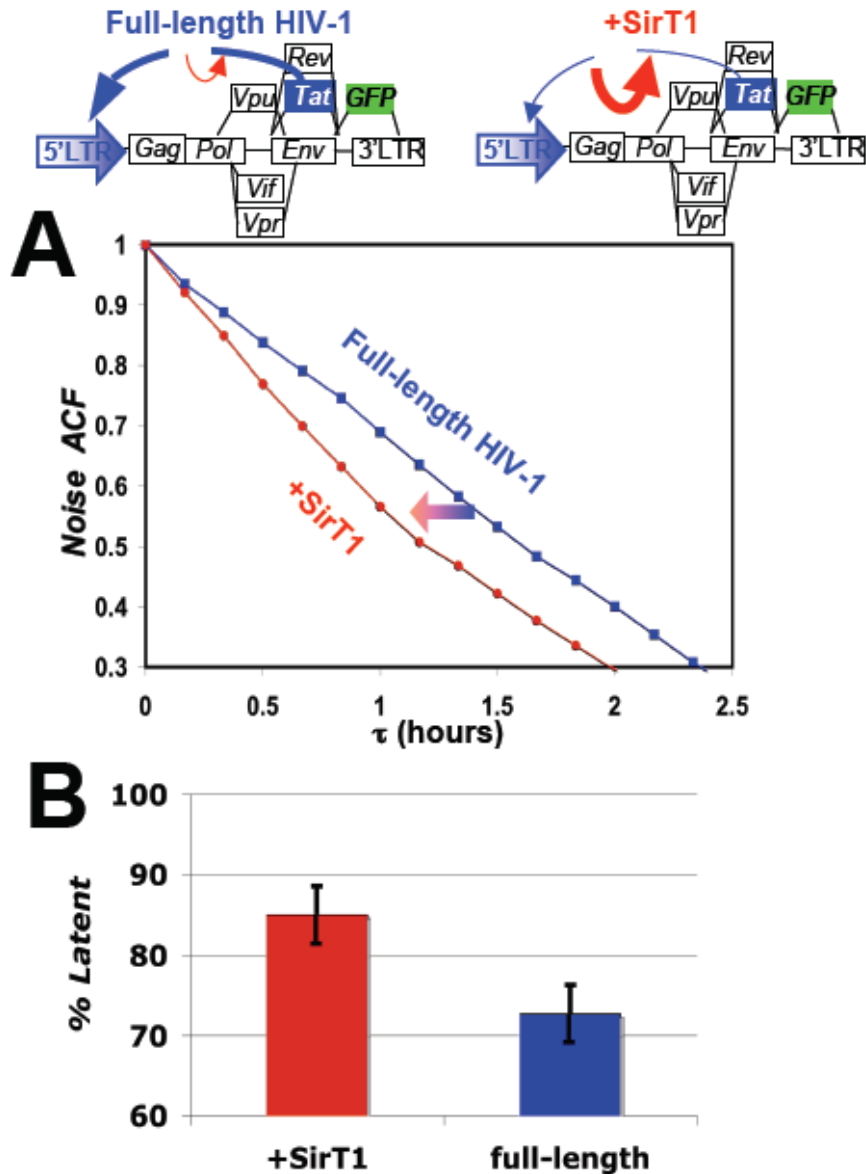


Figure 3.16 SirT1 over-expression, in full-length HIV-1 decreases positive-feedback strength and increases the probability of latency. **a.** Noise ACF for full-length HIV-1 (blue) and SirT1 over-expression in full-length HIV-1 (red). SirT1 over-expression yields weaker positive-feedback strength compared to full-length HIV-1 alone ($\tau_{1/2} = 1.35 \pm 0.08$ vs. $\tau_{1/2} = 1.76 \pm 0.08$, respectively). **b.** Analytical flow cytometry data showing % of latent cells (i.e. % of cells not expressing GFP) from triplicate sorts collected 96hrs post FACS sorting of TNF- α -activated populations of SirT1 overexpressing (red) and full-length HIV-1 (blue) sorts. SirT1 over-expression results in a significantly higher percentage of latent cells post-transactivation. Error bars are ± 1 s.d., as found from triplicate runs of the same experiment.

3.3.3.5 Summary of positive autoregulation effects on noise

For modest feedback strength ($T < 1$), +AR follows the same theory as -AR, but with reversed effects. That is, +AR increases the noise magnitude and lowers the frequency of the fluctuations. The HIV-1 Tat circuit has evolved into a modest feedback strength +AR circuit that generates stochastically-timed transient pulses. The baseline pulsing behavior is likely to be controlled by the chromosomal integration site (see Chapter 4), but the Tat +AR interacts with this background pulsatile behavior to create a probabilistic decision circuit that usually chooses active infection, but can choose a long-lived latent state that confounds therapeutic intervention. This is an example of both the use of noise for a functional advantage (bet hedging by generation of a latent state), but also of the active regulation of noise structure (through +AR) to achieve the functional objective. As such, this example will be followed out in more detail using a novel noise analysis technique in Chapter 4. However, first it is important to take a closer look at the origins of the stochastic expression bursting that initiates the HIV-1 active infection-latency decision.

3.4 Transcriptional Regulation

As a final example of gene circuit architecture shaping the resulting noise structure, the section considers the two-state model of bursty transcription. For example, transcription may be controlled by a protein-DNA interaction at an operator site within the promoter that can either activate or repress transcription (Fig. 3.17). In eukaryotes, the large genome is compacted by wrapping around nucleosomes that may make the gene inaccessible for transcription for some periods of time. At other times, the gene may be released from the nucleosome, which allows access to the gene promoter for transcription. In both these examples, transcription switches between active and inactive states, and a two-state transcription model has been established in the literature (Fig. 3.17) [10, 74]. The two-state model presented here will provide the analytical backbone for studies presented in chapters 4 and 5.

3.4.1 Two-state model of transcriptional bursting

Transcriptional bursting is a model of gene expression where the expression rate is controlled by switching between discrete high and low transcriptional rates (Fig. 3.17). The average rate is determined by the fractional amount of time spent in each of the two states. The model is illustrated by introducing equations adapted from an earlier analysis [10] with the simplifying assumptions that the low expression rate is 0, and that other than burst dynamics there is only one dominant time constant (usually either protein or mRNA decay dilution) represented by the rate constant γ_d . These assumptions are only made to lead to simple analytical expressions that aid in developing an intuitive understanding of the system.

The transcriptional bursting is represented by three model parameters (Figs. 3.18 and 3.19): (1) the transcription rate in the high expression state, α ; (2) the fraction of time spent in the high expression state, O , also referred to as the ‘on fraction’; and (3) the kinetics of the switching between off and on expression states, which is represented by k (referred to here as the burst kinetic rate), the sum of k_{ON} and k_{OFF} (Figures 3.18 and 3.19). Finally the burst frequency (f_B) defined as the inverse of the total time in the on and off states $f_B = 1/(\overline{\tau_{on}} + \overline{\tau_{off}})$ (Fig. 3.19).

As previously shown [47], with these assumptions, the autocorrelation function of the noise, $\Phi(\tau)$, is

$$\Phi(\tau) \approx \frac{\alpha O b}{\gamma_d} b e^{(-\gamma_d \tau)} + \left(\frac{\alpha O b}{\gamma_d} \right)^2 \frac{(1-O)}{O k} \left(\frac{\gamma_d}{\left[1 - \left(\frac{\gamma_d}{k} \right)^2 \right]} e^{(-\gamma_d \tau)} + \frac{k}{\left[1 - \left(\frac{k}{\gamma_d} \right)^2 \right]} e^{-k \tau} \right), \quad (3.12)$$

where b is the translational burst rate (average number of proteins translated from each mRNA). The average protein population, $\langle p \rangle$, is

$$\langle p \rangle = \frac{\alpha O b}{\gamma_d} \quad (3.13)$$

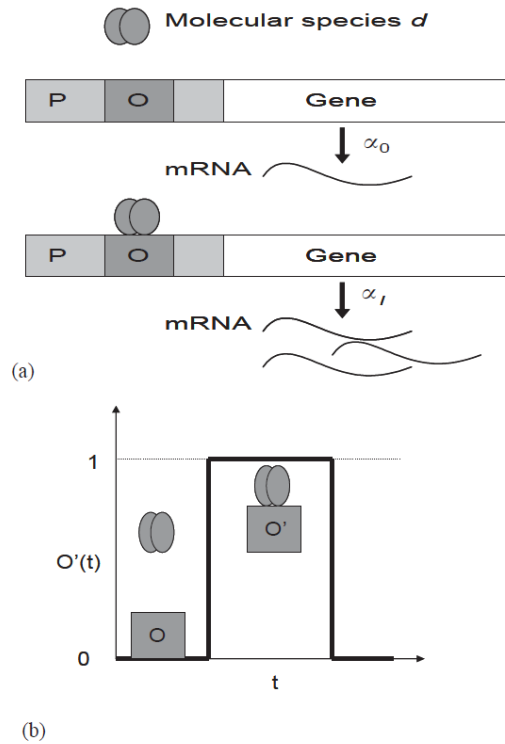


Figure 3.17 Transcriptional regulation and bursting (a) A gene is regulated by a molecular species at its operator site within the promoter region via protein-DNA interactions. The gene switches between two transcription rates (α_0 and α_1) depending on the activation state of the operator. (b) The operator population of the bound state (O') as a function of time. Figure reproduced from Simpson et al, 2004 [10].

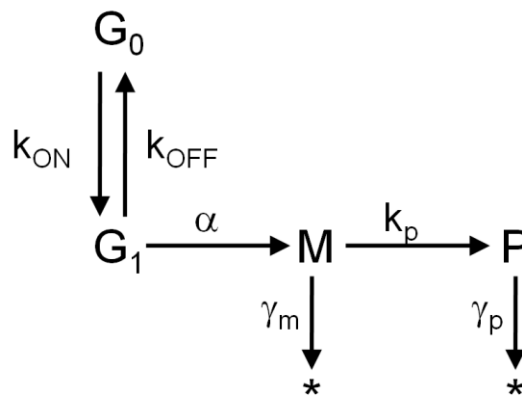


Figure 3.18 The 2-state transcription model. The gene transitions between active (G_1 ; transcription rate $=\alpha$) and inactive (G_0 ; transcription rate $=0$) states. The fraction of time spent in state G_1 , O , is given by $O = \frac{k_{on}}{k_{on} + k_{off}}$.

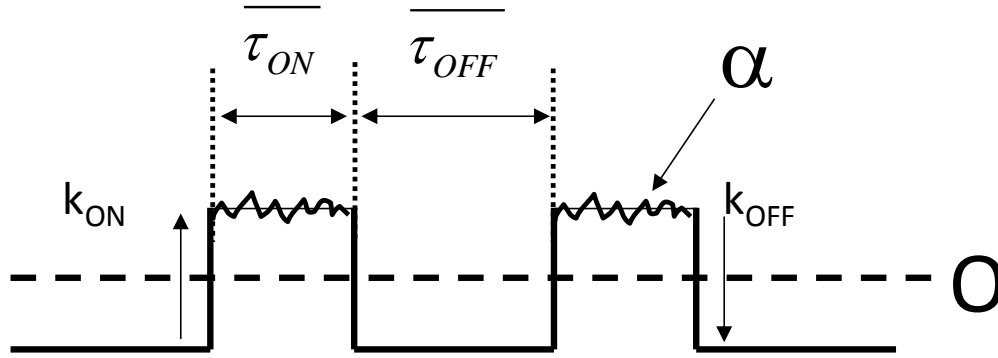


Figure 3.19 Diagram of operator state and 2-state transcription model. The diagram has O , k_{on} , k_{off} , α , τ_{ON} and τ_{OFF} labeled.

and

$$CV^2 = \frac{\Phi(0)}{\langle p \rangle^2} \approx \frac{b}{\langle p \rangle} + \frac{(1-O)}{Ok} \left(\frac{\gamma_d}{\left[1 - \left(\frac{\gamma_d}{k}\right)^2\right]} + \frac{k}{\left[1 - \left(\frac{k}{\gamma_d}\right)^2\right]} \right). \quad (3.14)$$

The first term on the right, referred to as the shot-noise term[10], is dominant at (i) low protein population; (ii) values of O that approach unity (constitutive expression); or (iii) if $k \gg \gamma_d$ (fast switching between expression states). Conversely, the second term on the right, referred to here as burst noise, is dominant for (i) low on fraction; (ii) high protein population; or (iii) slow switching between transcriptional states ($\gamma_d \gg k$).

The effect of transcriptional bursting is to increase both the noise magnitude and the half correlation time. This effect can be quite modest, giving noise that is essentially the same as constitutive expression, if O approaches unity or if k is large (fast switching between states). However, for low values of O and k , transcriptional bursting can dominate noise behavior, leading to well defined and separated bursts of expression. It is

this latter case that would seem to be of most interest with respect to the HIV-1 Tat circuit, as these stochastic bursts of expression would seem to be the seeds of the active infection-latency decision. The topic of the next chapter will be to look at this bursty behavior in some detail, and in particular for the HIV gene circuit.

CHAPTER 4: Noise Mapping

The previous chapter described in detail the coupling between gene circuit and noise structure (Equation 3.1). This chapter will describe a formal method, called noise mapping, for showing this coupling. Furthermore, noise mapping opens up the ability to investigate the use of noise measurements for their probative value. That is, this chapter will explore the use of noise measurements to elucidate the structure and function of the underlying gene circuit.

This chapter will begin with a description of noise mapping methodology [47], followed by a discussion of the experimental realities of noise mapping, and finally will demonstrate a novel experimental noise mapping approach for the investigation of transcriptional bursting across the human genome. This final piece of experimental work will be used to address the role of noise in the establishment of latency in the retrovirus HIV-1.

4.1 Noise Maps as a Gene Circuit Discovery Tool

Some of the initial noise studies mentioned in Chapter 3 used well-defined synthetic gene circuits and demonstrated the probative value of noise in their characterization [15, 70]. Like all new methods, this was a crucial step in the development of a gene expression noise spectroscopy science. However, it would be much more useful if noise measurements could be employed in characterizing gene circuits whose structure is only partially known. The systematic methodology to characterizing gene circuits and discovery of their structure-function relationships was recently reported by Cox *et al.* in 2008 [47]. In general, the approach is based on how much the experimentally measured noise traits (noise magnitude (CV^2) and correlation ($\tau_{1/2}$)) of a gene deviate from the theoretically predicted noise traits of a canonical constitutively expressed transcription-translation circuit using known kinetic parameters describing the gene circuit [47] (Figure 4.1).

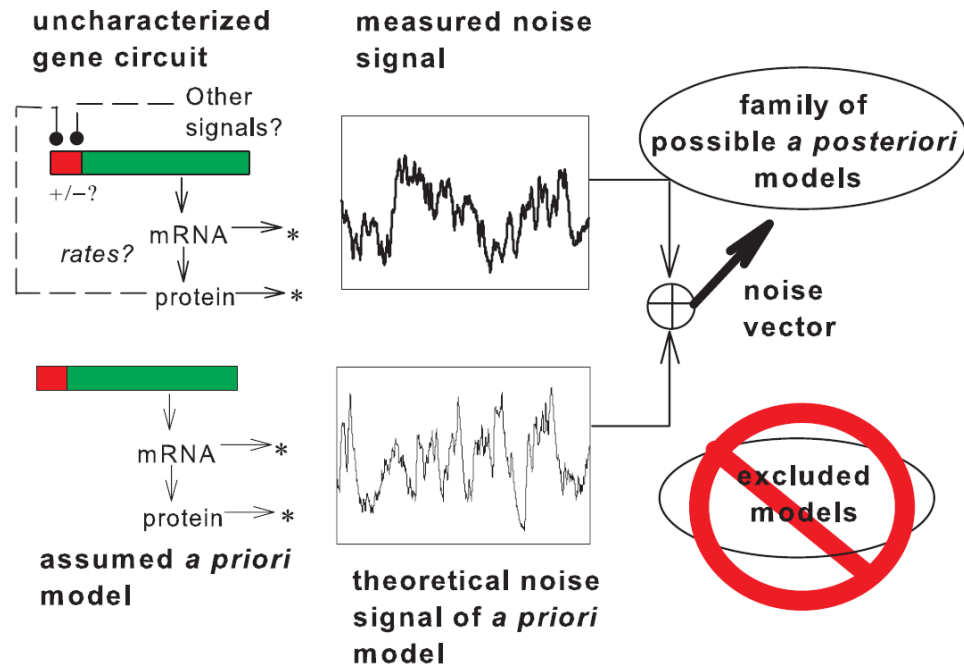


Figure 4.1 Noise mapping as a gene circuit discovery tool. The noise regulatory vector for an uncharacterized gene circuit is determined by comparison of its experimental noise structure to the noise structure of an assumed model. The vector points toward a family of gene circuits that includes the true gene circuit, and away from inappropriate models. [Figure adapted from Cox et al., 2008[47]]

4.1.1. Defining the noise regulatory vector and 3-D noise map space

Chapter 3 defined a noise vector, which in general is an m -component vector that describes the noise structure. For example, from the work described in chapter 3, a noise vector could have components that relate to noise magnitude, noise correlation, and single cell distributions. Each of these elements could contribute more than one dimension to the noise vector. For example, correlation could be characterized not only by the half correlation time, but also by the $1/10$ correlation time, and single cell distributions could be characterized by mean, variance, skew, or any other moments of the distribution. But for simplicity and for graphical representation, it is convenient to stick with the noise vector definition stated in chapter 3:

$$\vec{N} \equiv \log CV^2 \hat{m} + \log \tau_{1/2} \hat{c} \quad (4.1)$$

where \hat{m} and \hat{c} are orthogonal unit vectors.

Every gene circuit can be broken down into two components: (1) a constitutively expressed core; and (2) regulatory arrangements (e.g. +AR, -AR, transcriptional bursting, etc.) that cause the circuit to stray from constitutive behavior. Assuming that the noise vector for the constitutively expressed core is known (or can be estimated), a regulatory vector ($\Delta \vec{N}_{reg}$) may be defined as

$$\Delta \vec{N}_{reg} = \vec{N} - \vec{N}_{const} = \Delta \log CV^2 \hat{m} + \Delta \log \tau_{1/2} \hat{c} = \log(CV^2 / CV_{const}^2) \hat{m} + \log(\tau_{1/2,m} / \tau_{1/2,const}) \hat{c} \quad (4.2)$$

where the subscript *const* indicates constitutive expression. This regulatory vector describes how regulation has altered the noise structure of the gene circuit.

Both CV_{const}^2 and $\tau_{1/2,const}$ are found from the autocorrelation function for constitutively expressed protein, $\Phi_p(\tau)$, [9, 47, 70]:

$$\Phi_p(\tau) = \langle p \rangle \left\{ \left(1 + \frac{b}{1 - (\gamma_p / \gamma_m)^2} \right) \exp(-\gamma_p \tau) + \left(\frac{b(\gamma_m / \gamma_p)}{1 - (\gamma_m / \gamma_p)^2} \right) \exp(-\gamma_m \tau) \right\} \quad (4.3)$$

where γ_p and γ_m are the decay rates of protein and mRNA, $\langle p \rangle$ is the mean protein abundance, and b is the translational burst rate. CV_{const}^2 is inversely proportional to the mean protein abundance, $\langle p \rangle$ (Equation 3.6), and the correlation time, which is dominated by protein dilution and decay, is invariant to changes in transcription rate. Therefore, for a given protein and mRNA decay rates the regulatory vector can be determined for any protein population graphically as shown in Figure 4.2. The constitutive expression line in the 3-d space of $\langle p \rangle$, CV^2 , and $\tau_{1/2}$ (shown in two different 2-d projections in Fig. 4.2) is referred to as the bias line.

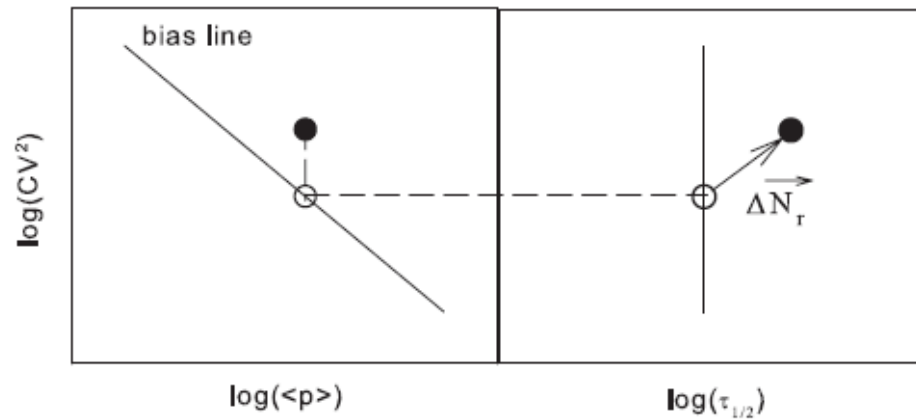


Figure 4.2 The noise regulatory vector and its relationship to the 3D noise map. Graphical definition of the bias line and 2-component regulatory vector ΔN_{reg} . The bias line represents the behavior of the *a priori* model. To determine ΔN_{reg} for a protein with measured coordinates on the noise map (filled circle), one first locates the bias point (open circle) by projecting vertically to the bias line. ΔN_{reg} is defined by the 2D vector connecting the bias point to the measured point in the $\log(\text{CV}^2)$ — $\log(\tau_{1/2})$ plane. [Figure adapted from Cox et al., 2008[47]]

4.1.2. Theoretical noise maps of main regulatory motifs

The theoretical noise map methodology presented here relies on an ideal world picture in which infinite time duration measurements are possible. Under this assumption, the measured and theoretically predicted (bias line and noise map origin) noise space coordinate are truly a single point in the 3D space. Using exhaustive and long duration stochastic simulations, it is possible to simulate various regulatory motifs of interest, with a variety of parameters, to see how the noise regulatory vector moves in the 3D noise map space. This picture changes significantly in real world experiments which will be discussed later.

4.1.2.1 Theoretical noise map of transcriptional regulation

Slow gene activation kinetics, where the rate of switching between active and inactive transcription states is comparable to the rate of protein and mRNA decay in the gene circuit, was explored in the noise map space using the 2-state model autocorrelation

function and noise magnitude presented in Chapter 3 (Equations 3.12 and 3.14). Here the slow activation adds a noise term (referred to as either burst or operator noise) to the gene expression noise and increases the half correlation time. As a result, the regulatory vectors for transcriptional bursting are found in the first quadrant ($+\Delta\tau_{1/2}$, $+\Delta CV^2$) of the noise map space (Fig. 4.3). The mean gene activity level as defined in chapter 3 is $O = k_{ON}/(k_{ON} + k_{OFF})$ and deviations from bias line behavior are largest at intermediate levels of O and smallest as O approaches 0 and 1. Also two ratios that characterize the DNA-binding kinetics were defined and scanned: $\kappa_1 = k_{OFF}/\gamma_p$ and $\kappa_2 = k_{OFF}/\alpha_0$. In general κ_1 has a larger effect on the direction of the vector while κ_2 has a larger effect on its overall magnitude.

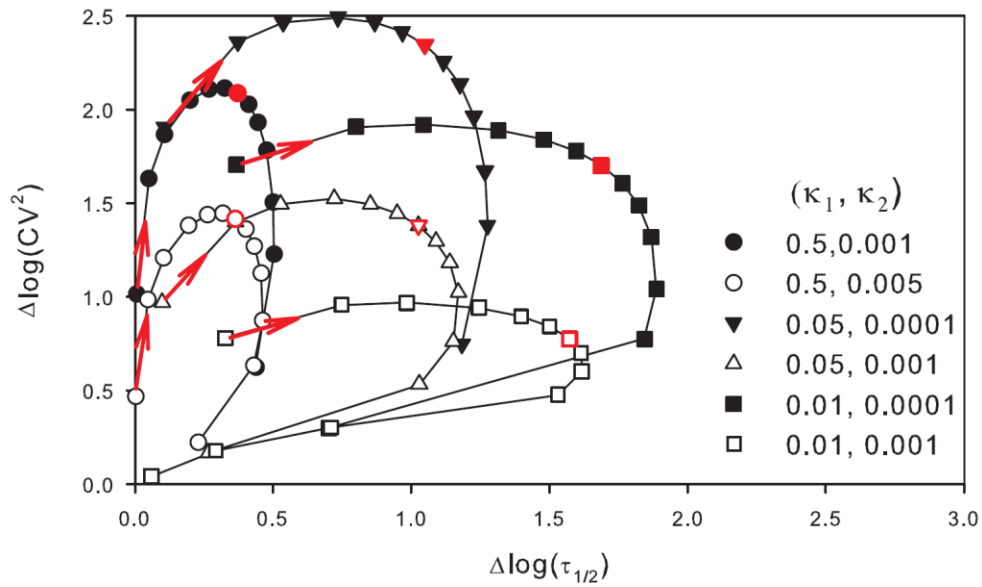


Figure 4.3 Noise regulatory vectors for slow gene activation kinetics. Points indicate ΔN_{reg} for O values of (starting from and moving in the direction of the red arrow) 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5 (red point), 0.6, 0.7, 0.8, 0.9, 0.95, and 0.99. The effect of gene activation kinetics is captured in the ratios $\kappa_1 = k_{OFF}/\gamma_p$ and $\kappa_2 = k_{OFF}/\alpha_0$. [Figure adapted from Cox et al., 2008[47]]

4.1.2.2 Theoretical noise map of negative autoregulation

The previous section showed that parameter deviations from the *a priori* (constitutive) model are quantitatively captured by the noise regulatory vector. To investigate the noise regulatory vectors of a negatively autoregulated protein, long-duration stochastic simulations scanning a range of parameter space was implemented (Fig. 4.5). Negative autoregulation can be separated into two dynamic and parallel processes (Fig. 4.4): (1) The regulation loop where the regulator represses its own expression, and (2) transcriptional repression at the operator sites within the promoter where autoregulator-DNA interactions switch the promoter between active and inactive, two-states of transcription (i.e. the model in the last section and section 3.4.1). Chapter 3 described a model of negative autoregulation that does not account for the kinetic rate of autoregulator-DNA binding and predicted a suppression of noise magnitude and extension of noise bandwidth (Section 3.3.1). In the current, more detailed model, when transcriptional regulation is accounted for, the predicted noise shift only occurs at fast rates of binding kinetics (Fig. 4.5). Using the same ratios as the previous section to describe the binding of auto-regulator and DNA it is found that slower binding kinetics moves the noise regulatory vector outside the third noise map quadrant and can even lead to an increase of noise magnitude and correlation when binding kinetics are slow enough (black triangles, Fig. 4.5). This finding may explain contradictory reports in the literature with some reporting that $-AR$ decreases noise, while other have suggested that $-AR$ increases noise compared to constitutive expression (e.g. [98]). In fact, the complete view of $-AR$ is that it both removes noise and decreases correlation time through the loop transmission effect (section 3.3.1) and increases noise and correlation time through the operator noise effect (section 3.4.1). The net effect is totally dependent on which of these two mechanisms is dominant.

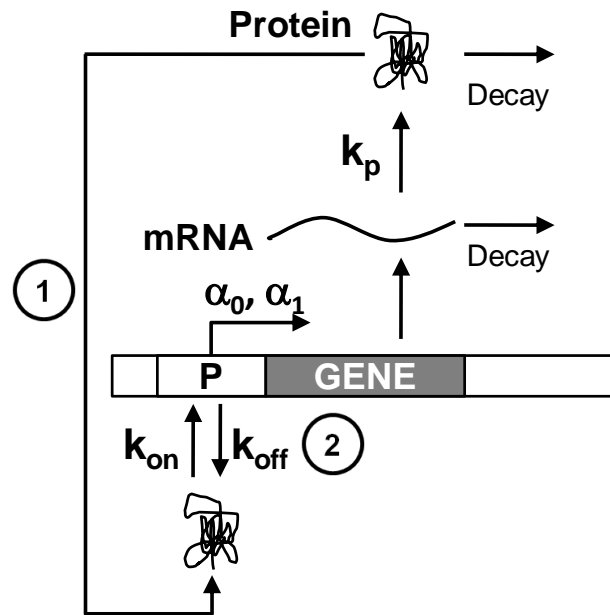


Figure 4.4 Negative autoregulation and autoregulator-DNA binding. A detailed view of negative autoregulation shows the opposing effects of two regulatory motifs: as labeled in the figure, (1) –AR loop transmission effect, predicted to suppress noise magnitude and lower correlation times when autoregulator-DNA binding is fast, and (2) transcriptional repression and switching between 2-transcription states based on the autoregulator-DNA binding kinetics can increase both noise magnitude and correlation when binding kinetics are sufficiently slow.

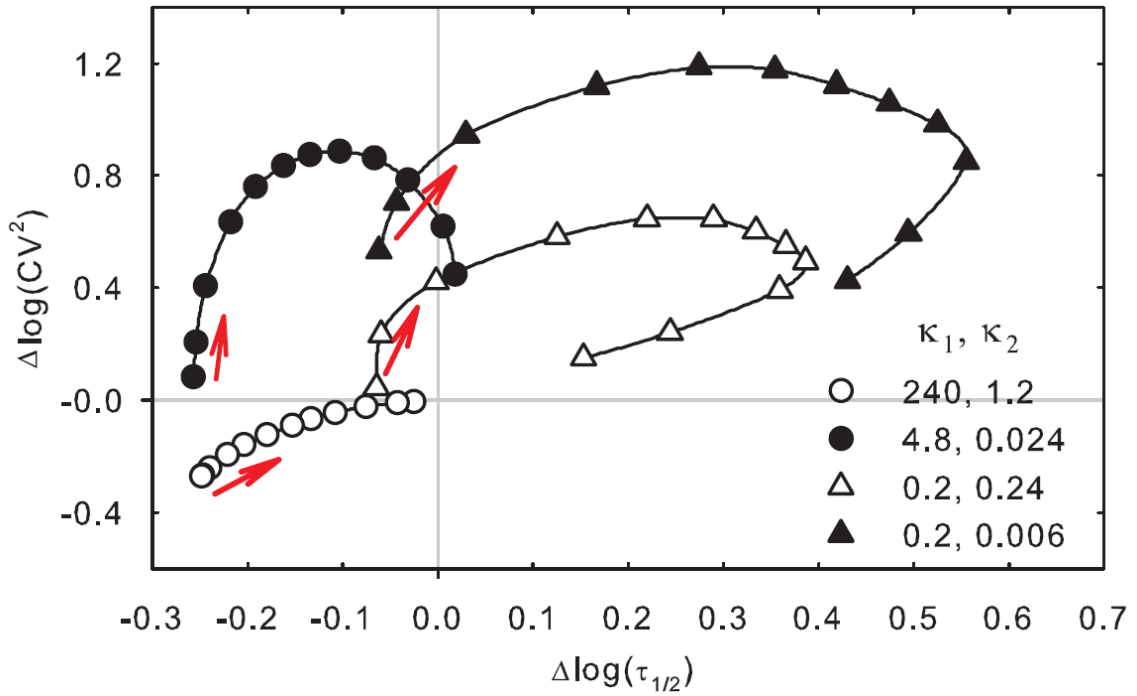


Figure 4.5 Noise regulatory vectors for negative autoregulation. Points indicate ΔN_{reg} at gene activation levels of (starting from and moving in the direction of the red arrow) 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.95. The effect of decreasing κ_2 is to increase $\Delta \log(\tau_{1/2})$ when $\kappa_1 > 1$ and to increase both $\Delta \log(CV^2)$ and $\Delta \log(\tau_{1/2})$ when $\kappa_1 < 1$. [Figure adapted from Cox et al., 2008[47]]

4.1.3. Noise vector domains for various regulatory motifs

The previous section presented theoretical noise regulatory vector movement in the noise map space for two regulatory motifs and parameters of interest. The noise regulatory vector approach can elucidate behaviors that include positive and negative autoregulation, slow gene activation kinetics, differences in translational burst rate (k_p/γ_m) and protein decay rate (Fig. 4.6). Figure 4.6 provides a legend of motif occupancy in the noise map space. After discovering the region of the noise map space occupied by a specific type of regulation, additional modeling and analysis is needed to constrain the parameters present in the unknown gene circuit. The clustering of a group of genes with common functions in the noise vector space may provide evidence of common regulatory motifs or kinetic parameters. Finally, resolution in the noise map space becomes extremely important as different regulatory motifs overlap in the space (e.g. quadrant 1 with slow gene activation, negative, and positive autoregulation) and for a known motif, different sets of parameter values can overlap and occupy the same noise map space coordinates (e.g. Figures 4.3 and 4.5). The regulatory vector curves presented above are for ideal ACFs (i.e. those found from infinite duration observations), which is far from the experimental reality.

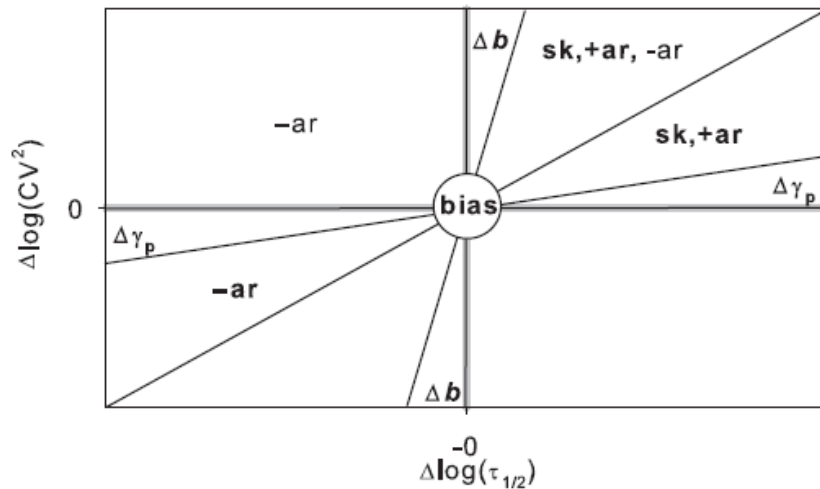


Figure 4.6 Summary of noise vector domains for various regulatory motifs. (-ar, negative autoregulation; +ar, positive autoregulation; sk, slow gene activation kinetics). Bold font denotes domains of primary influence. [Figure adapted from Cox et al., 2008[47]]

4.2 The experimental reality of noise maps

The ideal world, in which infinite duration gene expression signals are acquired, is far from realistic for several reasons. First, biology is difficult to observe over long periods of time. In the case of single cell fluorescence microscopy, the extended optical probing of cells can affect their health and behavior; cells may move throughout the experiment and enter/exit the imaging window, thereby becoming difficult to track; and finally, the local and global environment of the cells may change significantly with media depletion. An *in vivo* biochemical gene circuit is far from an ideal *in silico* version of this gene circuit where composition, structure, function, and environment are controlled and constant.

These experimental considerations lead to the reality that single-cell time-lapse gene expression measurements have limited duration observation windows that will limit the visibility of low frequency fluctuations. If these processes are assumed to be ergodic, then ensemble averages could be used to recover the low-frequency information as was done for the *E. coli* experiments in chapter 3. However, because of the issues discussed in chapter 2 (see Figure 2.20), ensemble averaging will not recover the low-frequency information lost through the high-frequency processing of noise trajectories. In such cases, it is preferable to look at the distribution of single cell behaviors instead of ensemble averages.

To demonstrate the single cell noise map spread of limited duration observations, 1600 single-cell gene expression snapshots were simulated using different observation window durations for a constitutive gene expression model with consistent kinetic parameters (Figure 4.7). The distribution of individual cells is spread around the origin, and the distribution condenses toward the origin with increasing window size. Although the distribution would collapse to a single point at the origin for infinite duration observations, even for very long windows, (e.g. 120 hours (or 5 days)) there is a marked variation in the behavior of individual cells. These simulated single-cell noise map spreads have a diagonal orientation in noise map space. That is, individual cells exhibiting a period of high noise magnitude also tend to exhibit a longer correlation time

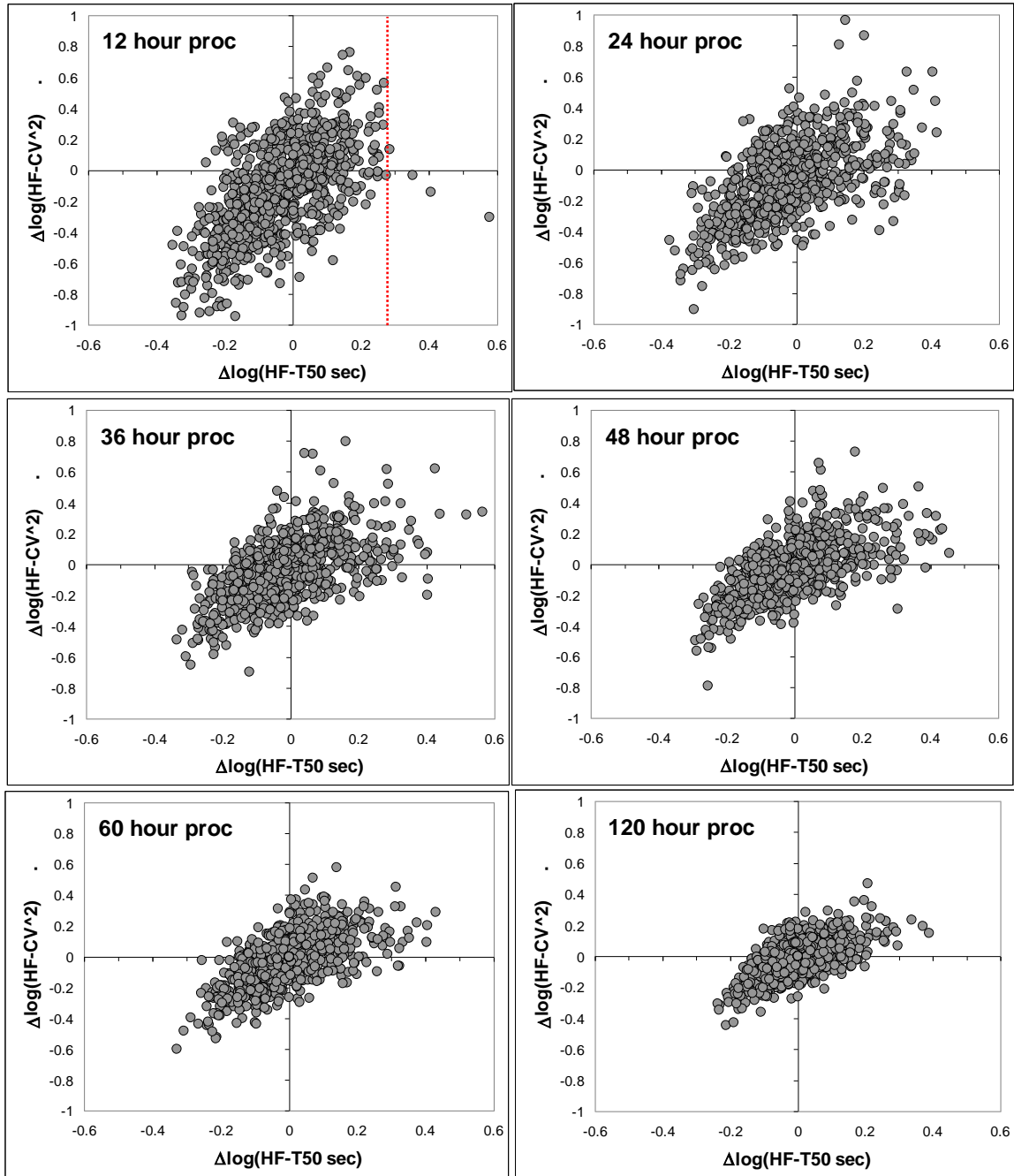


Figure 4.7 Convergence of noise map spread with increasing experimental duration. Simulated constitutive expression noise maps were generated for experiment durations of 12 to 120 hours. The single-cell scatter about the origin condenses but still occupies a diagonal region even at longer experiment durations.

during this period. This result is at least somewhat intuitive as it indicates that regression to the mean takes longer for large excursions from the mean.

4.3 Noise maps to study HIV latency

As already explained in Chapter 3, HIV-1 infected human T-cells can enter two different fates: active replication and cell lysis or proviral latency, a long-lived quiescent state where viral gene expression is turned off [86, 87]. The latent cell pool is of great importance because it is the main reservoir preventing effective HIV eradication from patients [88].

The Weinberger, Dar, and Simpson 2008 study investigated a transient-mediated fate decision in a transcriptional (positively autoregulated) circuit of HIV-1 [15] (see chapter 3). This study demonstrated that the role of +AR in this circuit is to extend the duration of stochastic pulses of expression. If these expression bursts persist long enough, the infected cell proceeds down the path of active replication and cell lysis. Conversely, if the expression bursts terminate too early, the cell enters the latent reservoir. What this study did not address was the origin and characteristic of the noisy expression bursts that were extended by +AR.

As discussed in chapter 3, gene expression can be an episodic process characterized by bursts [99-101] in transcription (see section 3.4) and translation. Evidence for transcriptional bursting has been found in yeast [34, 102], fly [103], and for specific human promoters [14, 104-106], and recent models suggest that bursts arise due to stochastic ‘waiting times’ inherent in the formation of active transcriptional complexes [104]. The physiological importance of transcriptional bursting lies in its ability to generate beneficial noise for stress responses [107] and fate determination [12, 14, 108], and provides a mechanism for regulatory control over gene expression via modulation of burst frequency [109]. It seems likely that such transcriptional bursting is the source of the noisy burst of expression that initiate the HIV-1 cycle, and it is the interplay between these transcriptional burst and the +AR circuit that ultimately drive the active infection/latency decision.

However, to date, transcriptional bursting has not been demonstrated to be a predominant mode of gene expression and genome-wide surveys of transcriptional burst frequency have not been performed. The following sections describe how the noise-mapping concept described above was applied (Fig. 4.8) to globally survey real-time single-cell expression kinetics and tests (i) whether transcriptional bursting is widespread throughout the human genome and (ii) how this transcription bursting is distributed in burst parameter (O and K) space.

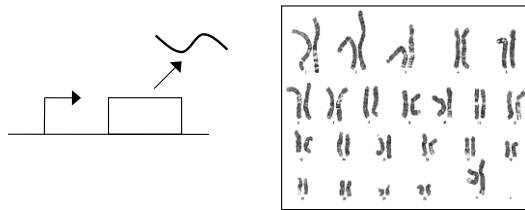
4.3.1 Experimental investigation of genome-wide transcriptional bursting in humans

4.3.1.1 Methods

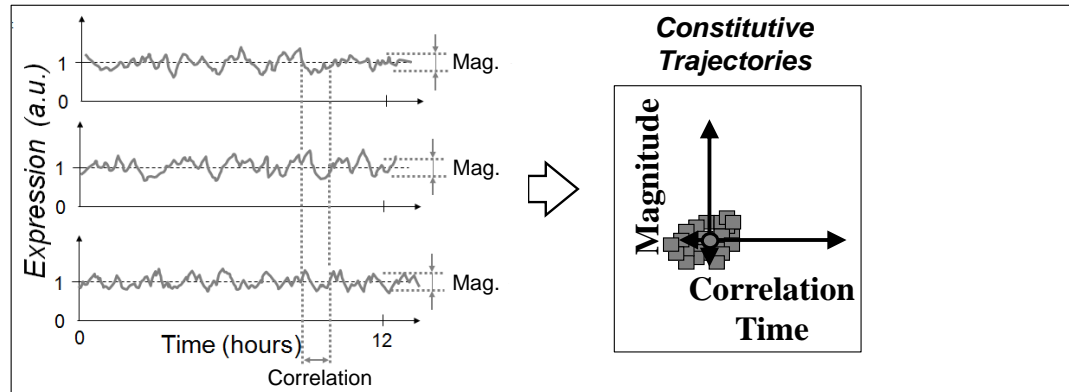
Genome-wide gene expression measurements have proven to be both labor intensive and scientifically valuable. Previous studies have elegantly used flow cytometry to measure noise magnitude in dozens or even thousands of genes in yeast [102, 110], but flow cytometry measurements are unable to provide information on fluctuation kinetics (see chapter 2). Conversely, time-lapse fluorescent imaging of individual cells provides both noise magnitude and dynamics but, to this point, the technique has required long imaging experiments for each gene circuit [35, 69]. To overcome these problems, a genome-wide noise mapping approach was employed (Figures 4.8 and 4.10) that produced a graphical representation of the distribution of noise behaviors for each individual cell in a population from relatively short-duration imaging experiments. Noise mapping allowed for extraction and identification of global expression characteristics, including gene regulation and the dynamics of transcriptional bursting, by exploiting the inherent clonality of each cell in a polyclonal population (Fig. 4.9A). Thus, noise mapping circumvents the requirement of subcloning and expansion of isogenic populations followed by long-duration imaging experiments for each clone [35, 106]. To screen for transcriptional bursting across the human genome, the semi-random pattern of integration exhibited by the HIV-1 lentivirus, where the vast majority of integrations (~69%) occur within transcriptionally-active regions, is exploited [111, 112]. Jurkat T cells were infected with HIV-based lentiviral vectors encoding a short-lived fluorescent

Figure 4.8 Scheme for probing transcriptional bursting across the genome. **(Top)** Genome-wide, transcription may be constitutive exhibiting small stochastic fluctuations. These fluctuations would result in noise maps that cluster around the origin of CV vs τ . **(Bottom)** Alternatively, transcription may be bursty and lead to large stochastic fluctuations in gene expression. Transcriptional bursting may be described by different “on fractions” (O) and burst kinetic rates (k). Noise mapping of single-cell expression trajectories can be used as a direct measure of single-cell transcriptional bursts and enable scanning for overrepresented O - k parameter ranges. Noise maps can be converted to O - k heat maps (probability burst map landscapes), which provide a direct probe for overrepresented burst dynamics.

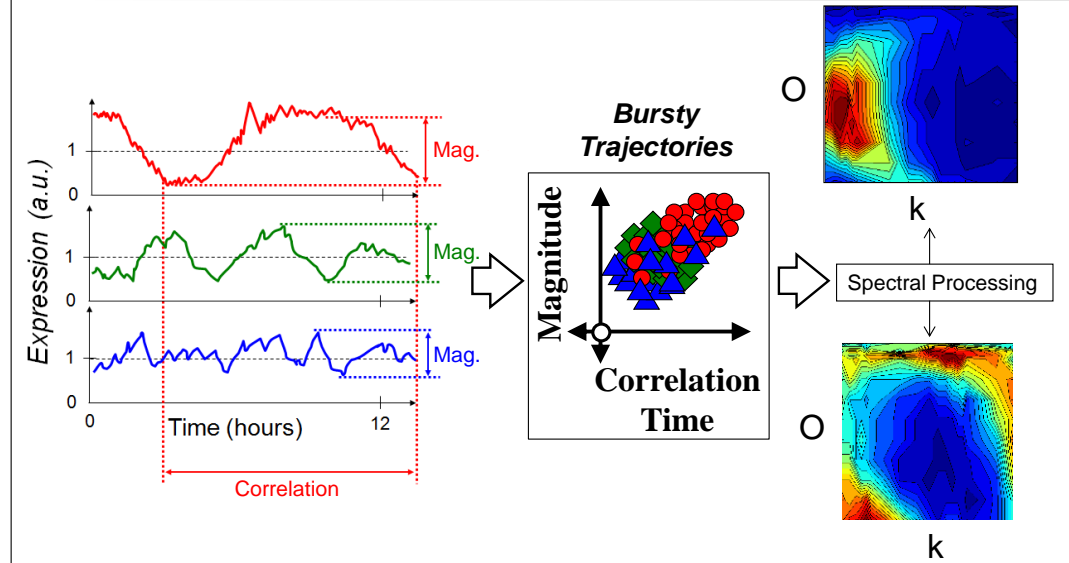
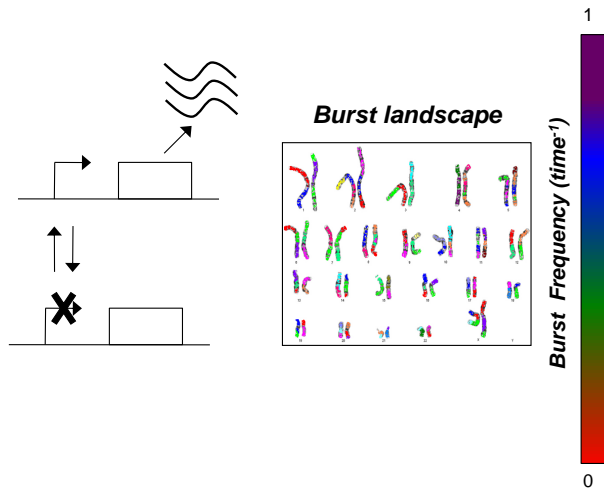
Constitutive Transcription



Noise mapping



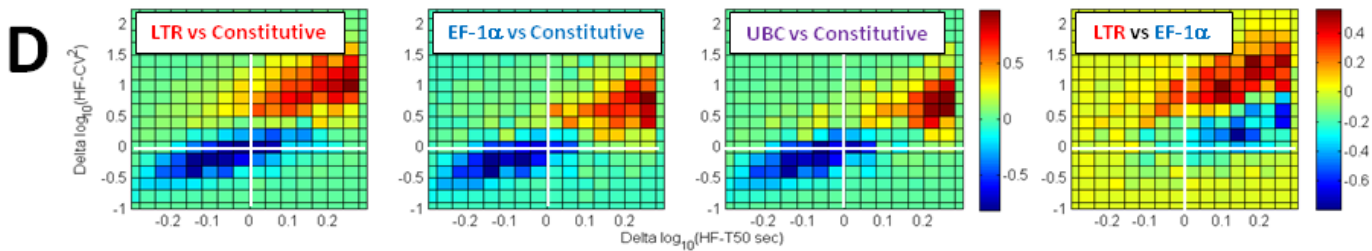
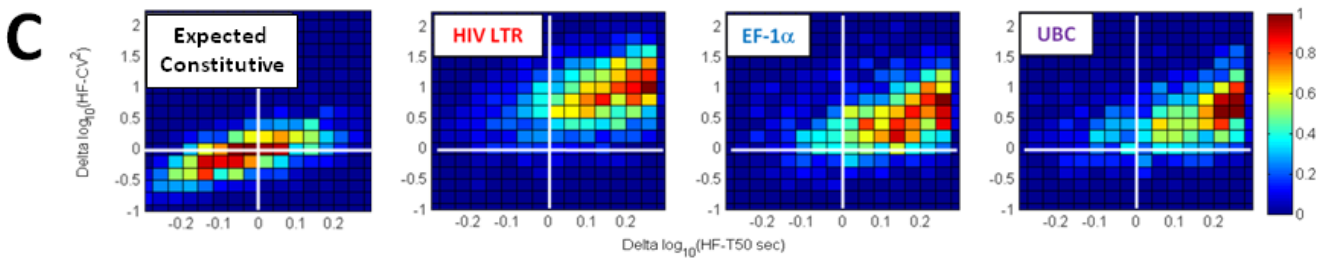
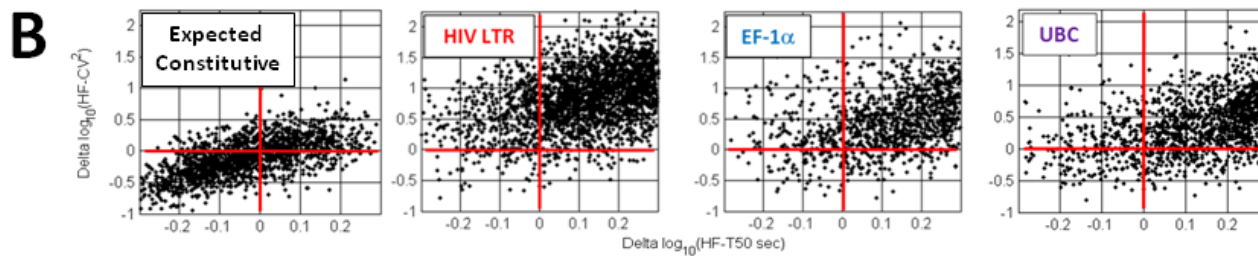
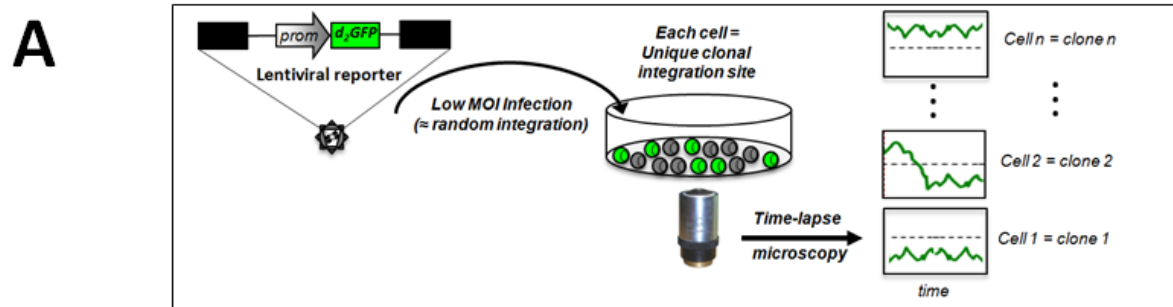
Transcriptional Bursting



protein (i.e. the two-hour half-life version of GFP, d₂GFP) to generate a polyclonal library where each individual cell contained the vector integrated at a distinct genomic position. Cells were then fluorescently imaged for 12-hours and the resulting fluorescence intensity trajectories were used to construct a noise map (Fig 4.9B and 4.10). Over 14,000 individual cells and over 8,000 distinct genomic loci with three different promoters integrated throughout the genome were analyzed (Fig 4.9B-D).

Initially, to focus on measuring the intrinsic fluctuation dynamics of genomic loci surrounding the vector integration site, a vector encoding the HIV LTR promoter driving expression of d₂GFP was utilized [90, 113]. To control for LTR-specific artifacts or vector-specific artifacts, self-inactivating lentiviral vectors encoding either the human elongation factor 1 α promoter (EF1A) or the human Ubiquitin C promoter (UBC) driving d₂GFP were also tested. UBC and EF1A are both essential cellular housekeeping genes—UBC promotes the ubiquitination cascade marking proteins for proteosomal degradation and EF1A promotes the GTP-dependent binding of an aminoacyl-tRNA to ribosomes—both are among the most abundant proteins in eukaryotic cells, and their promoters exhibit robust high-level expression across integration sites in different cell types[114, 115].

Figure 4.9 Bursty gene expression dominates across the human genome; constitutive expression is exceedingly rare. **A.** Schematic of experimental approach to measure gene expression frequencies across the human genome. Cells are infected with a lentiviral vector expressing a short-lived GFP reporter (d2GFP) so that each cell represents a unique clone harboring a single semi-random integration of reporter. Cells are tracked for 12 hours by time-lapse microscopy and noise maps are constructed for resulting trajectories. **B.** Scatter plot noise maps representing over 7000 individual cell trajectories for the HIV LTR promoter, Ef1A promoter, and UBC promoter and a noise map of simulated constitutive gene-expression trajectories (right). **C.** Noise probability density (NPD) maps which act as two-dimensional histograms of panel B, showing the probability of finding the noise of any individual cell at particular noise map locations. **D.** NPD difference maps that compare the LTR, Ef1A, and UBC promoters to constitutive expression. Compared to constitutive expression, all three promoters exhibit noise that is shifted to the upper right of the noise map (high CV^2 , high $\tau_{1/2}$). **D, far right.** NPD difference maps comparing the LTR to Ef1A. Both LTR and Ef1A exhibit almost identical shifts in $\tau_{1/2}$ for more see Figures 4.6 and 4.7.



4.3.1.2 Creating an experimental noise map

Starting with time-lapse fluorescent imaging, the creation of a noise map has three steps: (1) image processing to create time histories of reporter protein concentrations in individual cells (Fig. 4.10A and Chapter 2); (2) analysis of time histories to determine noise magnitude and correlation (Fig. 4.10B-C and Chapter 2); and (3) graphical determination of the noise vector for each individual cell (Fig. 4.10D-F). For the experiment shown (~350 cells), fluorescence was measured at 10 minutes intervals for 12 hours. For convenience, two trajectories (blue and red) are followed from measured intensities in panel A to final noise map coordinates in panel F. Chapter 2 has the details pertaining to deducing the noise in gene expression which requires separating out the stochastic fluctuation components from the deterministic components and uses a multiple-step signal-processing algorithm (Equation 2.6). After extracting gene expression noise trajectories, each trajectory in panel A is high-frequency (HF) processed by base-line suppression as described in chapter 2. The resulting noise of panel B preserves the higher-frequency components of the noise while removing much of the lower frequency noise (see Chapter 2). As described earlier, HF processing (1) prevents the calculation of erroneously long autocorrelations due to inaccuracies in the determination of the true average expression level[116], and (2) focuses on the analysis of the higher frequency fluctuations of intrinsic noise—which are directly coupled to the structure and function of the underlying gene circuit—while de-emphasizing the lower frequency fluctuations of extrinsic noise that are tied to global factors affecting gene expression[27, 70] (for additional details see Chapter 2).

HF autocorrelation functions (HF-ACFs; Fig. 4.10C) are derived from panel B and allow extraction of single-cell noise variance and half-correlation times (Fig. 4.10C). As shown in figure 4.2, determination of the noise regulatory vector requires the establishment of a bias line, which is an estimate of the noise behavior of the “constitutive core” of the gene circuit. Although in principle it may be possible to arrive at a theoretical bias line, in this work an experimental approach was adopted. Half correlation times were measured for six different monoclonal populations carrying the

Ld2G circuit. The single-cell distributions of these half correlation times were examined in search of integration site(s) where transcriptional bursting was having the least pronounced affect on noise behavior, which

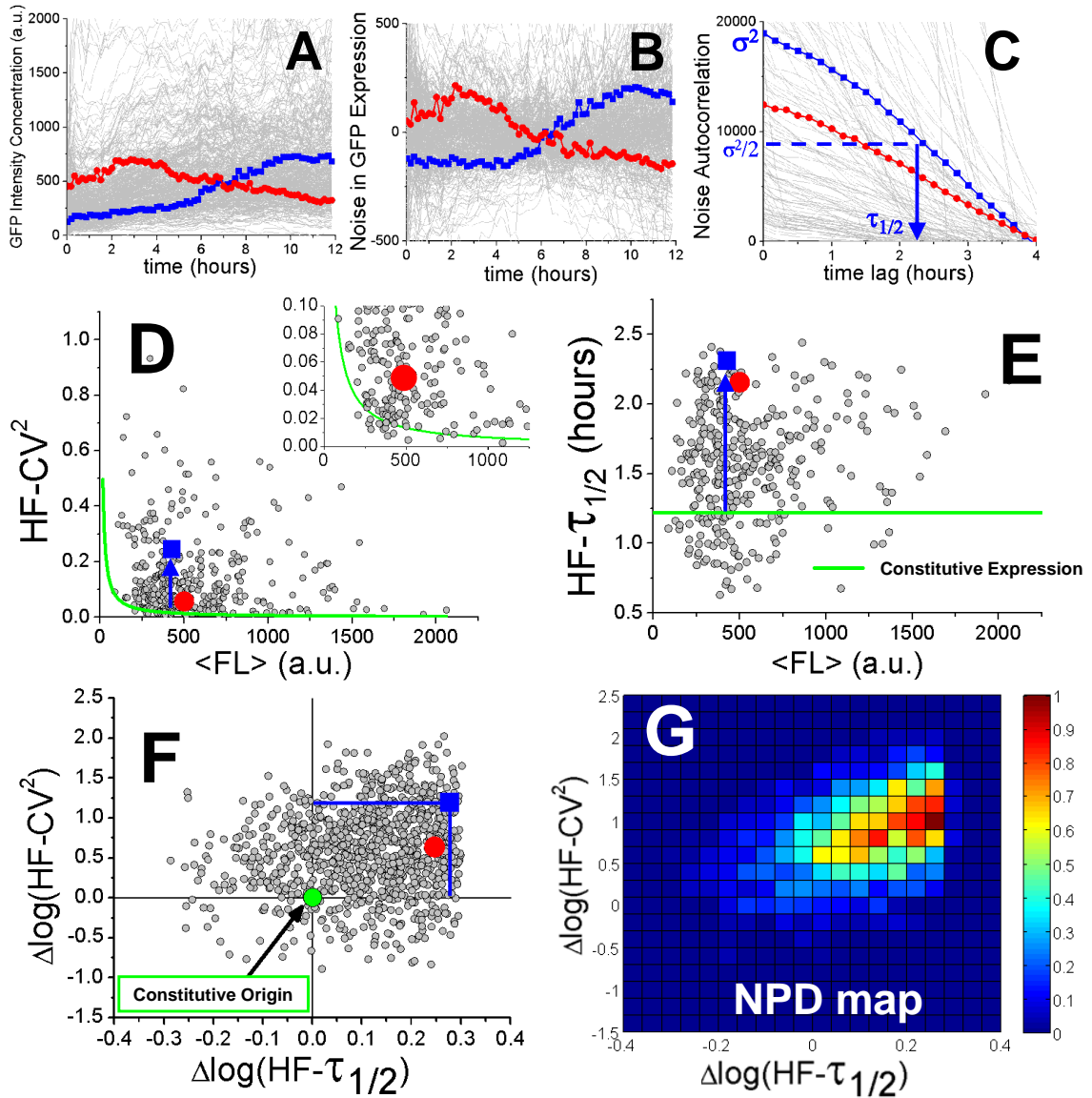


Figure 4.10 Creating a Noise Map. Detailed descriptions of how to generate a noise map are found in the above main text. Grey lines and circles represent individual single cells (or unique integration/clone). Red and blue highlight the full noise mapping process for two selected cells.

would be seen as low half correlation times (See the Appendix for LTRd2GFP isoclone processing). Two clones were identified (denoted as C32 and D36) with low half correlation times compared with other clones and with the polyclonal measurements. Accordingly, the bias line was based on these two clones (see figures 4.11 and 4.12 below), which yielded

$$CV_{const}^2(\langle fl \rangle) = \frac{0.45}{\langle fl \rangle} \quad (4.4)$$

$$T_{1/2const\ bias} = 1.22 \text{ hours.}$$

The HF- $\tau_{1/2}$ of clones C32 and D36 exhibit (to different degrees) bimodal distributions with one mode peaking between 0.97 and 1.17 hours, and the other mode peaking between 1.76 and 1.96 hours (Fig 4.12). The upper mode is evidence of transcriptional bursting in these clones, but the strong lower mode indicates that transcriptional bursting is not as pronounced in these clones as it is in others. Accordingly, the $\tau_{1/2}$ bias line was based on the lower mode and the simulation model of constitutive expression was constructed (discussed below) to fit this lower mode. Using the noise bias line, the resulting combined noise map of the C32 and D36 clones is essentially the same as the simulated constitutive expression map (Figures 4.9B and 4.13).

The simulation model of constitutive expression was constructed to be as simple as possible while remaining consistent with the experimentally determined bias line. Accordingly, a simple transcription/translation model was chosen where transcription and translation rates were selected to be consistent with the measured CV^2 bias line. The GFP half-life was set to 2 hours in agreement with its reported value [105, 117], and the mRNA half-life was selected to achieve a simulated HF- $\tau_{1/2}$ value consistent with the measured value. GFP maturation times were scanned to match the constitutive monoclonal noise map scatter shape and correlation range. Parameters for the model are given in the Appendix. This model was used to generate the constitutive expression NPD and noise maps, which are seen to be consistent with the measured noise map of the most constitutive clones (Fig. 4.13).

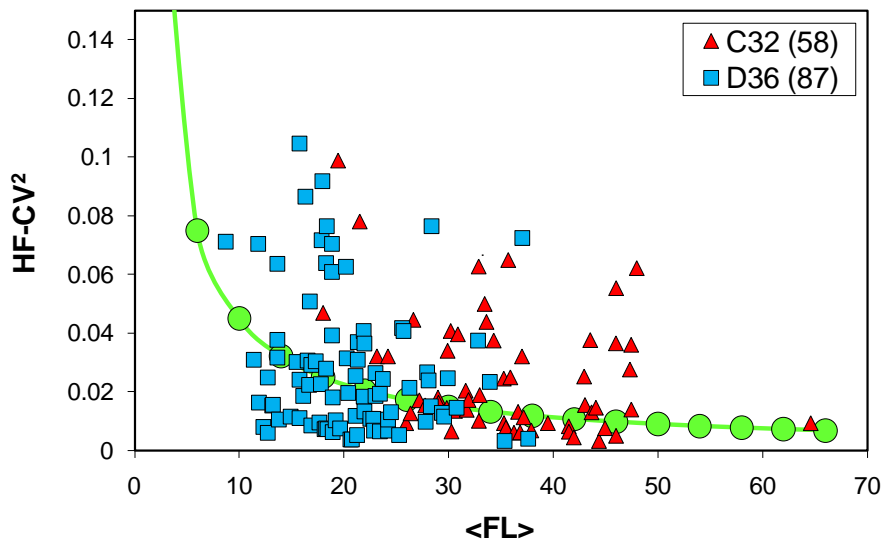


Figure 4.11 HF-CV² vs. average fluorescence level for LTR d2GFP monoclonal cells C32 (58 cells) and D36 (87 cells). From these measurements the CV² component of the bias vector (green line) was found as 0.45/<fl>.

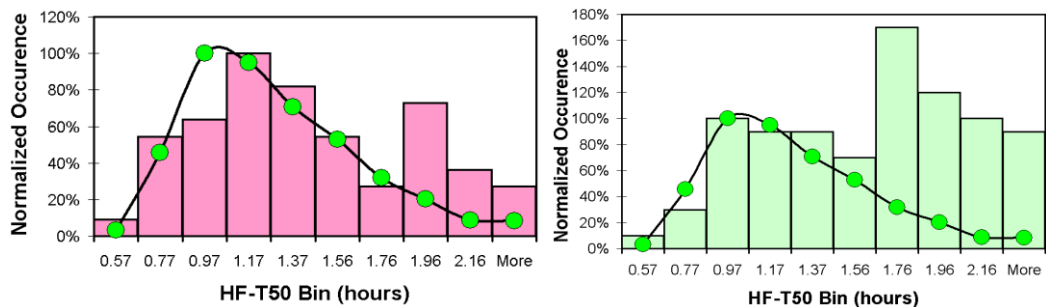


Figure 4.12 Distributions of HF- $\tau_{1/2}$ s measured for LTR d2GFP monoclonal cells C32 (pink) and D36 (green). The bias vector value of HF- $\tau_{1/2}$ was selected as 1.22 hours. The green points show the simulated HF- $\tau_{1/2}$ distribution for constitutive expression and HF- $\tau_{1/2}$ = 1.22 hours, which is seen to fit well with the lower modes of the C32 and D36 HF- $\tau_{1/2}$ distributions. The higher HF- $\tau_{1/2}$ modes are indicative of some transcriptional bursting in these clones.

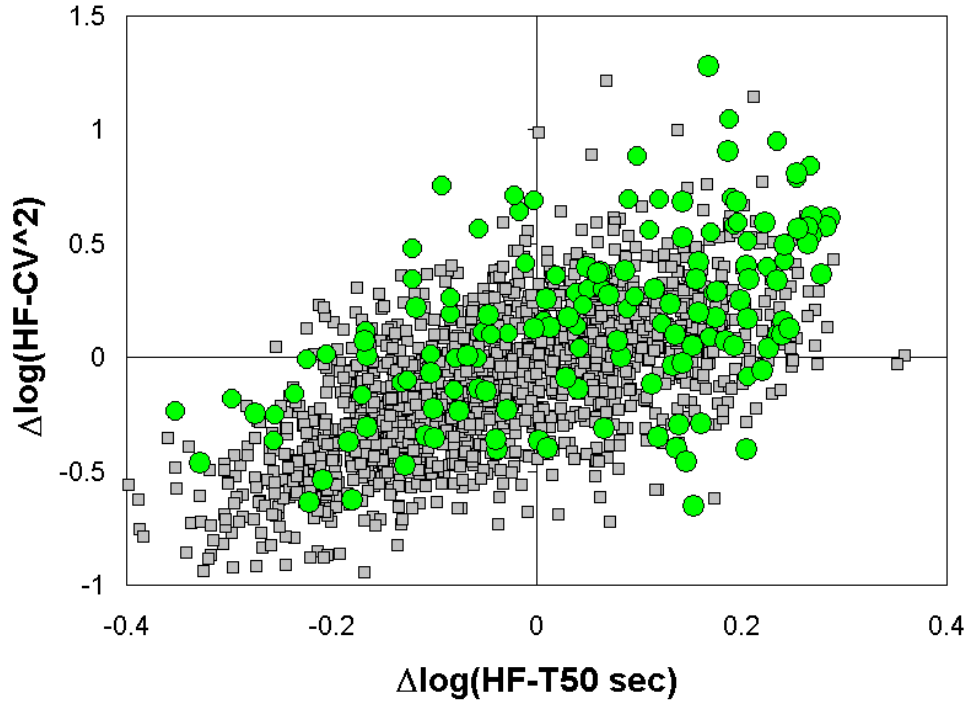


Figure 4.13 Combined noise map for monoclonal cell lines C32 and D36 using the bias vector described in this section (green circles). This noise map closely resembles the simulated constitutive expression noise map (gray squares in the background) seen in figure 4.9.

Using the experimentally determined bias lines the experimental noise map can be completed. The noise vector components of magnitude (ΔCV^2 ; y-axis; Fig. 4.10D) and correlation ($\Delta \tau_{1/2}$; x-axis; Fig. 4.10E) on the noise map (Fig. 4.10F) could be established for each cell (or clone). Finally, the noise map may be used to estimate a noise probability density (NPD) map that shows the likelihood of finding the measured noise of any randomly selected individual cell within a particular region of the map (Fig. 4.10G).

4.3.1.3 Results

The noise maps exhibit shifts to the upper right (Fig. 4.9B-D), indicating significant bursting kinetics at virtually all genomic loci. These data differ significantly from the theoretically expected noise map of fluctuating constitutive expression that lacks transcriptional bursting (Fig. 4.9B, left panel). To determine the probability of finding any individual cell at a particular noise map location and visualize distributions in the noise map, a Noise Probability Density (NPD) map was also constructed (Fig. 4.9C) and revealed a clustering of all genomic integration sites in the upper right region of the noise map – where the two-state model of transcriptional regulation would predict their presence (Figures 4.3 and 4.6 and sections 3.4.1 and 4.1.2.1). The NPD map allows convenient comparison between the measured promoter noise maps to the theoretical constitutive map and the calculated difference (Fig. 4.9D) provides a measure demonstrating the lack of constitutive expression across integration sites. These data, for both the weak (LTR) and strong (EF1A and UBC) promoters, argue for transcriptional bursting, described by the two-state model (Section 3.4) [10, 29, 74, 99], at the vast majority of expressed genomic loci. Surprisingly, NPD map analysis shows that LTR, EF1A, and UBC all exhibit the same distribution of correlation times and mean correlation time (Figures 4.14 and 4.15), suggesting that the integration site sets the transcriptional burst behavior, although this baseline behavior may be modulated by the specific promoter (see discussion below of the LTR transcriptional stall).

To explore the parameters of the transcriptional bursting, experimental noise maps were calibrated against a library of computationally simulated noise maps that span a spectrum of values for the burst kinetic rate parameter (k) and on fraction (O) (Fig 4.16). This calibration library of two-state expression allows for differentiation between different transcriptional bursting motifs (Fig. 4.16). More details on the simulated calibration library are included in the Appendix. Using a resampling algorithm (see Appendix), the noise maps in CV^2 - $\tau_{1/2}$ space were used to determine the prevalence of different bursty behaviors in O - k space (Fig 4.17). High scores indicated that an O - K pair represented the experimental dataset well, while a low score indicated an O - K pair that is

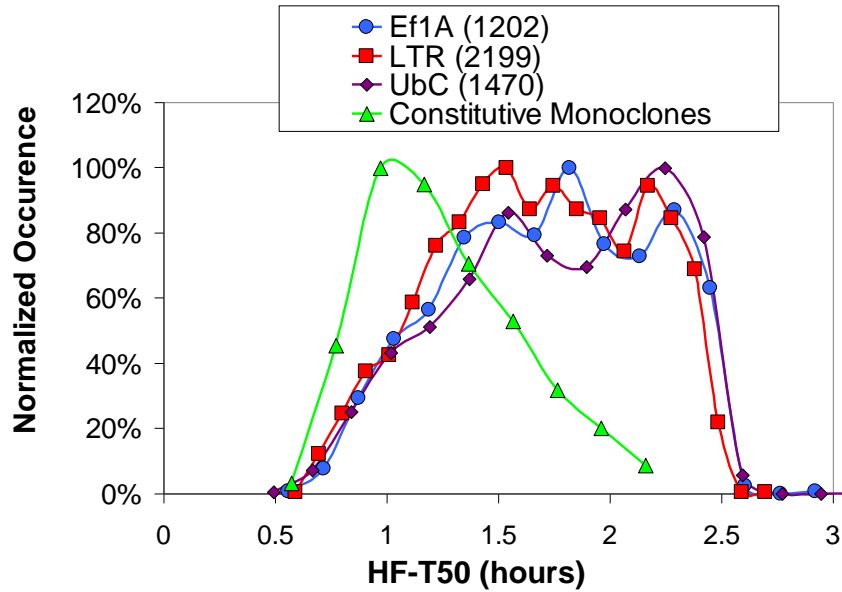


Figure 4.14 Equivalent polyclone correlation time distributions for the 3 promoters measured. The three polyclone correlation distributions are shifted to bursty correlation time ranges compared to the constitutive monoclones (green trend, and more on monoclonal measurements later). Mean correlation times for the three distributions are 1.65 hours for the LTR d2G poly + nothing, 1.7 hours for Ubc d2G poly + nothing, and 1.66 hours for Ef1A d2G poly + nothing.

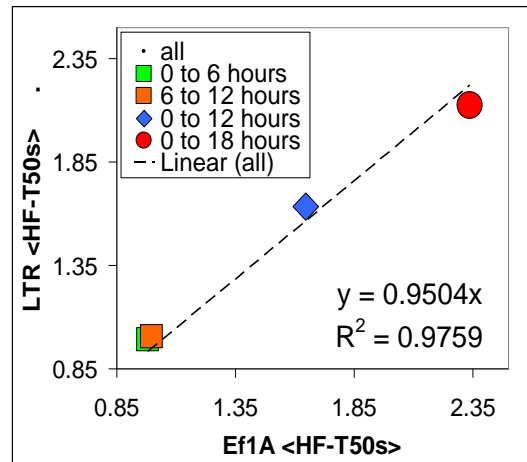
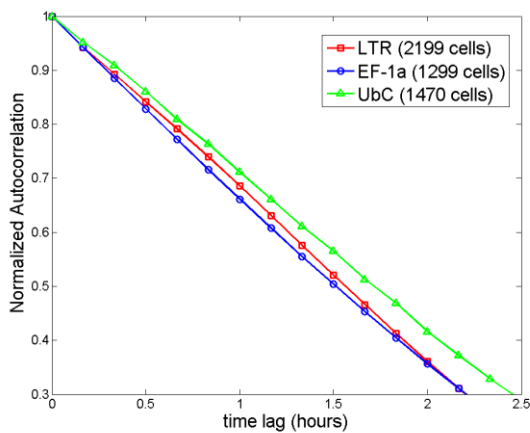


Figure 4.15 Equivalent normalized composite autocorrelations for the 3 promoters. The three polyclone correlation functions and half-correlation times were found equivalent (or very close to equivalent) for 6, 12, and 18 hour experiment durations.

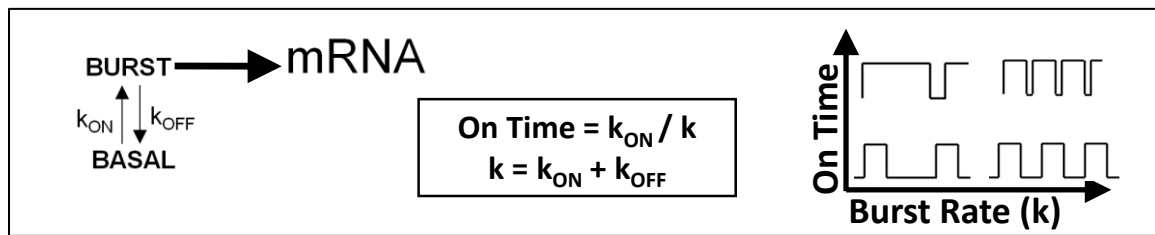


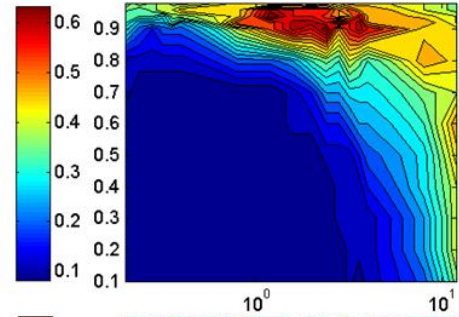
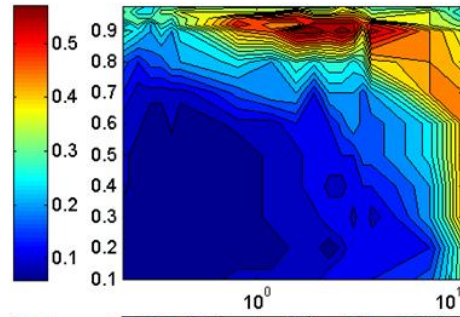
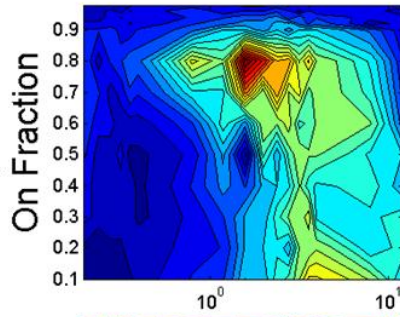
Figure 4.16 A model of two-state transcriptional bursting. The two-state ‘bursting’ model where the promoter fluctuates between an active expression state that generates multiple mRNAs (i.e. bursts) per unit time and a non-expressive basal state. Two parameters, burst kinetic rate (k) and duration in the ‘on’ state (O), can be varied to account for different integration site and promoter behaviors corresponding to different two-dimensional calibration map positions (right panel).

Table 4.1 Characterization of Transcriptional Burst Landscapes

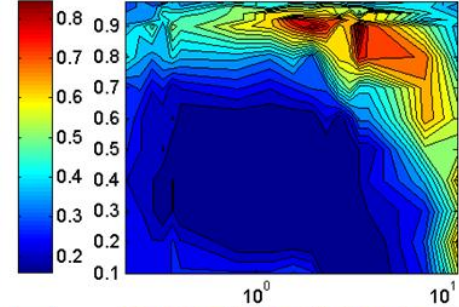
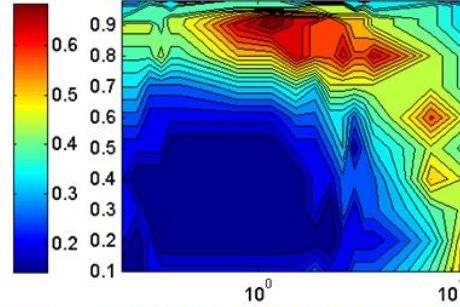
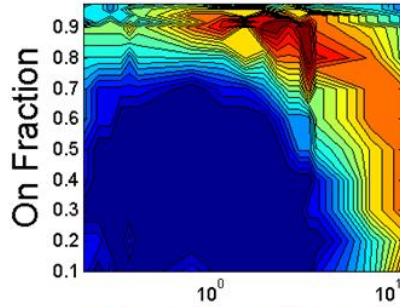
Location in Figure 4.10	Experiment	Highest Scoring O- k range	Landscape movement and remarks
Upper left	Ld2G poly + nothing	$k = 2 \text{ Hr}^{-1}$, $O = 0.8$	
Middle left	Ld2G poly + TNF α	$k = 0.4-8 \text{ Hr}^{-1}$, $O = 0.8-0.9$	Increase in k_{on} increases O and k
Upper center	Ef1Ad2G poly + nothing	$k = 0.8-8 \text{ Hr}^{-1}$, $O = 0.9$	
Middle center	Ef1Ad2G poly + TNF α	$k = 0.8-8 \text{ Hr}^{-1}$, $O = 0.8-0.9$	Reduced expression decreases O and k by reduction of k_{on} , ~25% of sites with NF-Kb get shifts of increased k_{on} (depleted region in difference map, Middle low)
Right column	UbC + nothing and + TNF	Similar to Ef1A case	Similar effects to Ef1A case.

Figure 4.17 Modulations of genomic transcriptional bursting landscape by integration site, promoter type, and signaling molecules. (left) O-K landscapes for the LTR polyclonal experiment, upper-left, LTR + TNF α , middle-left, and the difference landscape between LTR+TNF α minus LTR, lower-left (middle) Same convention as left column except with the Ef1A promoter experiments (right) Same convention of the left 2 columns except with the UbC promoter experiments.

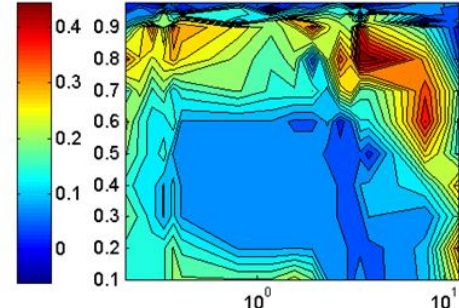
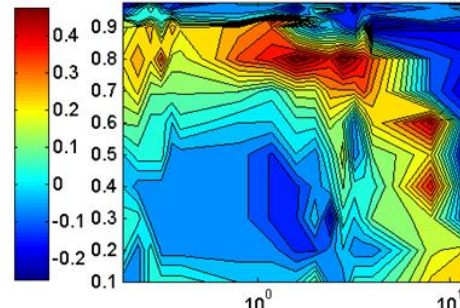
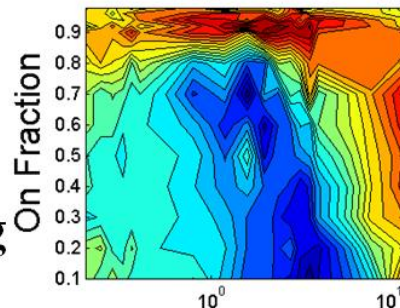
uninduced



+ $\text{TNF}\alpha$



**+ $\text{TNF}\alpha$
minus
+ nothing**



K (Hr^{-1})

K (Hr^{-1})

K (Hr^{-1})

unlikely to have contributed to behavior seen in the experiment.

The resampling algorithm revealed specific O-K landscapes for the uninduced LTR, Ef1A, and UbC promoters (Fig. 4.17 and Table 4.1). LTR exhibited enrichment at $O=0.8$, $K = \sim 2 \text{ Hr}^{-1}$, while Ef1A and UbC exhibited a wider K range at a higher on fraction of $O=0.9$. Looking at the data in $O-f_B$ space (Fig. 4.19, and section 3.4), it appears that although different in on fraction ($O=0.8$ versus 0.9), the frequency range of LTR and Ef1A exhibited enriched bursting frequencies in the $\sim 0.2\text{-}0.6 \text{ Hr}^{-1}$ range. Much like the very similar distributions of $\tau_{1/2}$ times, this common burst frequency band suggests that the transcriptional bursting kinetics are set by the integration site. It would seem that the major effect of the LTR promoter is to lower the on fraction.

Understanding this effect requires a closer look at the function of the LTR promoter. As described in chapter 3, the Tat +AR works by relieving stalled transcription from the LTR promoter. In the absence of Tat, transcription stalls after transcription of a small portion (known as the TAR sequence, which is the Tat binding site) of the gene. This transcriptional stall is due to a nucleosome, with high affinity for a stretch of DNA at the nuc-1 position within the LTR, blocking RNAP II elongation (Fig. 4.17 and 4.19), thereby delaying the start of the LTR transcriptional burst (Fig. 4.18). RNAPII stalling has been recently reported as widespread across the genome and of importance to transcriptional regulation [118]. In the absence of Tat, the stall is relieved either by RNAP II falling off the DNA strand and terminating transcription, or by reading through and completing transcription. This latter case would describe how the initial burst of Tat would be produced in an infected cell after integration of the viral genes into the genome.

To illustrate the effect of the LTR transcriptional stall burst behavior, here the EF-1A and LTR promoters are modeled with the same base transcriptional bursting behavior differing only in the transcriptional stall of the LTR promoter. Ef1A transcription is modeled with stochastic burst that on average are on for a duration of $\overline{\tau_{on}}$ followed by off periods of average duration of $\overline{\tau_{off}}$. The LTR transcriptional stall is modeled as a stochastic delay (average duration = $\overline{\tau_d}$) between the leading edge of the transcriptional

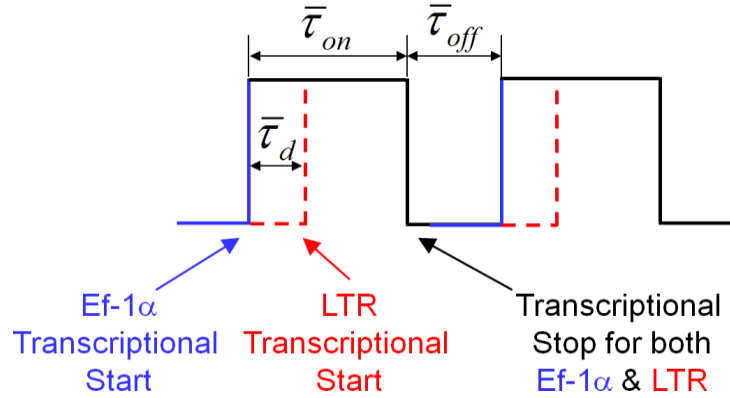


Figure 4.18 LTR promoter transcriptional start is delayed by stalling of RNA polymerase.

burst and the actual start of LTR transcription (Figure 4.18 above). To simplify the analysis, this model uses a lumped delay at the beginning of the transcriptional burst. Using this approximation,

$$\begin{aligned}
 O_{EF-1\alpha} &= \frac{\overline{\tau_{on}}}{\overline{\tau_{on}} + \overline{\tau_{off}}} \\
 O_{LTR} &= \frac{\overline{\tau_{on}} - \overline{\tau_d}}{\overline{\tau_{on}} + \overline{\tau_{off}}} = O_{EF-1\alpha} - \frac{\overline{\tau_d}}{\overline{\tau_{on}} + \overline{\tau_{off}}}, \quad (4.5)
 \end{aligned}$$

and the major effect of the transcriptional stall is seen as a reduction in the measured O for LTR compared to EF1A.

Although the period $(\overline{\tau_{on}} + \overline{\tau_{off}})$ and therefore f_B remains unchanged by the stall (assuming the stall time does not exceed $\overline{\tau_{on}}$), the stall may affect k . Using the model above

$$\begin{aligned}
k_{on} &= \frac{1}{\overline{\tau_{off}}} \\
k_{off} &= \frac{1}{\overline{\tau_{on}}} \\
k_{on_s} &= \frac{1}{\overline{\tau_{off} + \tau_d}} \\
k_{off_s} &= \frac{1}{\overline{\tau_{on} - \tau_d}}
\end{aligned} \tag{4.6}$$

where the s subscript indicates the effective values for k_{on} and k_{off} including the transcriptional stall. For this case,

$$k_s = k_{on_s} + k_{off_s} = \frac{\overline{\tau_{off} + \tau_{on}}}{(\overline{\tau_{on} - \tau_d})(\overline{\tau_{off} + \tau_d})}, \tag{4.7}$$

and

$$k_s \approx k \tag{4.8}$$

if $\overline{\tau_d}$ is small compared to $\overline{\tau_{on}}$ and $\overline{\tau_{off}}$.

Using equation 4.5 above and the resulting enrichment at $O=0.9$ and 0.8 for Ef1A and LTR experiments without induction, respectively, it is possible to estimate a lower limit of the stall providing a 10% reduction in the on fraction (O_{LTR}) or

$$\overline{\tau_d} \approx \frac{1}{10} \cdot (\overline{\tau_{on}} + \overline{\tau_{off}}) \tag{4.9}$$

The measured LTR f_B (figure 4.19) spectrum does not extend as high in frequency as the EF1A f_B spectrum, which may indicate cases where $\overline{\tau_d}$ has become significant compared to $\overline{\tau_{on}}$ (Fig. 4.19), and skipped bursts (bursts where transcription remained stalled for the entire on time) are causing fast bursting to appear to be slow bursting (Fig. 4.19).

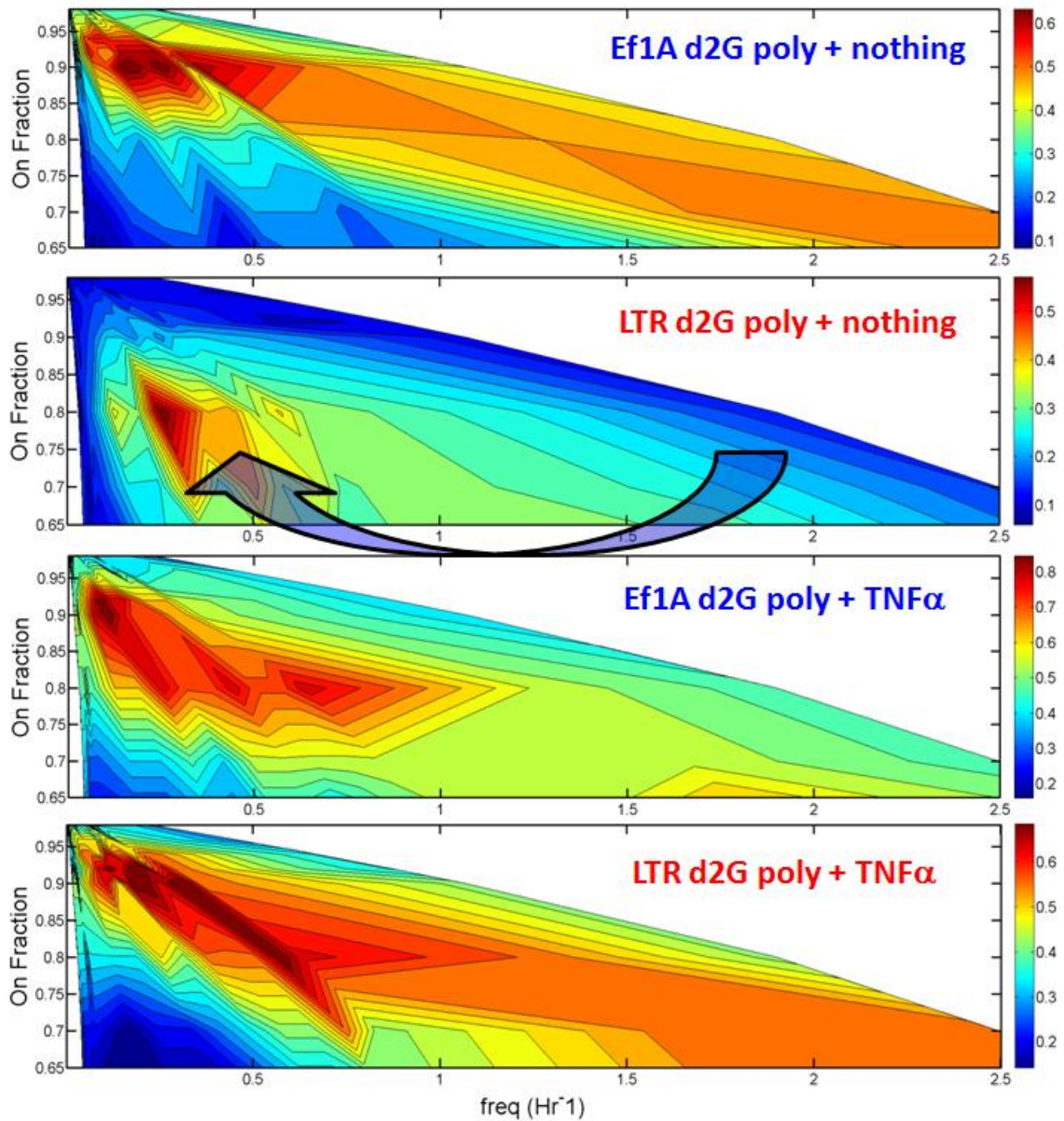


Figure 4.19 A robust genome-wide frequency-band of transcription. Experimental O-K landscape represented with burst frequency ($\text{freq} = f_B = O * k_{off}$) yields an enrichment of frequency band range of ~ 0.2 - 0.6 for the Ef1A and LTR promoters. The blue arrow in the LTR case shows evidence of high frequency pulse skipping. Upon adding $\text{TNF}\alpha$ the frequency band gets widened to $\sim 1 \text{ Hr}^{-1}$.

Transcriptional burst modulation with a signaling molecule

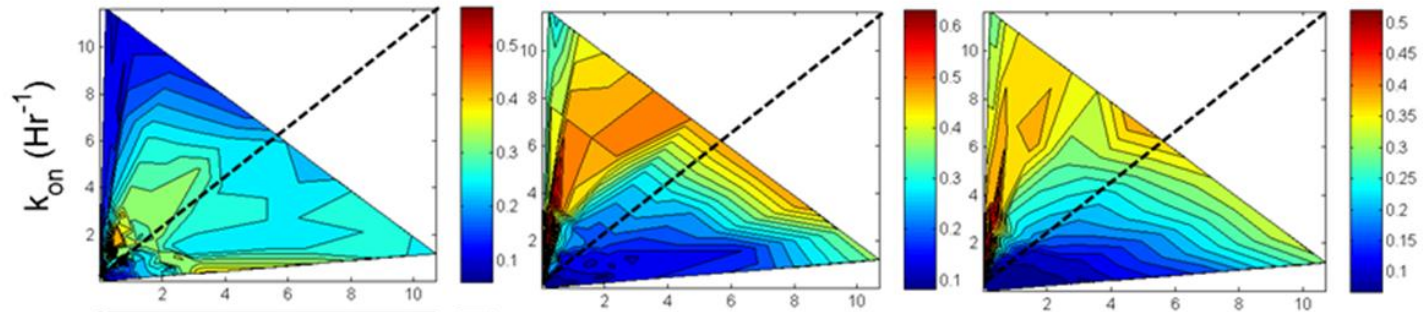
The next experiment examined how cell-signaling molecules that modulate gene expression influence the global transcriptional-burst landscapes. For these experiments, the signaling molecule Tumor Necrosis Factor alpha (TNF α), which enhances expression across a broad range of promoters by stimulating recruitment of a p65-RelA heterodimer to Nuclear Factor κ B (NF κ B) binding sites was used [119]. The HIV LTR provides a convenient analysis system since it encodes multiple NF κ B sites and is potently activated by TNF α across all integration sites (Fig. 4.21 of LTR promoter) [90]. The experiment resulted in an obvious LTR O-k landscape shift to higher O and k by increasing k_{on} (Fig. 4.17, left column, and Fig. 4.19).

In contrast to the LTR promoter, the EF1A and UbC promoters do not contain NF κ B binding sites, and TNF α would not be expected to have a direct effect on their expression. Surprisingly, the noise maps seem to indicate that both O and k are reduced for these promoters in response to TNF α (Fig. 4.17 and 4.19), a result that appears to be connected to the shared resources and plasticity themes of this thesis. NF κ B binding sites are found at ~25% of human promoters, and TNF α would be expected to up regulate the expression of these genes, thereby consuming more of the shared resources of the cell. As a result, fewer of these resources would be available for the expression of genes controlled by the EF1A and UbC promoters. This is seen in the TNF α mediated reduction of the ‘on’ time of the EF-1 α promoter bursts (Fig. 4.17 middle column, O from 0.9 to 0.8), resulting in a slight decrease in EF-1 α expression after TNF α exposure (Fig. 4.22).

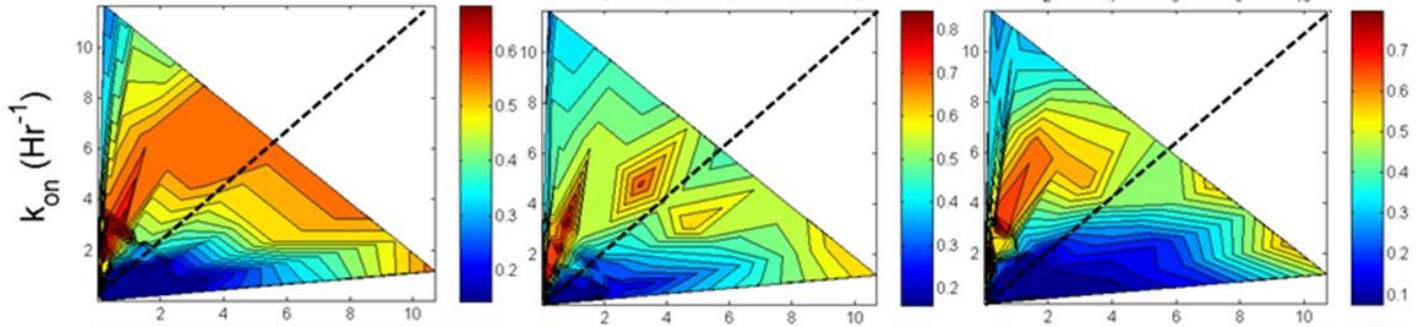
The global analysis of gene-expression fluctuations demonstrates that idealized constitutive gene expression, where promoters continuously emit transcripts over time, is rare in the human genome. A transcriptional burst frequency band with pulses every ~2-5 hours was detected and modulated using an exogenous chemical inducer (TNF α) to higher frequency 1 Hr⁻¹ bursts. The frequency band may have important implications in cellular signaling and function as regulating proteins can only work in specified ‘windows of opportunity’. The transcriptional burst landscape measurement and

Figure 4.20 Modulations of burst transition rates with $\text{TNF}\alpha$. (left) $k_{\text{on}}-k_{\text{off}}$ landscapes for the LTR polyclonal experiment, upper-left, LTR + $\text{TNF}\alpha$, middle-left, and the difference landscape between LTR+ $\text{TNF}\alpha$ minus LTR, lower-left (middle) Same convention as left column except with the Ef1A promoter experiments (right) Same convention of the left 2 columns except with the UbC promoter experiments. Without $\text{TNF}\alpha$ the LTR has a low k_{on} while Ef1A and UbC have higher k_{on} . After addition of $\text{TNF}\alpha$ a marked increase (decrease) in k_{on} is observed for the LTR (Ef1A and UbC) promoter.

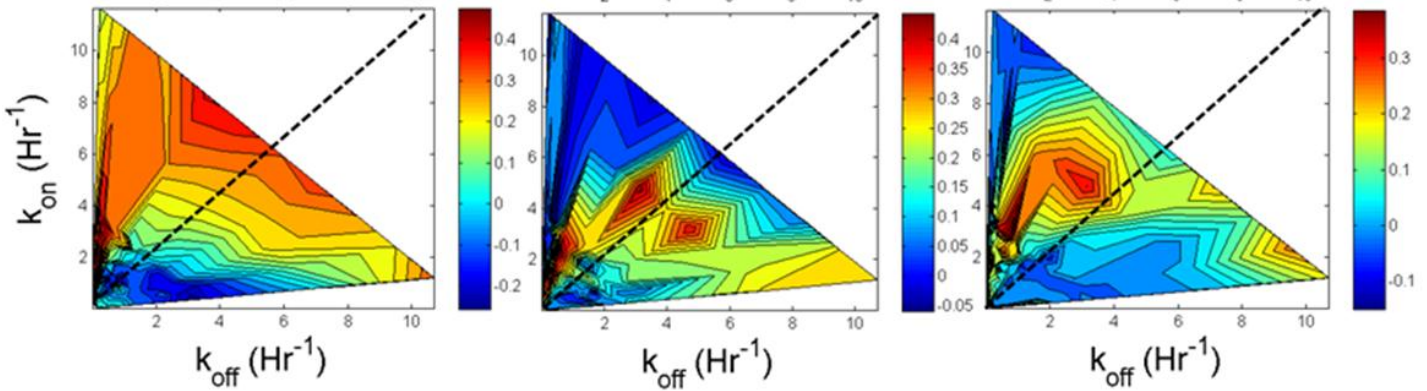
uninduced



+ TNF α



**+ TNF α
minus
+ nothing**



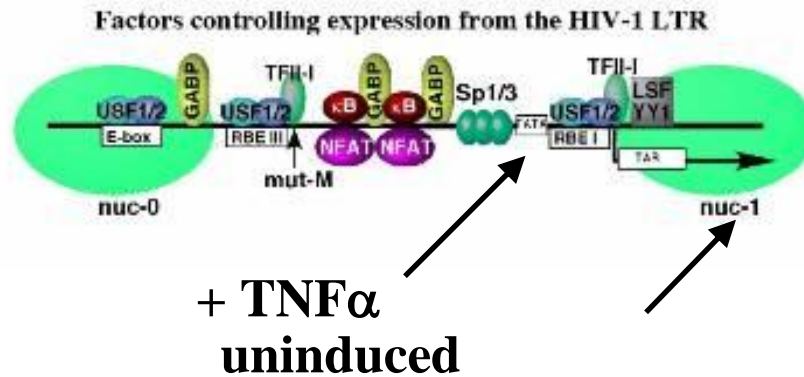


Figure 4.21 Detailed representation of the LTR promoter. The LTR is weak and bursty, it has several activation sites including several NF-K β , the TATA box, and nuc-1 at the transcriptional start site, TSS.

[Figure from http://www.biochem.ubc.ca/fac_research/faculty/sadowski.html]

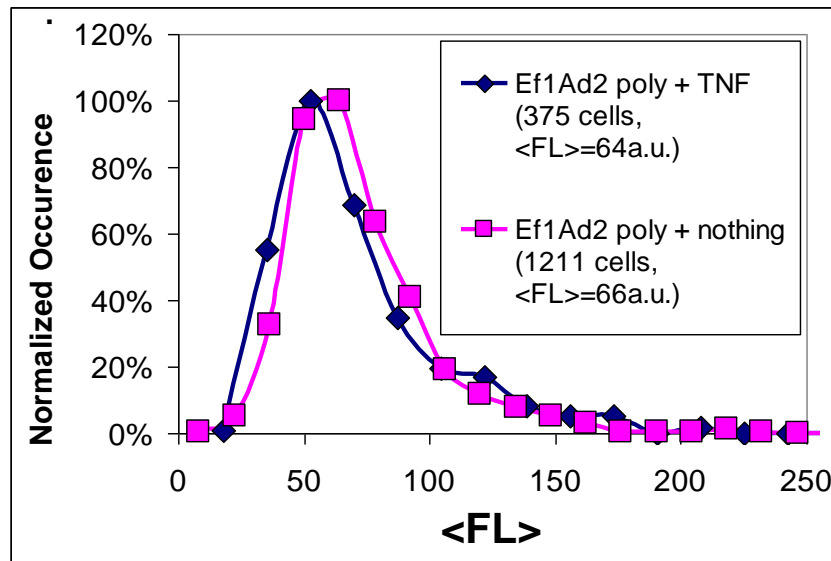


Figure 4.22 EF-1A expression slightly decreases with TNF α addition. Using microscopy the above compares normalized distributions of single cell mean fluorescence (over 12 hours) for Ef1A d2G poly + nothing (pink) and Ef1A d2G poly + TNF α (blue).

modeling for RNAP II stalling in the LTR may have strong biological relevance as this stalling has recently been reported as widespread, and of regulatory importance across the *Drosophila* genome [118]. In addition, the characterized genomic transcriptional bursting background may have implications in the HIV latency problem, the study of information transport in complex systems, and future synthetic biology efforts that integrate into the genome. Overall, these results suggest that different cell-physiological states may exhibit different transcriptional bursting landscapes and noise mapping provides a high-throughput systems-level method to profile these states and enable dynamic, genome-wide measurements on the effects of transcription factors and other biomolecules on cellular state.

4.4 Deterministic implications of the two-state model of transcription

The current chapter focused on characterizing the transcriptional burst dynamics of specific promoters across the integration landscape of the human genome by analyzing gene expression stochasticity signatures (through noise mapping). The LTR promoter is similar in many ways to stress gene promoters as it has a TATA box, nucleosome occupancy impeding its expression at the transcriptional start site, and is considerably noisy (Fig. 4.21). Stress genes have been reported to be excessively noisy in both healthy [33] and stressful [34] environments. As stress genes deterministically respond to perturbations and stressful environments there exists the possibility that their excessive noise in gene expression is in some way coupled to their ability to strongly respond to the randomly timed environment. The following chapter derives a coupling between transcriptional noise and plasticity using the two-state model [10] and uses a comprehensive data mining effort to elucidate organizational insights to this behavior across the genomes of both *S. cerevisiae* (budding yeast) and *E. coli*.

CHAPTER 5: The Coupling of Stochastic and Plastic Response

As explained in chapter 1, the expression of every gene can be decomposed into stochastic and plastic (deterministic) components. The limited resource constraint led to a sharing and *distribution of resources* among the system information carriers, which in turn led to a *conservation and distribution of stochasticity*. That is, noise can be placed in different parts of the system, but it cannot be avoided. The intriguing issue is how is homeostasis maintained at the system (i.e. cell) level, given these large and unavoidable fluctuations at the component (protein, RNA) level? On at least some level, the variability in individual components as described in earlier chapters is understood, yet there is no understanding of how to integrate these fluctuating components together to achieve complex function at the system level.

As a first step in beginning to understand the roles of stochastic and plastic gene expression at the system level, this chapter explores how these two different responses are coupled in convenient model organisms such as *Saccharomyces cerevisiae*, which has been characterized as a system to such a level as to allow the investigation of the noise structure-function relationships sought here. This chapter begins with a derivation of the relationship between noise and plastic response during transcriptional regulation with slow gene activation/repression processes and then presents an in-depth data mining effort covering over 10 genome-wide databases in *S. cerevisiae* and *E. coli* to address the above objectives.

5.1 Transcriptional two–state model describes coupling between stochasticity and plasticity

The previous chapter established that two-state transcriptional bursting is a ubiquitous motif of gene expression in the human genome, suggesting that this may be a common gene expression arrangement. Using the two-state transcriptional bursting model of chapter 3

$$\langle p \rangle = \frac{\alpha O b}{\gamma_d} \quad (5.1)$$

Assuming transcriptional control of protein population (i.e. $\frac{b}{\gamma_d}$ remains constant), $\langle p \rangle$ is modulated (e.g. in response to environmental stimuli) through variations in O , α , or both. Neglecting extrinsic noise, the autocorrelation function for transcriptional bursting (section 3.4.1) is

$$\Phi(\tau) \approx \frac{\alpha O b}{\gamma_d} b e^{(-\gamma_p \tau)} + \left(\frac{\alpha O b}{\gamma_d} \right)^2 \frac{(1-O)}{O k} \left(\frac{\gamma_d}{\left[1 - \left(\frac{\gamma_d}{k} \right)^2 \right]} e^{(-\gamma_p \tau)} + \frac{k}{\left[1 - \left(\frac{k}{\gamma_d} \right)^2 \right]} e^{-k \tau} \right) \quad (5.2)$$

and

$$CV^2 = \frac{\Phi(0)}{\langle p \rangle^2} \approx \frac{b}{\langle p \rangle} + C_k \frac{(1-O)}{O} \quad (5.3)$$

where C_k varies between 0 (fast switching) and 1 (slow switching) and is given by

$$C_k = \left(\frac{\gamma_d/k}{\left[1 - \left(\frac{\gamma_d}{k} \right)^2 \right]} + \frac{1}{\left[1 - \left(\frac{k}{\gamma_d} \right)^2 \right]} \right) \quad (5.4)$$

The noise magnitude has a shot-noise term ($b/\langle p \rangle$) [10] which is dominant for high O , low $\langle p \rangle$, and $k \gg \gamma_p$ where C_k approaches 0, and a burst or operator noise component ($C_k \frac{(1-O)}{O}$), which dominates at low O , $k \ll \gamma_p$ where C_k approaches 1, and high $\langle p \rangle$. The CV^2 component of the regulatory vector for this circuit is simply

$$\Delta CV^2 = CV^2 - CV_{shot}^2 \approx \frac{b}{\langle p \rangle} + C_k \frac{(1-O)}{O} - \frac{b}{\langle p \rangle} = C_k \frac{(1-O)}{O} \quad (5.5)$$

Additionally, Newman *et al.* [102] have reported experimentally measured noise for *S. Cerevisiae* and they report a term called DM, which here is taken to be approximated by the relationship

$$DM = CV - CV_{shot} \approx \sqrt{\frac{b}{\langle p \rangle} + C_k \frac{(1-O)}{O}} - \sqrt{\frac{b}{\langle p \rangle}} \quad (5.6)$$

Plasticity (*Pl*) is a measure of how responsive expression is to external stimuli. For a particular gene, here it is quantified as a ratio between the highest and lowest levels of expression for that gene, or

$$Pl = \frac{\langle P \rangle_{\max}}{\langle P \rangle_{\min}} \quad (5.7)$$

Assuming, for the moment, that different expression levels are achieved only through modulation of *O*,

$$Pl = \frac{\frac{\alpha O_{\max} b}{\gamma_d}}{\frac{\alpha O_{\min} b}{\gamma_d}} = \frac{O_{\max}}{O_{\min}} \quad (5.8)$$

For maximum plasticity, $O_{\max} \rightarrow 1$ and

$$Pl \approx \frac{1}{O_{\min}} \quad (5.9)$$

It is then possible to show (see the Appendix) that DM and plasticity are related by:

$$\boxed{DM \approx \sqrt{(Pl - 1)}} \quad (5.10)$$

There are ways other than modulating O to generate plasticity. However, notice from equations 5.5 and 5.6 that O is the only one of these parameters that affects the excess noise (i.e. noise in excess of shot noise).

5.2 Distribution and regulation of stochasticity and plasticity in *Saccharomyces cerevisiae*

5.2.1 System-wide noise

One of the most extensive genome-wide noise magnitude studies to date was reported by Newman *et al.* in 2006 [33]. Using both rich (YPD) and minimal (SD+) growth media, Newman *et al.* measured the noise in the populations of more than 2000 proteins in yeast using flow cytometry and found that most proteins displayed the inverse relationship between protein abundance and noise as predicted by Poisson statistics (shot noise component of equation 5.3).

As mentioned above, the relationship between noise and plasticity focuses on proteins that exhibit noise that exceeds the level predicted by the protein abundance. Newman *et al.* quantified this ‘excess’ noise as the difference (DM) between the measured noise and the noise expected at that protein abundance. In this study the minimal medium (SD) DM measurements reported by Newman *et al.* are used.

5.2.2 System-wide plasticity

While stochasticity is defined as variation that occurs irrespective of the presence of a stimulus, plasticity is defined with respect to variation in gene expression that occurs in concert with stimuli. We begin by considering a system consisting of genes $i=1, 2, \dots, I$ whose transcription level (as measured by mRNA abundance) is determined in environments $j=1, 2, \dots, J$. In microarray data, the *relative transcriptional response* of gene i to environment j , e_{ij} , is quantified as

$$e_{ij} \equiv \log_2 \left(\frac{m_{ij}}{m_{i0}} \right) \tag{5.11}$$

where m_{i0} is the mRNA level in some reference (unstressed) environment. For reviews of microarray technology and applications see [120-122].

To define *environmental transcriptional plasticity*, Pl_i , of gene i , the following expression is used:

$$Pl_i = 2^{(\max_5(e_{ij}) - \min_5(e_{ij}))} = \frac{m_{i_max}}{m_{i_min}}, \quad (5.12)$$

where

$$\max_5(e_{ij}) \equiv \left(\sum_{Top5} e_{ij} \right) / 5 \quad (5.13)$$

$$\min_5(e_{ij}) \equiv \left(\sum_{Bottom5} e_{ij} \right) / 5$$

where m_{i_max} and m_{i_min} are average mRNA levels derived from the largest and smallest 5 mRNA ratio values across all environments. An averaging is used to reduce variability due to outliers in the genetic response. The max/min ratio in the definition is reached by subtracting the two e_{ij} average values so that the reference mRNA level, m_{i0} , cancels out.

According to this definition of plasticity there are 3 ways to reach large values of plasticity: (1) induction or an up-regulated response to stimuli, (2) repression or a down-regulated response to stimuli, or (3) induction in response to some stimuli and repression in response to others. Each high plasticity gene fits into one of these three categories. Low plasticity implies that expression either remains fairly constant regardless of environmental condition, or that any significant changes in expression occur in very few environments.

Measures of plasticity are calculated using a report from Gasch *et al.* that describes an extensive genome-wide stress response microarray study in yeast [123]. The study employed 13 stressors that included heat shock, hydrogen peroxide, amino acid starvation, and nitrogen depletion. All of the stressors were applied at various strengths for a total of 173 measured environments. Each environment had a separate microarray

containing the ~6200 genes of the yeast genome (vertical columns of Fig. 5.1). In brief, when the budding yeast is subjected to stress, it stops growth by repressing ~600 genes and reallocates its expression capacity to the induction of ~300 stress response genes. This large-scale switching between the growth and stress states was coined the environmental stress response (ESR) and is executed by highly coordinated gene regulation (Fig. 5.2). ESR execution is independent of stressor type as seen by the variety of stressors used in the study (Fig. 5.1).

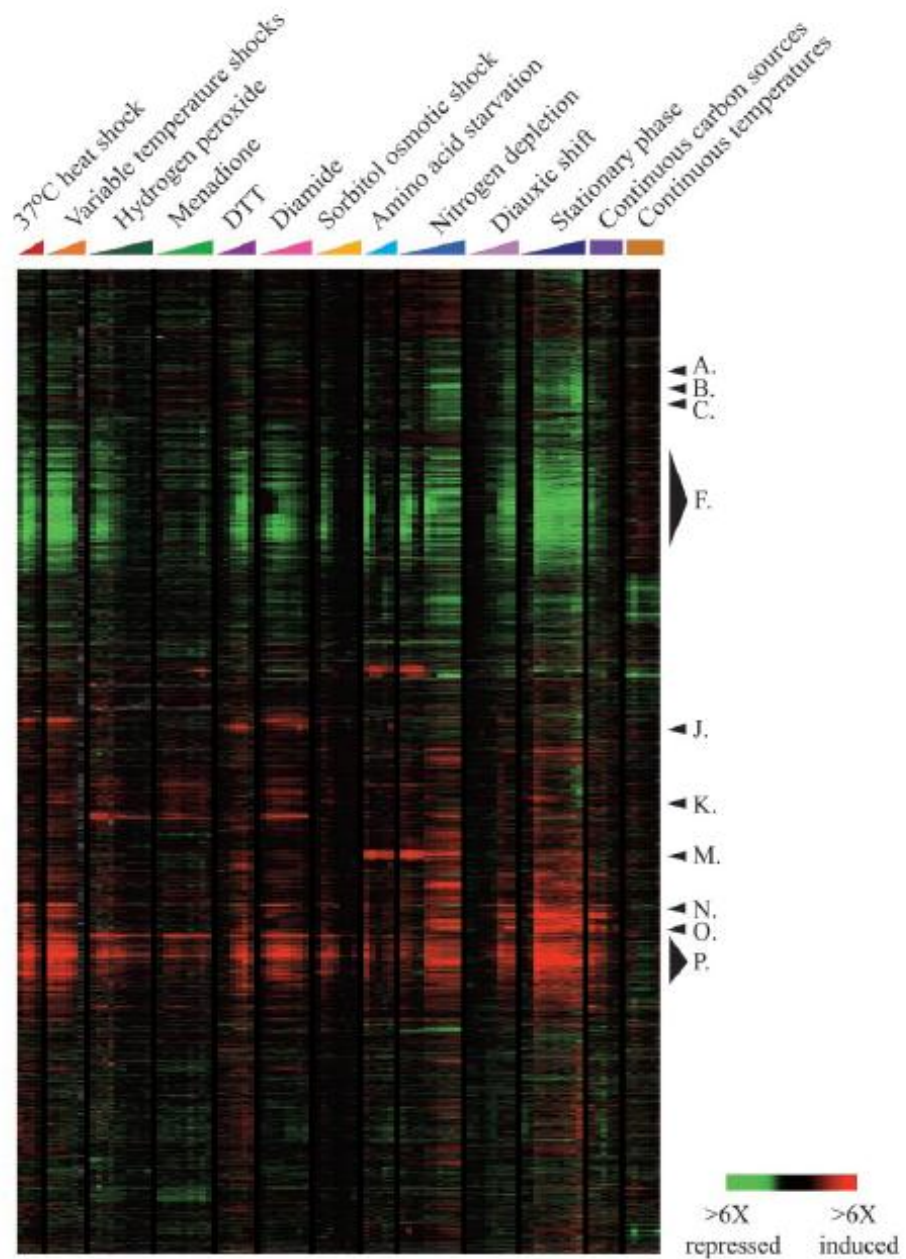


Figure 5.1 The yeast Environmental Stress Response (ESR). Genomic expression patterns in response to environmental changes. mRNA expression changes by microarray of 6200 genes (columns) of the budding yeast genome responding to 13 different stressors of varying strengths for a total of 173 environments (rows). There are 900 genes participating in the ESR (green and red regions). [Figure adapted from Gasch et al., 2000[123]]

5.2.3 Regulatory arrangements that control noise and plasticity

To begin exploring the relationships between stochasticity and plasticity, three different gene regulatory arrangements that control expression were considered (Fig. 5.2). Two of the arrangements are encoded in the DNA regions adjacent to the gene being expressed and the third regulatory arrangement is related to protein-DNA interactions and to the two main co-activation complexes found in yeast (Fig. 5.2).

The first regulatory arrangement is the presence or absence of a TATA box (5'-TATAAA-3' or other variant), which is a core promoter motif present in ~20% of all *S. cerevisiae* genes [124] and in ~24% of human genes [125]. The TATA box is a competitive binding site for histones and transcription factors and is generally associated with greater expression noise [33, 126]. A study by Basehoar *et al.* [124] was used to determine which yeast genes contain TATA boxes and which do not.

Next the chromatin structure surrounding the gene of interest was accounted for from the nucleosome occupancy pattern near the transcriptional start sites (TSS). A high resolution atlas of yeast nucleosome occupancy patterns was recently reported by Lee *et al.* in 2007 [127]. This study covered over 80% of the yeast genome and characterized nucleosome occupancy patterns at +/- 400 bp from the TSS of each gene. Here both occupancy upstream and downstream of the TSS was considered. Occupancy patterns were aggregated into four occupancy motifs (identified here as clusters 1-4; Fig. 5.3) Lee *et al.* generated with *k*-means clustering using the Euclidean distance metric. The four occupancy motifs were used to define the occupancy pattern for each gene.

Finally, the third arrangement, which is related to global protein-DNA regulation, considers the co-activation complexes involved in the coordinated regulation of the yeast stress response versus growth and housekeeping genes. TFIID is known to regulate ~90% of the measurable genome while the SAGA complex only regulates ~10% of the measurable genome [128]. Data reported by Huisinga *et al.* [128] is used to identify the co-activation complexes responsible for regulating the expression for each of the genes considered in this study.

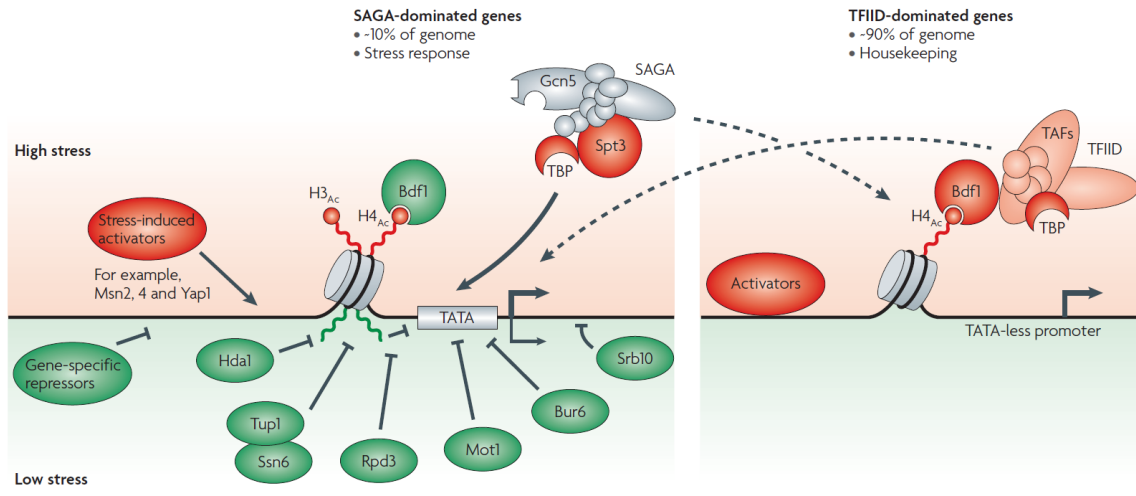


Figure 5.2 Distinct regulatory features of stress and growth related genes. (left) Yeast stress genes are heavily regulated. They are usually activated via the SAGA-complex and contain a promoter TATA box. In addition to a variety of repressors (green) and activators (red) there is competition between factors that acetylate and de-acetylate histones H3 and H4. (right) Housekeeping or growth genes are usually less regulated, TATA-less, and activated by the TFIID complex. The dashed arrows show that for some genes the SAGA/TFIID activation is switched [Figure adapted from [129] and [128]].

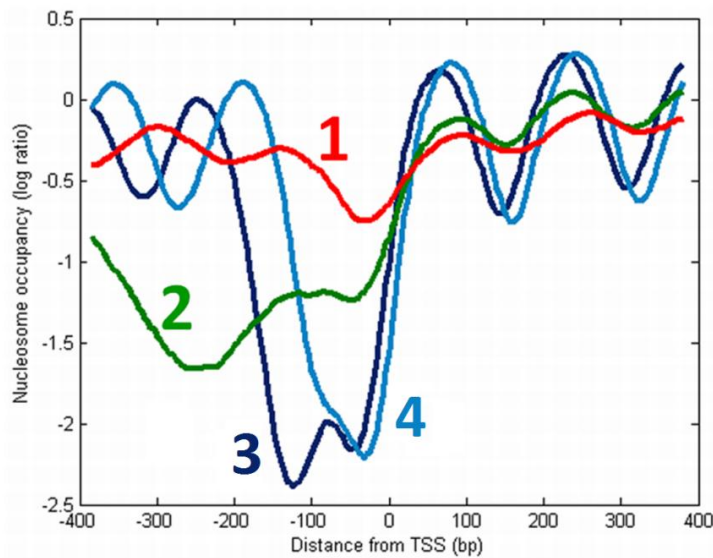


Figure 5.3 Primary nucleosome occupancy patterns reported by Lee *et al* [127]. Clusters 1-4 are labeled next to their respective curve in the plot and were calculated using a K-means clustering algorithm. Of note, at least half the genome has patterns 3 and 4 which have a vacant region just upstream of the transcription start site (TSS). Reproduced with permission from Lee *et al* [127].

5.2.4 Results

5.2.4.1 Noise-Plasticity coupling is widespread across the genome

The first observation accounted for all genes where both noise and plasticity had been measured (2021 genes, Figure 5.4). After clustering and averaging genes in bins of 150 genes in ascending plasticity order, a square-root relationship similar to that derived in section 5.1 emerges with strong agreement (Fig. 5.4, lower, $R^2=0.91$). This implies that noise-plasticity coupling is wide-spread across the genome and different strengths of coupling occur on a gene-by-gene basis (i.e. there is a distribution of coupling strengths across the genome). An additional analysis of genome-wide binning according to the 3 high plasticity categories mentioned above (up-regulation, down-regulation, and both) is provided in the Appendix. There, repressed growth genes follow a low DM-PI coupling trend and induced stress genes follow a high DM-PI coupling trend (in healthy conditions).

5.2.4.2 Noise-Plasticity coupling strength is dominated by regulatory arrangement

To investigate the relationship between noise and plasticity further, the yeast genes were segregated into 4 *main categories* (see section 5.2.3): TATA-SAGA; TATAless-SAGA; TATA-TFIID; and TATAless-TFIID. Each of these were further divided into 1 of 4 *sub-categories* (nucleosome occupancy pattern clusters) such that there were 16 distinct grouping of genes. An average DM and plasticity were calculated for genes where both their TATA/TATAless and SAGA/TFIID architecture were reported in the datasets. These averages did not include any genes for which the architecture was unknown or ambiguous. The averages were also calculated separately for each of the 4 nucleosome occupancy pattern clusters.

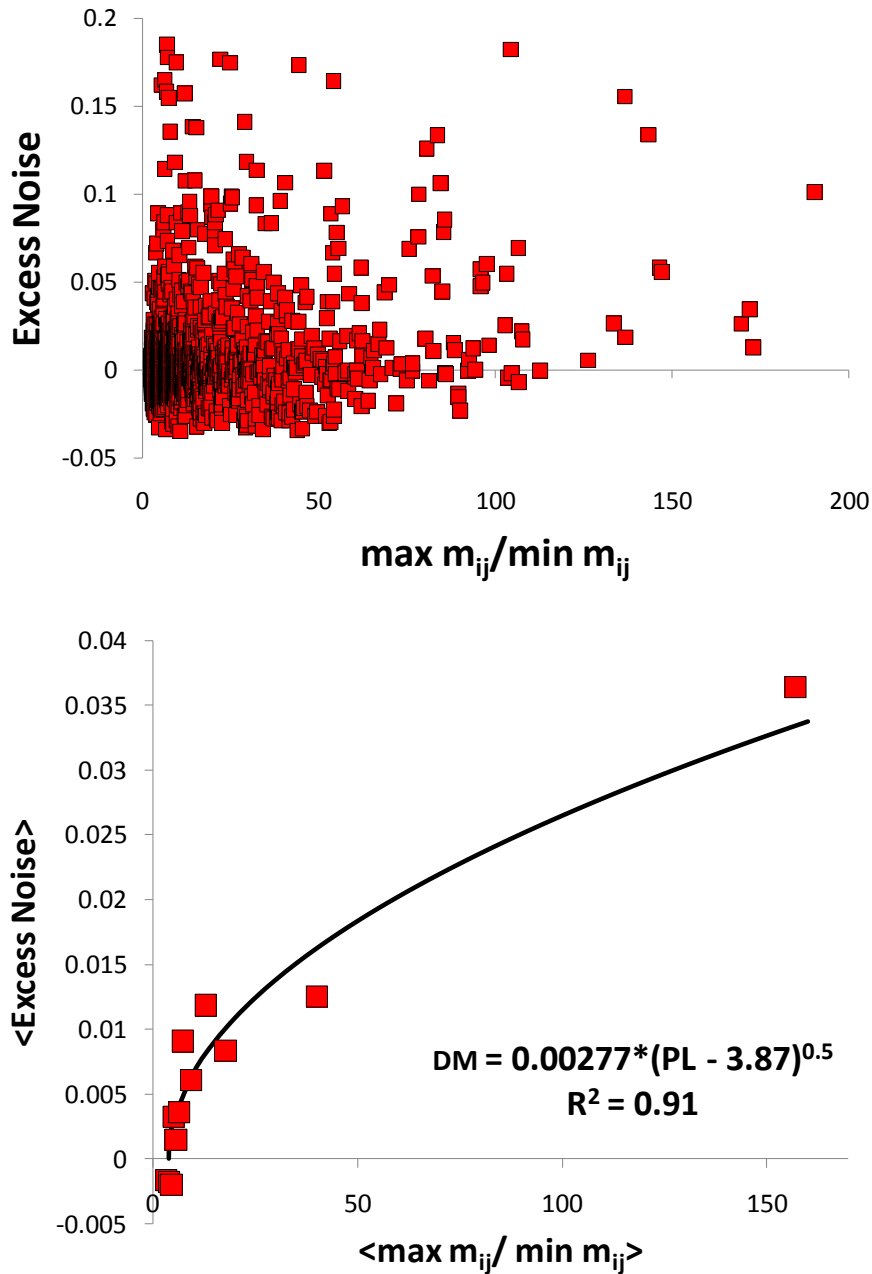


Figure 5.4 Widespread genome-wide noise-plasticity coupling in Yeast. (upper) The genome-wide yeast picture and (lower), clusters of 150 genes yield the expected noise-plasticity coupling. The black model line is consistent with the predicted two-state transcriptional regulation coupling relationship.

Excess noise was strongly dependent on the gene main category with the average DM of TATA-SAGA an order of magnitude greater than that for TATAless-TFIID (Fig. 5.5, upper). This result is not surprising as genes containing a TATA box tend to have large transcriptional bursts that accentuate the noise [33, 126]. TATA-containing promoters are often associated with stress-response functions [124], and it may be that the TATA-generated DM provides a noise-mediated benefit for the response to acute environmental stress [126]. Conversely, TATA-less promoters are often associated with housekeeping functions [124] where noise may be detrimental or have little impact on function. The SAGA or TFIID classification had an even greater effect on DM than TATA, with TATAless-SAGA having higher noise than TATA-TFIID.

For the case of nucleosome occupancy clusters, cluster 1 had ~4x the excess noise as compared to clusters 3 and 4, which had similar low levels of excess noise. Using gene ontology (GO) analysis, Lee *et al* reported an enrichment in stress response genes in cluster 1 [127] consistent with the high excess noise in TATA architectures (Fig. 5.5 upper).

Intriguingly, high noise architectures were also high plasticity architectures, as plasticity follows exactly the same pattern as DM (Figures 5.5 and 5.6, upper). Accordingly, there is a positive correlation between DM and plasticity, consistent with additional findings in recent studies [130, 131].

With respect to nucleosome occupancy cluster and the 16 distinct architecture subgroups (Fig. 5.6, lower), DM and plasticity followed a similar pattern for all main categories except TATAless-TFIID, with clusters 1 and 2 associated with higher DM and plasticity than clusters 3 and 4 (Fig. 5.6). The TATAless-TFIID genes had very low DMs that had little or no correlation with plasticity.

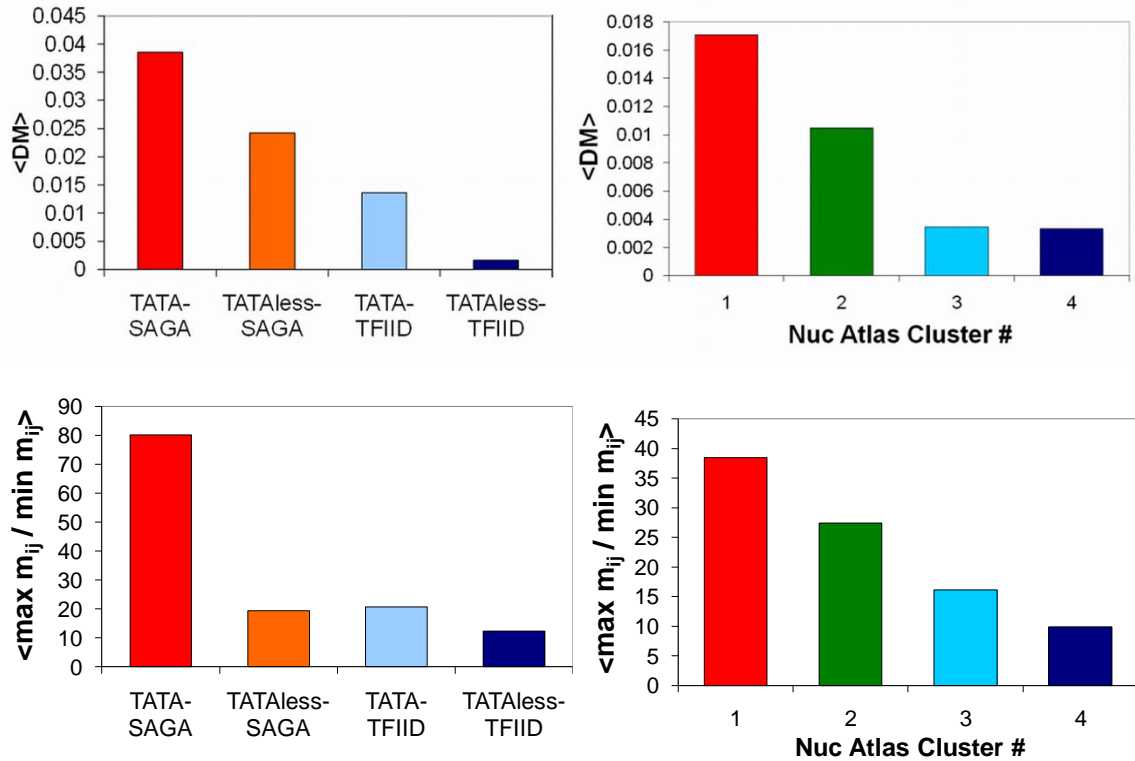


Figure 5.5 Excess noise and plasticity are related and strongly dependent on gene regulatory architecture. **(upper)** Mean excess noise for the 4 main regulatory categories (TATA/TATAless, SAGA/TFIID) and 4 main nucleosome occupancy patterns. **(lower)** Mean plasticity for the 4 main regulatory categories and 4 main nucleosome occupancy patterns. High plasticity architectures share high excess noise and low plasticity architectures share low excess noise.

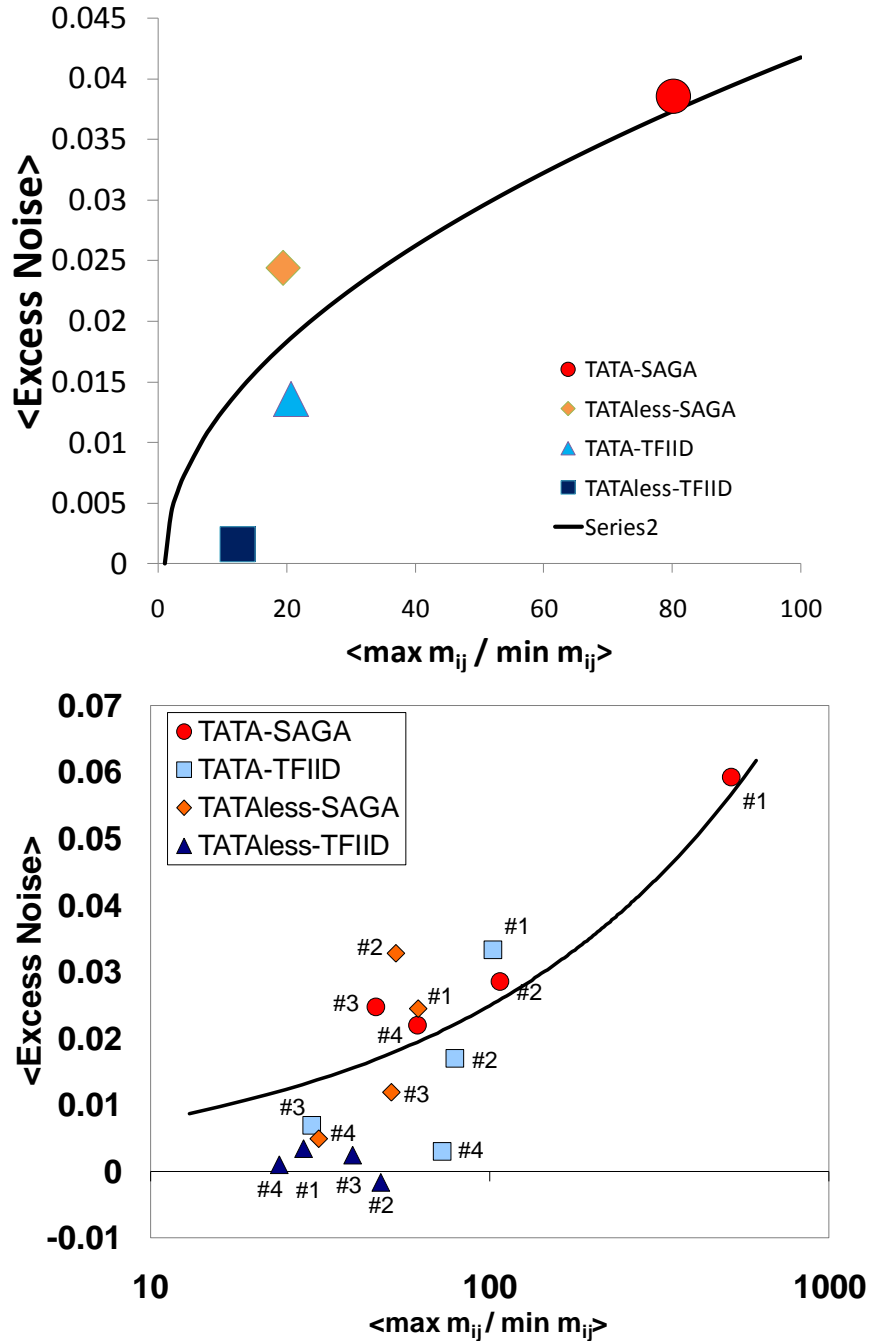


Figure 5.6 Excess noise and plasticity are positively correlated across 16 distinct groupings of genes. (**upper**) The 4 main regulatory arrangement categories have a strong correlation between stochasticity and plasticity. (**lower**) Labeling of the 16 sub-categories including the 4 nucleosome occupancy patterns of Figure 5.3. The nucleosome occupancy motif determines much of the coupling along the model line. The model line in both plots is $DM = 0.0025 \cdot \sqrt{Pl}$.

5.2.5 Discussion

On first inspection, the strong positive correlation between DM and plasticity is surprising as one might have expected a large degree of plasticity (expression varying in a deterministic way in response to environmental signals) to instead be associated with low levels of random variability (Fig. 5.7). This expectation would be consistent with a hypothesis where noise is used as a bet hedging strategy when the optimal expression level is unknown. However, closer inspection indicates that these results and this bet hedging hypothesis are not at odds. The TATA-SAGA genes that have the highest DM and plasticity are often associated with stress-response functions [124], and their plasticity implies that an optimal expression level is known, but only for stressful environments. In the non-stressed environment where DM was measured, the optimal expression level for stress genes is unknown, and noisy expression might provide an anticipatory response – the equivalent of occasionally sending a fire truck past a fire prone building – and the high level of noise is coincident with a high level of uncertainty in the timing of gene expression. There currently exists no genome-wide noise study under stressful environments to further investigate noisy gene transitions into the stressful state. In 2006 Bar-Even *et al.* measured high levels of excess noise in a small number of stress genes with TATA-SAGA architectures in a variety of stressful conditions [34] suggesting that bursty architecture genes are noisy in both healthy and stressful environments, perhaps a constraint of their genetic architecture (it may be that the excess noise is reduced but still present in the deterministically expressed stressful environment).

The relationship between DM and plasticity presented here are consistent with the derivation and predicted relationship from the beginning of the chapter and a gene expression model dominated by two-state transcriptional bursting [10, 14, 29, 74]. The results suggest that each of the distinct major and minor gene regulatory arrangements occupy a specific region of the stochastic-deterministic gene expression space and drive variable noise-plasticity coupling strengths.

On the whole, the heavily studied *S. cerevisiae* enabled the exploration of genome-wide plastic and stochastic gene expression relationships in a complex and

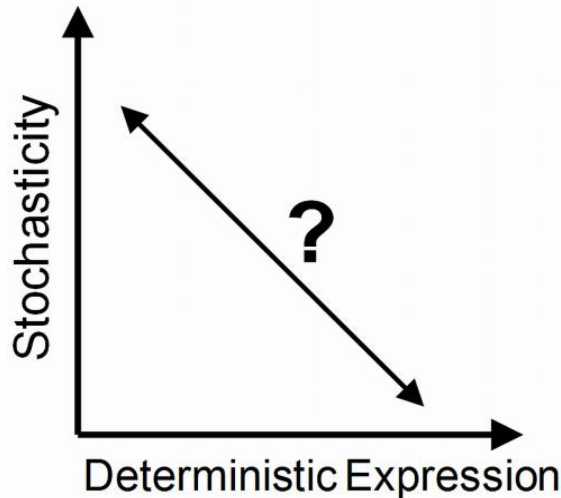


Figure 5.7 Expected inverse relationship between variability and deterministic expression. In this scenario high variability would be counter-productive to strong deterministic gene expression. The relationship explored is inverted because the noise database used is measured in the healthy environment where the timing for the optimal expression level is unknown and random.

highly dimensional system. Among the things learned are: (1) distinct regulatory architectures occupy and function in specified regions of the stochastic versus plastic gene expression space; (2) stochasticity and plasticity are positively correlated for many genes; and (3) stress genes with TATA-SAGA regulatory arrangements display the most noise in unstressed environmental conditions. This behavior is consistent with a two-state transcription burst model where the plasticity is achieved through modulation of the burst duration (on fraction), a condition that generates the greatest excess noise when the gene is expressed at its lowest level. Finally the relationship between excess noise and plasticity agrees well with the derived relationship from the two-state transcriptional bursting model.

Upon closer observation, it appears that genes that are dominated by the SAGA co-activation complex have higher noise-plasticity coupling strengths even without TATA boxes or with low noise nucleosome occupancy clusters 3 and 4. This suggests

that protein-DNA interactions alone may be capable of driving noise-plasticity coupling in a complex system. It is possible to explore this prediction in a prokaryotic bacterium that has no highly compact DNA architectures (chromatin) and where system-wide regulation, signaling, and function are all mediated primarily by protein-protein or protein-DNA interactions. The next section attempts to answer these questions and applies the above investigative reasoning to *E. coli*.

5.3 Distribution and regulation of stochasticity and plasticity in *E. coli*

5.3.1 Introduction

After observing and modeling the co-regulation and control of stochastic and plastic response in yeast, additional biological systems were explored for similar response coupling effects. *E. coli* emerged as a good candidate as it is one of the most studied model prokaryotes with many genome-wide datasets available. Conservation of similar noise-plasticity coupling in *E. coli* is not initially obvious. *E. coli* and *S. cerevisiae* have taken significantly different evolutionary paths in their development as they are on separate branches of the evolutionary (phylogenetic) tree (Bacteria vs Eucaryota). Compared to *S. cerevisiae*, *E. coli*: (1) has no cell nucleus, and therefore transcription and translation may proceed in parallel; (2) has a smaller genome; (3) has no chromatin or highly compact DNA architectures, and therefore does not have the obvious bursty behavior expected from the remodeling of such architectures; and (4) environmental regulation of transcription is mediated by sigma factor subunits in the RNA polymerase complex as opposed to direct regulatory protein – DNA interactions in yeast (or SAGA/TFIID type co-activation complexes).

5.3.2 System-wide noise

In a recent study, Taniguchi, *et al.* reported the construction of a chromosomal yellow fluorescent protein (YFP)-protein fusion library in *E. coli* [32]. They used a microfluidic chip and automated fluorescence microscopy for high throughput single cell screening to quantify both protein and mRNA levels. In the protein case, they successfully imaged ~1000 protein-YFP fusions. In their study, they did not perform any extensive investigation of excess noise, which will be pursued here.

Consistent with the yeast study, DM was defined as the difference between the measured CV for an individual gene circuit and the CV that would have been expected for a protein with the same abundance (see arrows in Fig. 5.8). Figure 5.8 shows the measured noise values supplied with the Taniguchi *et al.* paper [32]. A median line (in red) was calculated and used in excess noise calculations for genes with $\langle P \rangle < 10$, and

from an extrinsic noise floor constant of $CV = \sim 0.4$ for genes with $\langle P \rangle > 10$. Figures 5.9 and 5.10 compare system-wide noise and excess noise ranges for *E. coli* and the budding yeast. Compared to yeast, *E. coli* is small and has lower protein abundances and a higher range of excess noise.

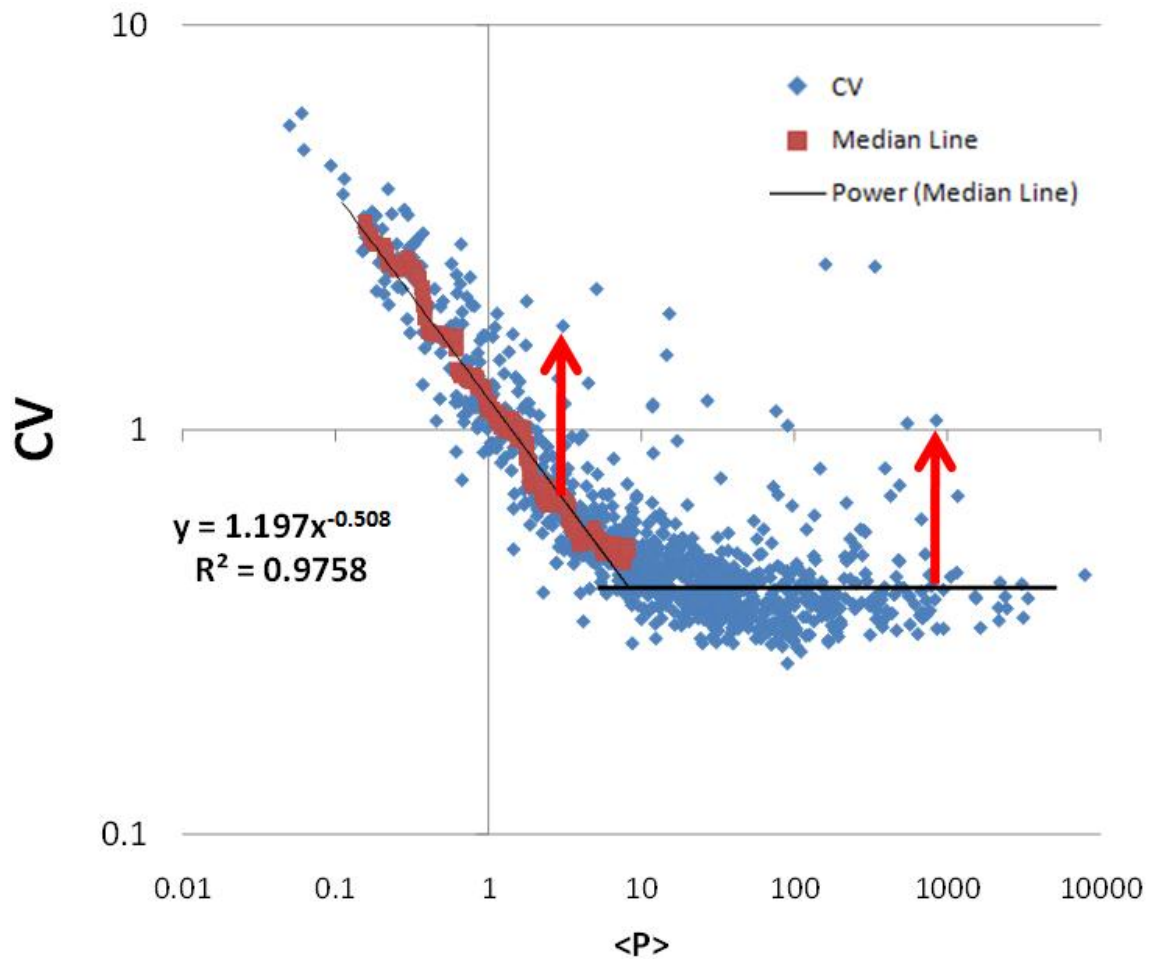


Figure 5.8 Noise in 1000 *E. coli* protein-YFP fusion strains. The above plot uses data supplied with the Taniguchi *et al* 2010 paper [32]. Excess noise was defined from a median line for $\langle P \rangle < 10$, and from an extrinsic noise floor of ~ 0.4 for $\langle P \rangle > 10$ and red arrows depict excess noise values sampled from these two separated regions.

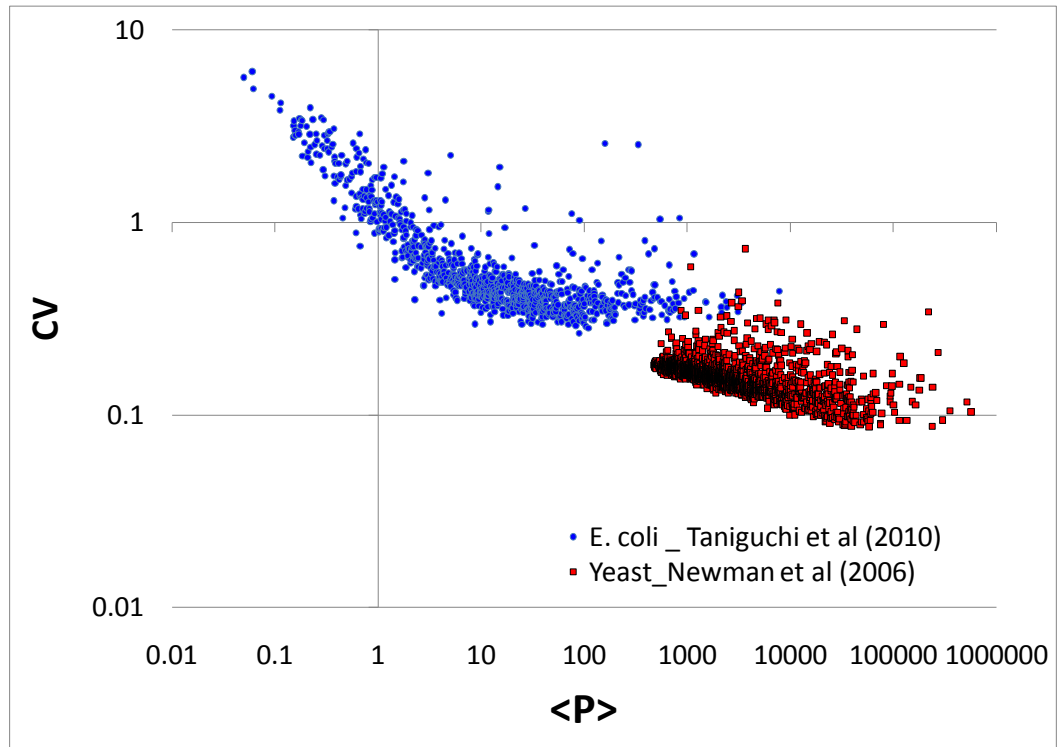


Figure 5.9 System-wide noise measurements in *E. coli* and Yeast. The two system-wide noise studies are plotted for comparison. Eukaryotic protein abundance levels are much higher than bacteria resulting in a lower Poisson noise range. This figure is reproduced from the supplementary information of Taniguchi *et al* [32], and with data from Taniguchi *et al.* [32] and Newman *et al.* [33].

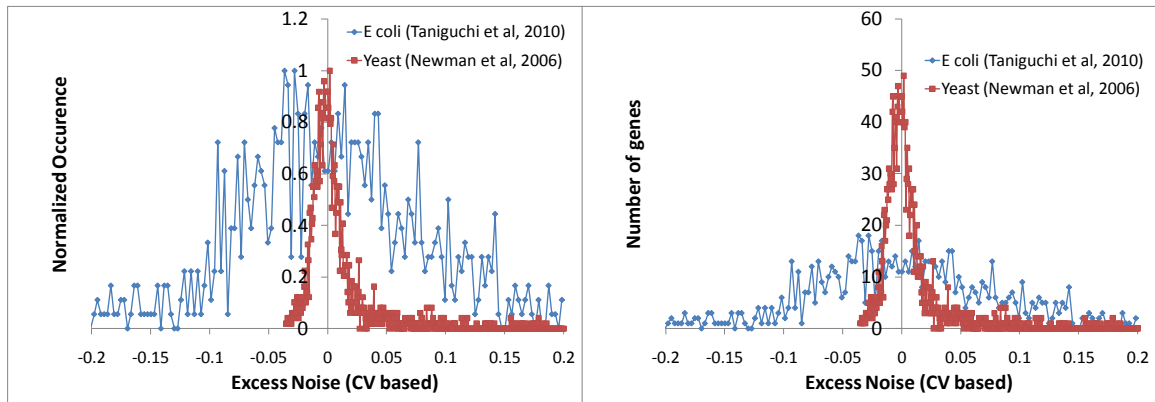


Figure 5.10 Comparison of excess noise in *E. coli* and Yeast. *E. coli* excess noise range is much wider than Yeast.

5.3.3 System-wide plasticity

To estimate the plasticity of every gene in the *E. coli* genome, microarray data from ~9 different studies that considered ~10 different stressors and a total of over 200 environments [132-140] were aligned and hierarchically clustered [141] on a gene-by-gene basis (Fig. 5.11). The data was retrieved from the *E. coli* Community's Gene Expression Database (GenExpDB), hosted by the University of Oklahoma (<http://genexpdb.ou.edu/>) [142]. Microarray data were selected for similar stressors that were used in the yeast study in the previous section. These include amino acid starvation [135], stationary phase growth [132], H₂O₂ exposure [132], DNA damage [134], glucose limitation [136], and more.

Similar to the yeast microarray data, the *E. coli* microarray data represented values of $\text{Log}_2(m_{ij}/m_{i0})$, where m_{ij} is the mRNA level of gene i in environment j , and m_{i0} is the reference mRNA level of gene i in a healthy environment. The heat map in figure 5.11 represents the ~4400 *E. coli* genes in each column and the 200+ environments on the horizontal rows. There are 2 immediate observations which can be made by this stress response heat map which distinguishes the *E. coli* response from the budding yeast: (1) *E. coli* does not have a pre-programmed Environmental Stress Response (ESR) -- a set of genes that up/down regulate regardless of the stressor type, as was present in the budding yeast (Fig. 5.1, 1/6th of the yeast genome is involved, ~600 growth genes are down-regulated and ~300 stress genes are up-regulated), and (2) Under stress, most of the genes are induced/repressed (shown as green or red in Fig. 5.11) to some extent, while fairly few remain at the same level of expression (shown as black in Fig. 5.11).

As a measure of a gene's plastic or deterministic response the maximal response of a gene under the various environments was used in the same plasticity definition (Equation 5.12) as before. Similar to before, to reduce error and variability in the plasticity calculation, an average of the highest and lowest 5 microarray values were used as an indication of maximal and minimal mRNA response respectively (Equation 5.13).

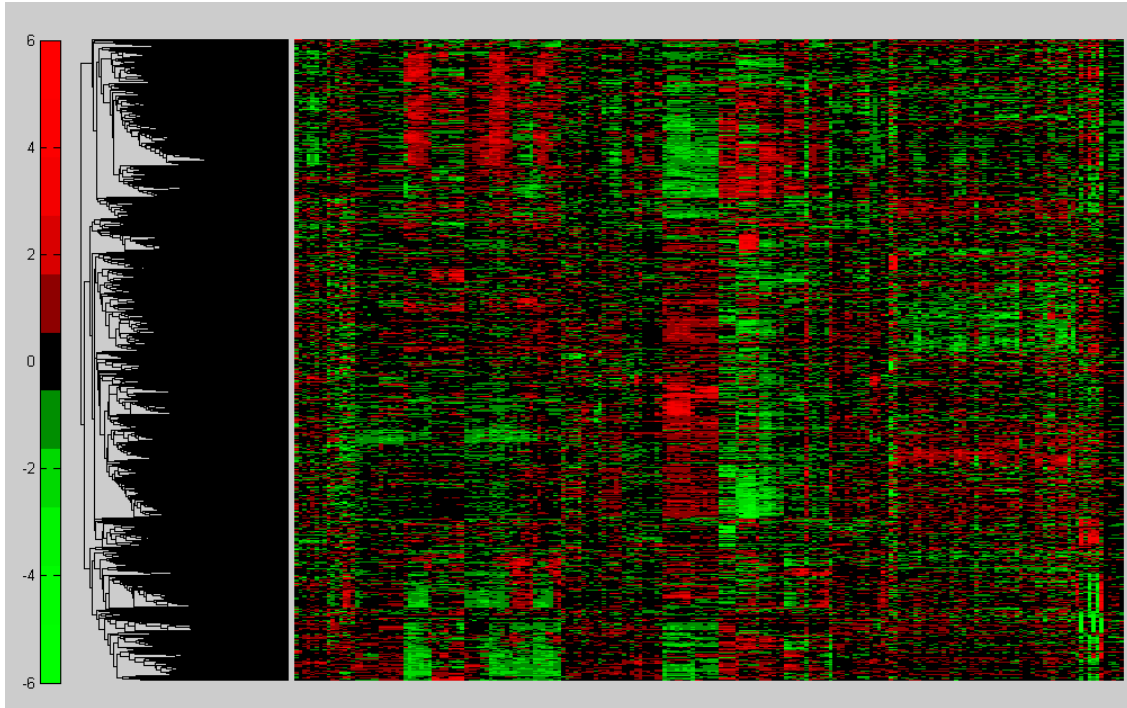


Figure 5.11 A constructed *E.coli* stress microarray compendium. $\text{Log}_2(m_{ij}/m_{i0})$ microarray values from a collection of studies [132-140] were aligned and hierarchically clustered for each gene [141]. The heat map has the ~ 4400 *E.coli* genes on the vertical axis, and ~ 200 environments on the horizontal. Red indicates the gene was induced and green that the gene was repressed. The scale-bar at left is of the fold induction/repression.

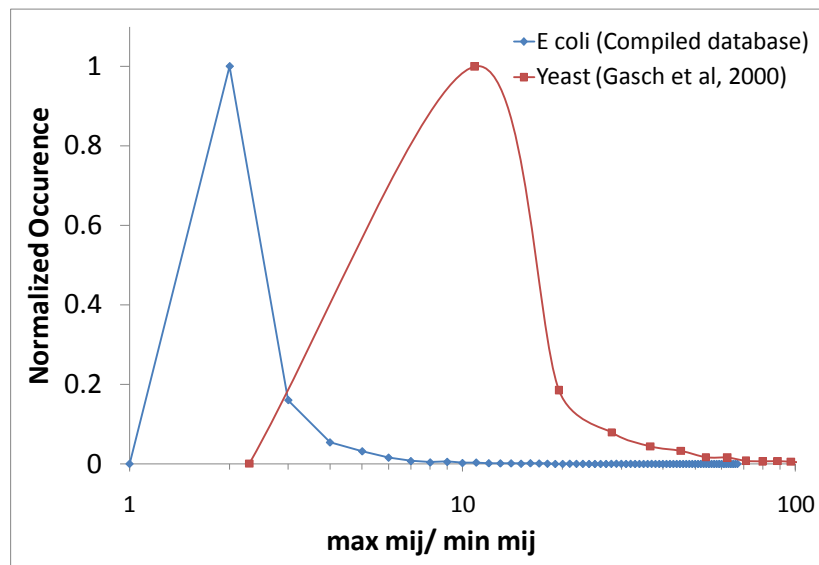


Figure 5.12 Comparison of plasticity range between *E.coli* and yeast. In yeast, the larger cell size, amount of resources, and bursty expression increases the plasticity range to an order of magnitude higher than *E.coli*.

5.3.4 Regulatory arrangements that control noise and plasticity

Sigma factors

In a similar fashion to the yeast investigation, it is of interest to find a generalized set of regulatory arrangements which regulate and control different strengths of noise-plasticity coupling. Since the *E.coli* chromosome is not compacted into chromatin, there is no *E. coli* parallel to the DNA structures found in the budding yeast (TATA/TATAless and Nucleosome Occupancy Pattern). Instead there are 7 *E.coli* sigma-factors, which control expression of specific genes in response to environmental stressors through protein-DNA interactions (somewhat analogous to TFIID versus SAGA co-activation complexes in yeast). The 7 *E.coli* sigma factors are: Nitrogen limitation (σ -21), flagella (σ -28), heat shock (σ -37), starvation or stationary phase (σ -38), extreme heat shock (σ -24), housekeeping (σ -70), and stress response (σ -S). Over 2200 Sigma Factor-Gene interactions found from data supplied by Freyre-Gonzalez *et al.* Genome Biology (2008) [143], were considered here.

5.3.5 Results

Figure 5.13 (upper) shows the single cell scatter of excess noise versus plasticity for *E.coli* and 5.13 (lower) shows the same relationship after binning and averaging 150 gene cohorts. This relation suggests that noise-plasticity coupling is widespread across the *E.coli* genome. There seem to be an abundance of low plasticity high excess noise genes that deviate on the upper side of the predicted noise-plastic coupling trend. This may suggest that genome-wide noise-plasticity coupling in *E. coli* is less widespread. Plots showing the differences in coupling for genes that are up-regulated versus repressed to stress are presented in the Appendix.

Next, genes were grouped together according to sigma factor control. Figure 5.14 shows that the sigma-controlled gene clustering also follows a noise-plasticity coupling (model line in black, similar to the model line in Fig. 5.13 lower) with stress-sigma (Sigma-S) having the highest noise-plasticity coupling strength. Flagella controlled genes by sigma-28 were not plotted as they only had 5 genes with noise measurements covered

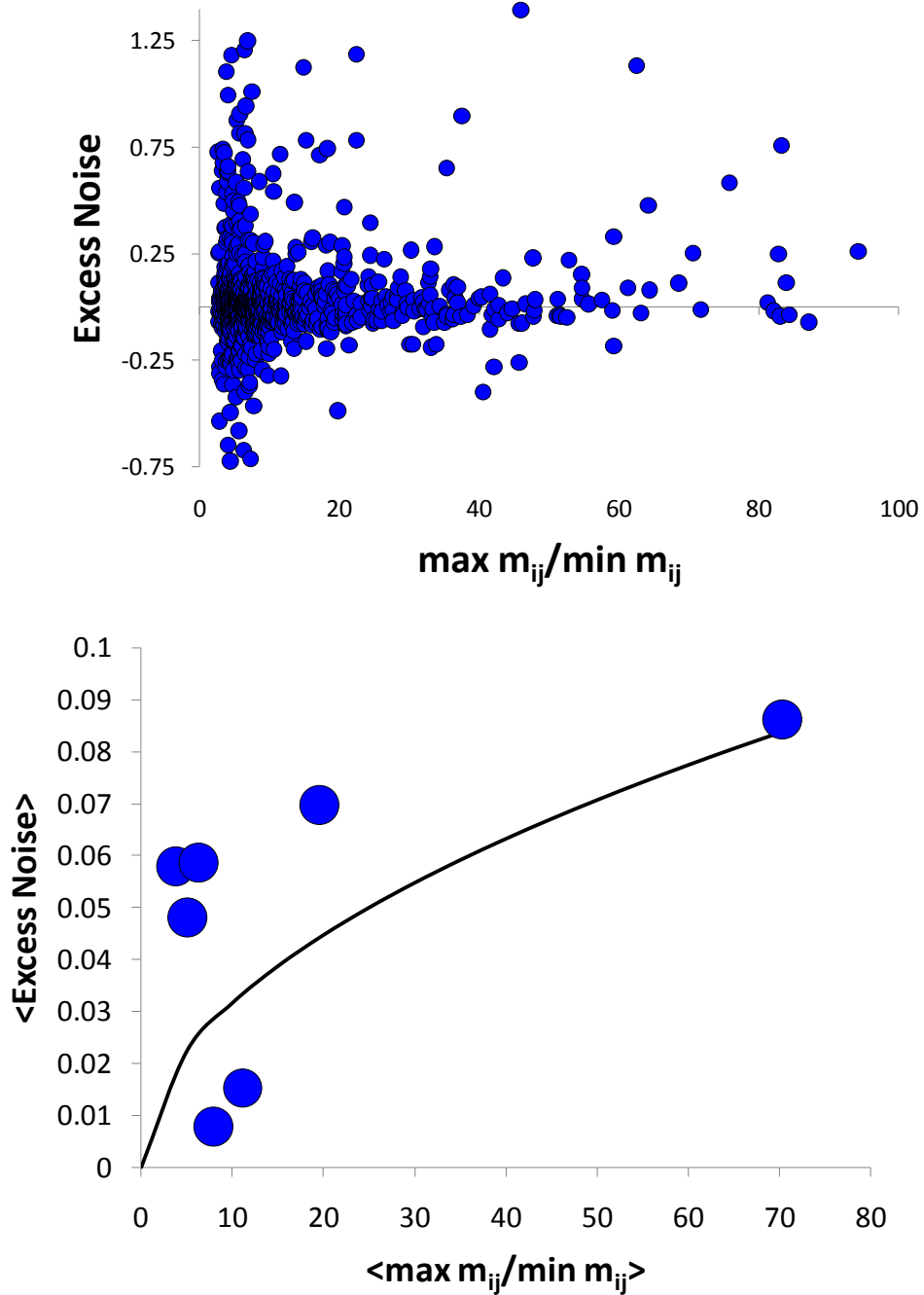


Figure 5.13 Widespread genome-wide noise-plasticity coupling in *E. coli*. (upper) 1017 *E. coli* genes. Genome-wide scatter does not reveal an obvious coupling. (lower) 150 gene bin averaging yields the predicted noise-plasticity coupling relationship. The black curve is the model line with equation $DM = 0.01 \cdot \sqrt{Pl}$.

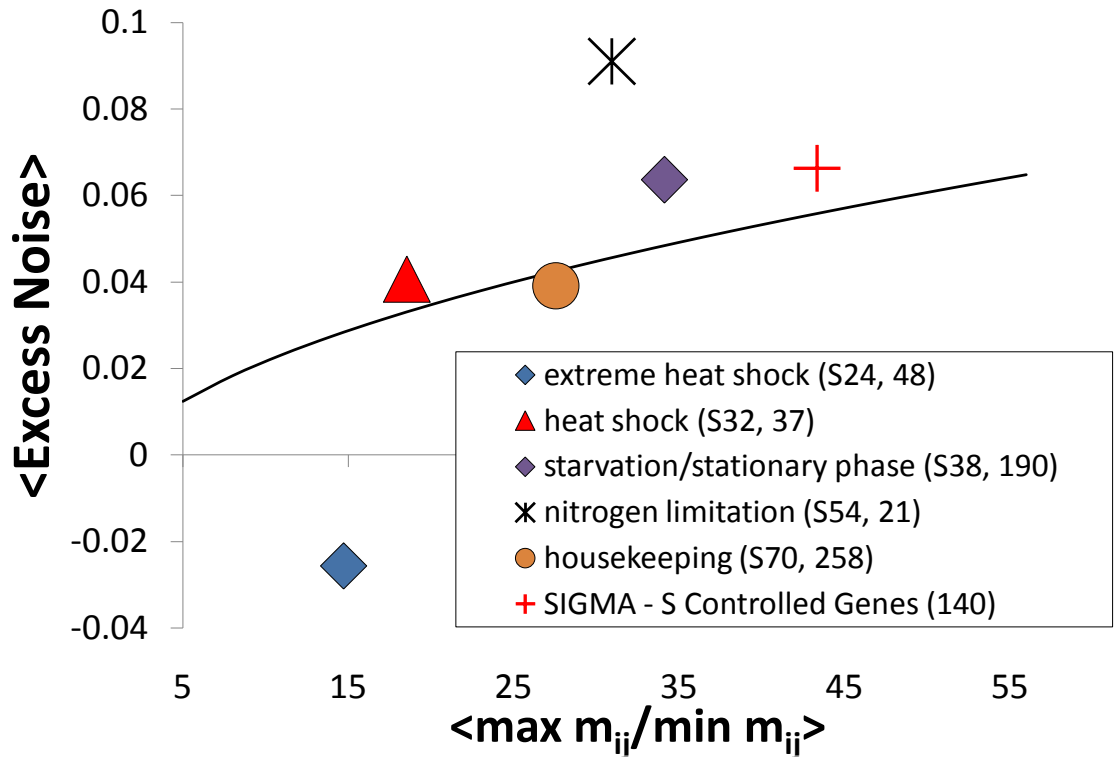


Figure 5.14 Noise-plasticity coupling among sigma-factor regulators. The legend specifies the sigma factor and the # of genes it controls that are accounted for in the gene clustering. Stress gene Sigma-S controlled genes are among the higher noise-plasticity coupled sigma factors. Extreme heat shock and heat shock genes are lower on the trend. Model line in black represents $= 0.01 \cdot \sqrt{Pl} - 0.01$.

in the Taniguchi *et al* study. Based on the 5 genes the plasticity and excess noise was very high (PI=38 and DM=0.24) which may be of interest as a particularly high noise gene cluster and may serve a functional role in flagellar synthesis, control, and overall cellular mobility in transitions between healthy and stressful environments. The limited ~1k DM measurements available in the database reduced the statistics in the sigma factor clustering, but enough noise measurements were available for the predicted relationship to emerge.

For additional gene clustering and noise-plasticity coupling maps, see the Appendix. As already mentioned, since high plasticity may be reached through different responses to stress, a response dependent clustering of genes was also performed in the Appendix.

5.3.6 Discussion

The finding of noise-plastic coupling in *E.coli* that follows a relationship similar to that found for yeast is intriguing as these two organisms have very different physiology and modes of regulation. The two-state transcriptional regulation model can describe both chromatin remodeling between open and closed states, and/or activation via protein-DNA interactions in the promoter operator region. The former (or a mixture of the two) may dominate and describe the yeast data whereas the latter is the main regulatory motif in *E.coli*. The results imply that although two very different types of transcriptional activation are occurring, their responses have been optimized to produce a very similar plasticity-stochasticity coupling. This may be an example of convergent evolution, where a similar need to respond to a fluctuating environment led to the selection of the same behavior in organisms that followed very different evolutionary paths.

5.4 A Novel Unicellular Noise-Plasticity Scaling Law

The main findings of this chapter may be summarized in Figure 5.15 which the stochasticity-plasticity coupling in both *E. coli* and *S. cerevesiae* on the same normalized scale. The figure suggests the possibility of a scaling law that describes a stochasticity and plasticity coupling optimized for the fluctuating environments seen by single-cell organisms. In an evolutionary sense, *E. coli* and *S. cerevesiae* diverged long ago, yet have arrived at a very similar relationship between noise and plasticity. It is still unknown if the coupling law applies to mammalian cells and multi-cellular organisms where evolution may have dictated different anticipatory demands and design as the range of environmental conditions to which such cells are exposed is controlled and limited (compared to a unicellular organism in an ‘unprotected’ environment). Certainly the number of genes participating in the stress response of human cells is much less than yeasts [144].

5.5 Noise-plasticity coupling is wide-spread but not all genes are coupled

Two-state model driven noise-plasticity coupling is wide-spread in both yeast and *E. coli* (Figures 5.4 and 5.13) but in no way applies to all genes in the system (genes were binned in groupings of 150 to detect the underlying behavior). In Simpson *et al*, 2009 [44] three consequences of noise are proposed: (1) noise is detrimental to a process and is minimized; (2) noise does not matter to the process and it is unimportant how much noise is distributed to it; and (3) noise has a functional use, is advantageous, and is exploited. Assuming all genes fall into these categories it is not surprising to observe a variety of noise-plasticity behaviors. Noise-plasticity couplers fall into case 3, and appear to be stochastic exploiters for some anticipatory advantage to randomly timed and unknown environments (more on this in the next section and Chapter 6).

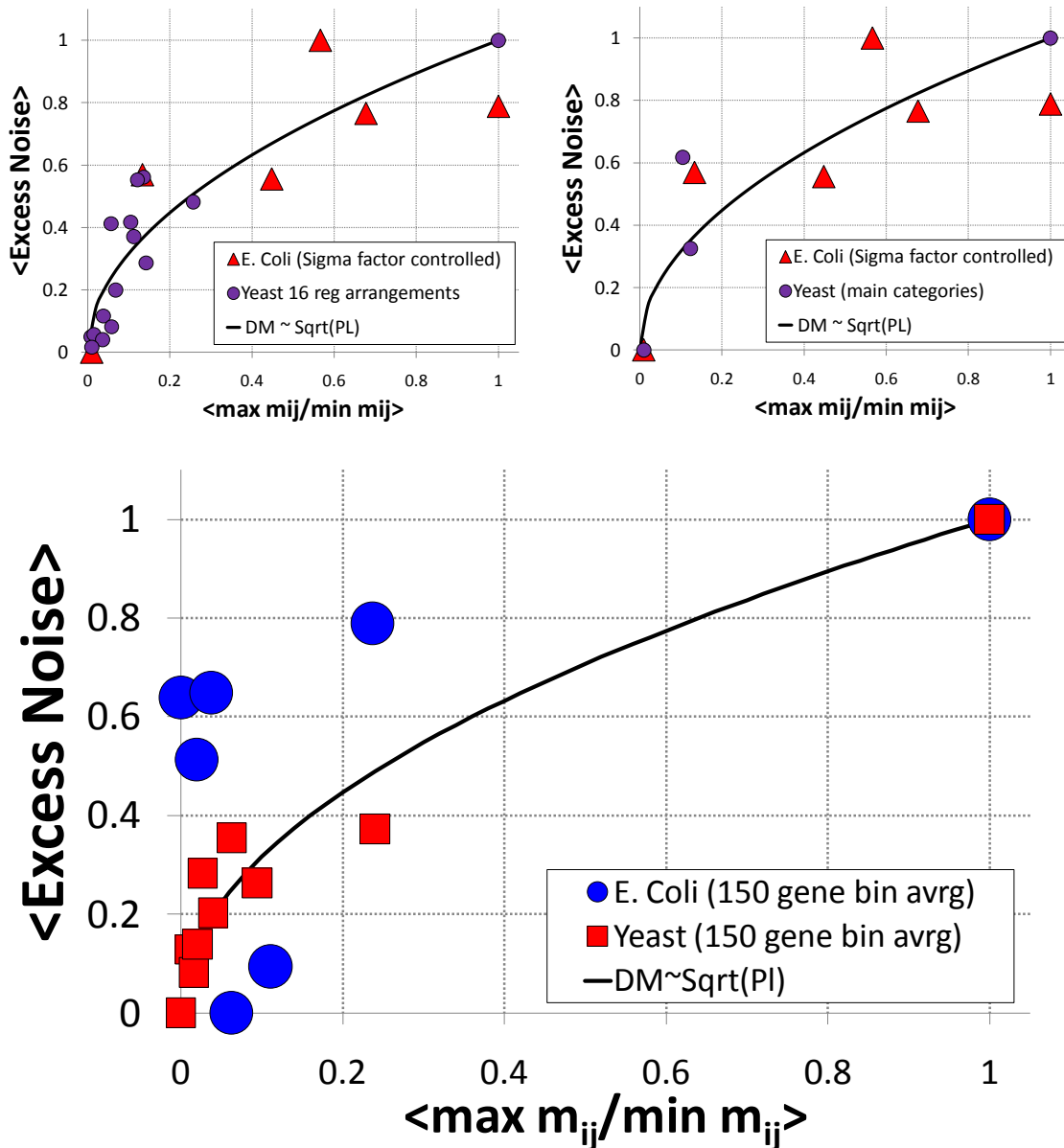


Figure 5.15 A Novel Unicellular Noise-Plasticity Scaling Law. (upper) Red triangles represent E.coli Sigma-factor controlled genes and purple circles are the main and sub-categories from the previous sections and Dar *et al*, 2010 [145]. The plots are scaled to the [0,1] range for comparison. The black line represents a Sqrt(PI) model line. (lower) Widespread genome-wide coupling. In this case no regulatory arrangement clustering is accounted for and genes in both organisms are clustered into bins of 150 genes in order of increasing plasticity. The predicted model line is shown in black.

Interestingly the two-state transcriptional regulation model can account for both the strong and weak coupling between noise and plasticity [10, 145] (Fig. 5.16). Genes that exhibit weak noise-plasticity coupling can either have (1) high excess noise with low plasticity, or (2) high plasticity with low excess noise. The former is observed at low plasticity values in both yeast and *E. coli* (Figures 5.4 and 5.13), and can be explained by a high noise architecture (e.g. low on fraction) where neither on fraction (O) nor transcription rate (α) responds to stress. The second case of high plasticity with low excess noise can be seen for certain highly responsive stress genes labeled in figure 5.16, lower. At least two mechanisms for such weak coupling are proposed in figure 5.16 (upper). The system may have frequent (high burst frequency, K) but short duration burst that may extend in response to stress. Alternatively, plasticity could be provided by a stress-responsive transcription rate (α). Overall, these observations of weakened noise-plasticity coupling supports a picture that high plasticity levels may be achieved without high excess noise and therefore noise is not necessarily just a byproduct of high plasticity [146].

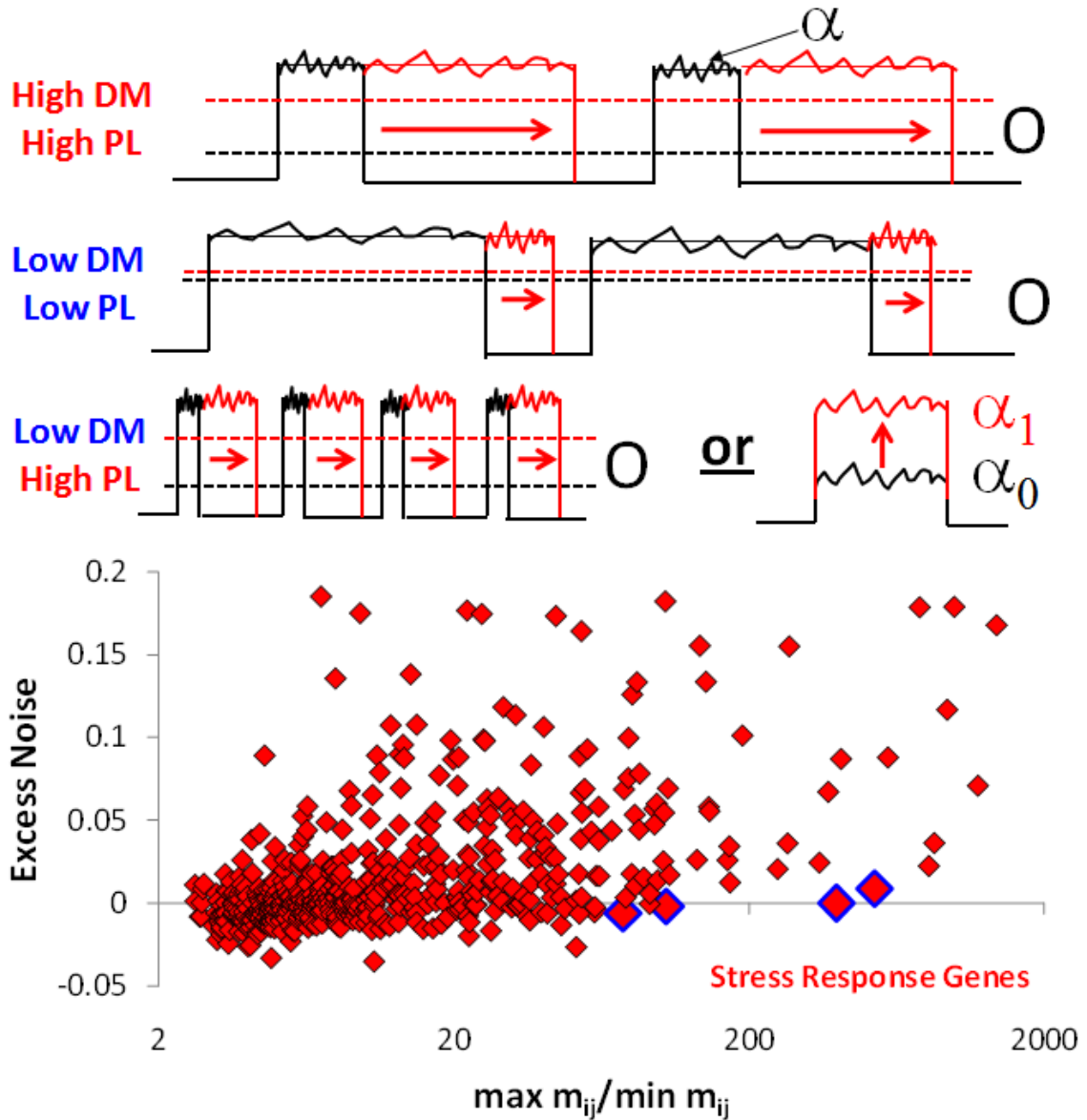


Figure 5.16 The two-state model can couple and uncouple noise and plasticity. (top) Depiction of the two-state model noise-plasticity driven by on fraction dependent modulation of plasticity with high noise (low O) or low noise (high O) in healthy environments. Noise-plasticity uncoupling may occur with short and frequent bursts (high burst frequency, K), which has low noise but can be on fraction modulated for high response and plasticity. The second uncoupling mechanism is a quite architecture that simply increases its transcription rate through activation. (bottom) examples of noise-plasticity uncoupling in the yeast stress response genes are seen along the plasticity axis. 4 high PL low DM genes are highlighted for clarity.

5.6 Noise-plasticity coupling of yeast and *E. coli* regulators

The findings of this chapter support that noise-plasticity coupling may form the basis of a type of stochastic exploitation with anticipatory implications. A unique subsystem of yeast genes that exploit noise-plasticity coupling not presented earlier are yeast regulators and the associated regulatory network. Yeast regulatory genes with larger numbers of downstream gene targets (K_{out}) appear to have increasing noise-plasticity coupling strengths (Fig. 5.17, upper). Transcriptional connectivity in yeast utilized data supplied by Milo *et al.*, (2002) [7] and accounted for regulators that were not themselves directly regulated by other proteins ($K_{in} = \sim 0$). Regulators with a K_{out} of 4 deviate from the predicted model line (black) and upon increasing the plasticity measurement statistics from 9 to 24 genes the outlier better fits the predicted model line (Fig. 5.18, upper) (the Newman *et al.*, 2006 [102] excess noise measurements are limited to 2k genes measured above background cell autofluorescence while plasticity measurements exist for the whole genome). This assumes that the excess noise calculated from the limited data is representative and in the right vicinity as the 24 gene statistic value (if all the noise measurements were available). The increase in coupling strength with increasing K_{out} is gene regulatory arrangement driven as the % of genes with TATA-SAGA architectures and nucleosome occupancy pattern #1 increase with increasing K_{out} while the % of genes with patterns #3 and #4 decrease (Fig. 5.17, lower). These are important gene regulators in the system. They are an even mixture of stress-induced and stress-repressed genes (data not shown), are related to the regulation of stress response, metabolism, and energy pathways, consume resources by their high abundances (Fig. 5.18, lower), and are stable with long protein half-lives in healthy conditions (Fig. 5.18, lower). It appears that evolution has consistently placed the strongest coupling strength regulatory arrangement in the highest K_{out} regulators suggesting a stochastic versus deterministic regulatory strategy for a randomly timed perturbation in the most highly connected regulators.

The above detection of increased coupling strength of yeast regulators with increasing K_{out} is not intuitive. This seems to negate the ~40% of *E. coli* transcription

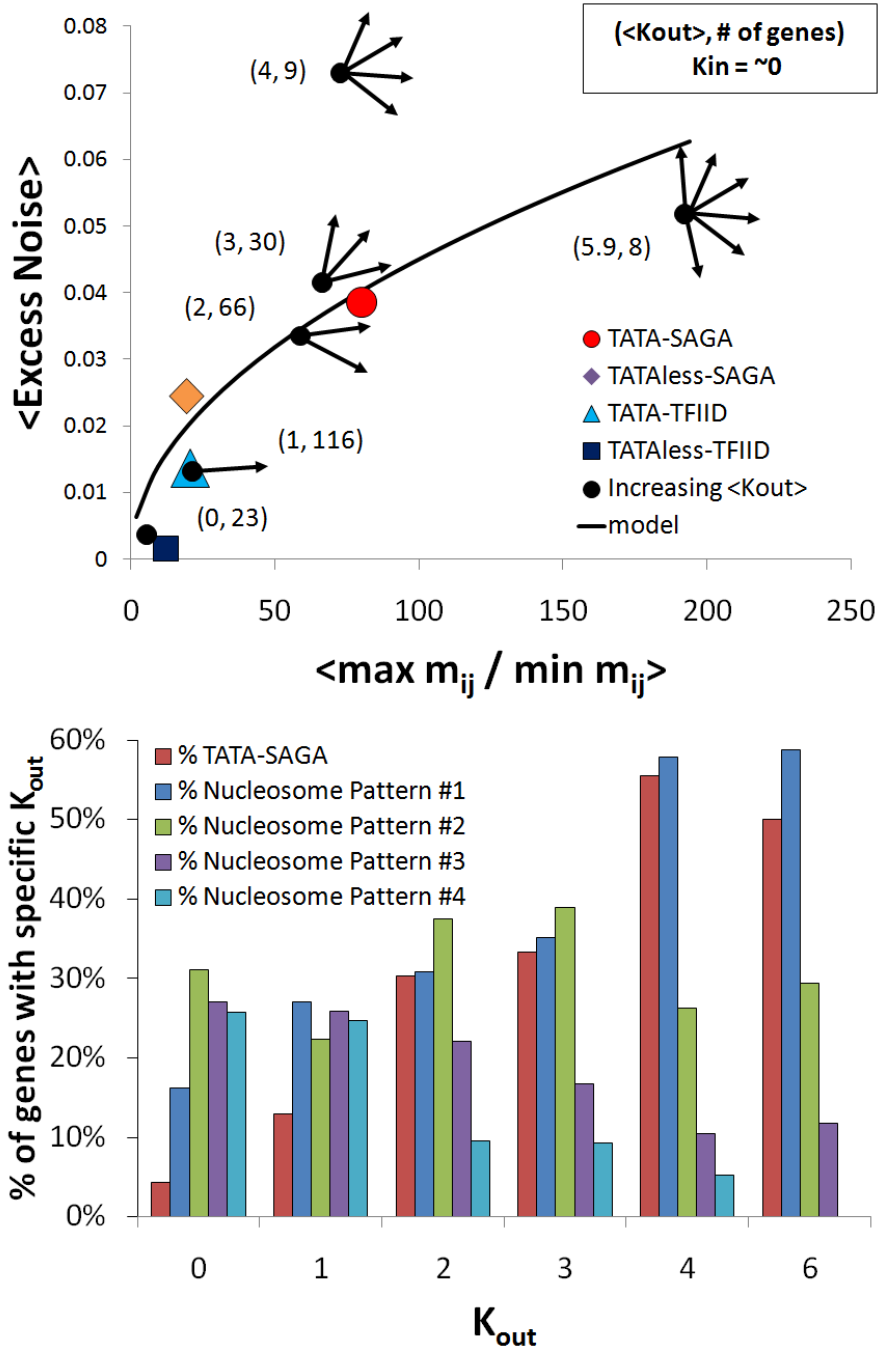


Figure 5.17 Noise-plasticity coupling of yeast regulators. (upper) Following the derived two-state noise-plasticity model line (black line), yeast regulators with the greatest number of downstream regulated gene targets (K_{out}) have the largest noise-plasticity coupling strength and TATA-SAGA #1 regulatory arrangements (upper and lower). Transcriptional connectivity in yeast utilized data supplied by Milo *et al.*, (2002) [7] and the other yeast data as described earlier in this chapter.

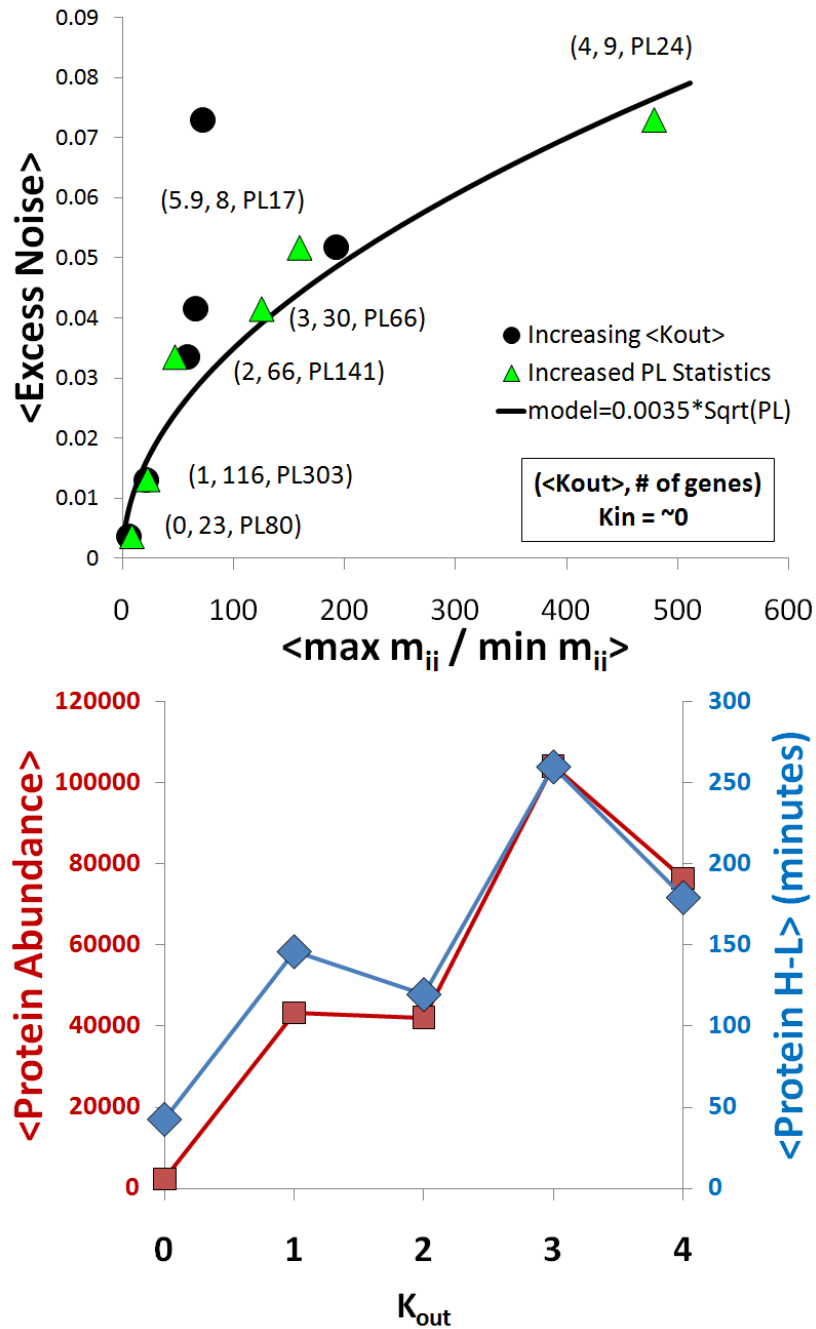


Figure 5.18 Yeast regulators are important stochastic exploiters. (upper) accounting for additional plasticity statistics (labeled PL# of genes in the figure) the yeast regulators with increasing K_{out} yield a better fit to the predicted model compared to figure 5.17. (lower) yeast regulators with increasing K_{out} have higher protein abundances and half-lives in healthy environments suggesting importance to the system along with strong noise-plasticity coupling.

factors (TFs) which are negatively autoregulated (-AR) and thought to stabilize system-wide regulation with noise magnitude suppression and frequency shifts to higher ranges for filtering (e.g. in a cascade) to produce high fidelity signals. To explore this enigma the noise-plasticity coupling strengths of *E. coli* TFs that are both -AR and non- -AR were plotted on the earlier *E. coli* genome-wide noise-plasticity coupling plot (Fig. 5.19). Updated TFs and identification of -AR TFs utilized RegulonDB version 7.0 [147] (found at <http://regulondb.ccg.unam.mx/>). A total of 14 -AR TFs and 29 non- -AR TFs with both DM and PL measurements were used. The results suggest that the -AR TFs have suppressed negative valued excess noise, as predicted from -AR noise analysis of sections 3.3.1 and 4.1.3, while the non -AR TFs have a weak noise-plasticity coupling strength. Additional TF connectivity analysis was hampered by limited excess noise measurements in the current dataset [32]. The results seem to suggest that *E. coli* and yeast regulators have different noise-plasticity design strategies but additional investigation is needed for a complete comparison.

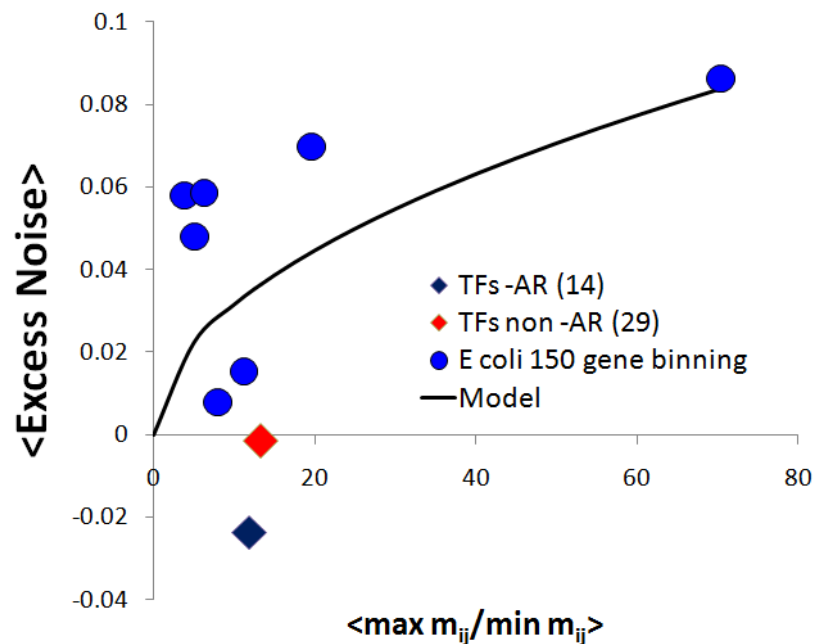


Figure 5.19 Non -AR *E. coli* TFs are noise-plasticity coupled. -AR *E. coli* TFs have negative excess noise as predicted by -AR noise modulation (Section 3.3.1). Other TFs that are non -AR appear to have weak noise-plasticity coupling. # of genes accounted for are labeled in the legend.

CHAPTER 6: Summary and Conclusions

As stated in the introduction, the vision of the nanoscience revolution is to create new systems with functionality that greatly exceeds that possible with microscale technology, but this cannot be accomplished using microscale strategies that do not scale downward. In particular, the approach of overpowering stochasticity is not a feasible nanoscale strategy, and instead we must embrace Nature's quite different paradigms of *exploiting* stochasticity, rather than overwhelming it.

Of course, the main difficulty with this approach is that we do not understand the rules of composition when noise is a major element of function. Accordingly, the central goal of this thesis was to begin to explore the lessons that biological cells can teach us about the emergences of function from deep within the noise.

To guide this exploration, this thesis tackled five major questions:

1. What is gene circuit noise?
2. What is its structure and how is it measured?
3. How is it regulated?
4. How can it be used to create function?
5. How can noise structure, distribution, regulation, and function be studied across all the components of a complex nanoscale system?

These questions were tackled using a variety of analytical, computational, and experimental approaches on three cell types (prokaryote, single cell eukaryote, and human cells). Although it created tremendous experimental challenges, this variety of cell types was essential for addressing the questions asked here. These cells have followed very different evolutionary pathways, and clearly have arrived at very different molecular mechanisms to achieve their various functions. Therefore, when these cells are found to have settled on very similar strategies, it becomes appropriate to ask: are these strategies part of the fundamental rules of composition of creating function from deep within the noise?

What have the explorations taught us about the five major questions asked above?

In the low (shot) noise limit, gene circuit noise is the natural consequence of the discrete nature and random timing of molecular events (synthesis, decay, binding, etc.). This noise would persist even if every gene were given all the resources required for its full expression. However, such a resource rich environment is not a realistic view of complex nanoscale systems, and is especially inapplicable to the harsh peer review (i.e. evolutionary selection) faced by cells. As Figure 6.1 illustrates, at some point – even for synthetic systems – the resource rich approach runs afoul of fundamental physical constraints. Getting more function in less space while using less energy creates a fundamental dilemma: the conservation of stochasticity (equation 1.4). Noise can be moved around between the components, but it cannot be avoided. So in summary, gene circuit noise is a consequence of confinement and the manifestation of evolutionary decisions about the distribution of resources. However, this noise is also an opportunity, that is, a functional component that can be used to create more function in less space.

Noise has a rich structure with components that describe its magnitude, correlation, and distribution, but measuring the full extent of this structure can be quite challenging. The highest throughput method – flow cytometry – only allows the measurement of noise magnitude and distribution. Time-lapse fluorescent microscopy can also measure noise correlation, but at the cost of following a limited number of cells over long time periods. These long imaging experiments are followed by a painstaking process of signal processing and analysis to (among other things) remove deterministic transient signals and deal with extrinsic noise and differences in basal expression levels. However, with these challenges addressed, noise structure can be measured in enough detail to elucidate the structure (and perhaps function) of the underlying gene circuits. So, the measurement of the noise in the expression of HIV-1 genes can be used to deduce the inner workings of this circuit, but the key to using noise structure in this manner is understanding the coupling between gene circuit and noise structure. Noise structure is regulated by the same mechanisms that regulate the other gene expression attributes (e.g. mean level, timing of expression). Indeed, every regulatory arrangement leaves its (not necessarily unique) signature in the noise, and in combination with other more traditional

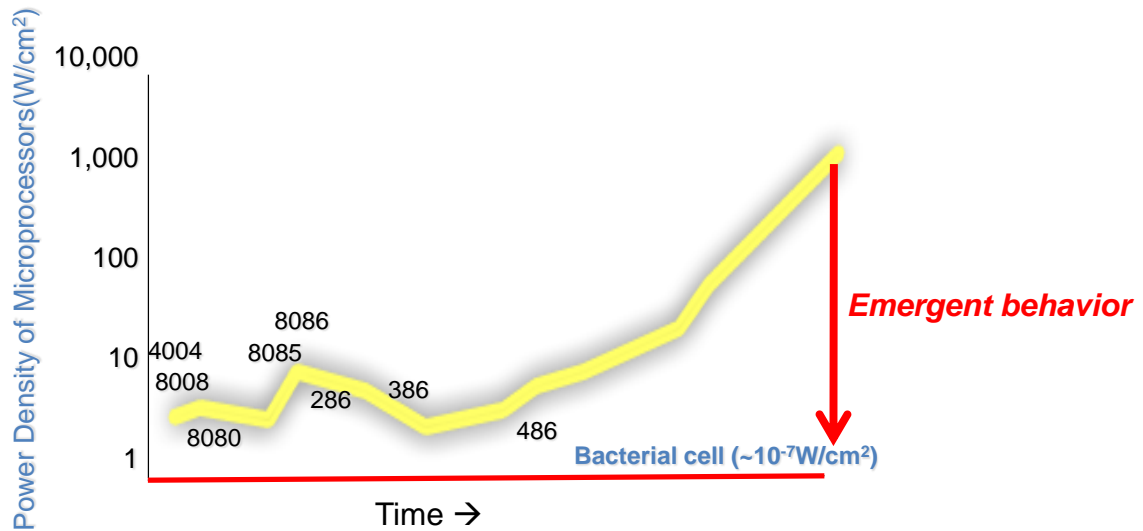


Figure 6.1 Problems with resource rich driven synthetic design. The graph depicts a plot of the microprocessor power density versus time for different microprocessor models developed over time and presented by Pat Gelsinger at an Intel Developer Forum, Spring 2004 (Pentium at 90 W). The power trend yields a microprocessor socket temperature that approaches that of the sun’s surface. The bacterial cell functions at a power consumption of eight to ten orders of magnitude lower than the microprocessor trend.

experimental methods, provides a new tool for unraveling a gene circuit’s secrets.

From a nanoscale science point of view, noise is an ideal component. It takes up no space and uses no energy, yet it can be used to produce greater functionality. The scientific community is only just beginning to understand how noise is used to create function in cells, and unfortunately, at present this is mostly done by example, rather than by the elucidation of underlying principles. This thesis adds to this growing list of examples by shedding more light on the noise driven strategy of the HIV-1 circuit and by describing the coupling between plasticity and stochasticity. The HIV-1 circuit strategy is the use of noise to provide a distribution of times for the virus to reproduce. Most infections will lead to rapid reproduction of the virus and death of the infected cell in short order. Yet a few contrarian infections will wait not truly latent, but delayed. These contrarian events are the main factor thwarting HIV-1 eradication from an affected individual – or from the viruses point of view, these contrarian events are the main factor

that allow survival in an unpredictable fluctuating environment. Similar issues might be at play in the coupling between stochasticity and plasticity. Many of the most plastic genes are those that respond to adverse environments, and their noisy behavior in non-stressful environments could be a noise driven strategy to respond to sudden unpredictable environmental fluctuations.

Studying the noise in individual elements has been the primary research activity in Noise Biology. However, the real challenge lies in studying the noise structure, distribution, regulation, and function across an entire system (e.g. organism). Previous work has provided organism-wide noise magnitude measurements using flow cytometry that has proved to be extremely useful. However, as demonstrated throughout this thesis, noise structure is much richer than magnitude alone, and within this richer structure lies greater detail about how noise is integrated into complex systems to produce function. The noise mapping technique described here is perhaps the first example of an experimental tool that gets at both the rich structure of the noise and the system-wide behavior of this noise.

With at least some answers to each of the five main questions, what picture emerges about the system-wide distribution of noise in a complex nanoscale system? One obvious observation is the central role of the 2-state transcriptional bursting motif. The importance of this motif would seem to be that it couples together competing interests of the cell. It couples the distribution of cellular resources for gene expression to the distribution of noise to the different gene circuits. It also couples together the plasticity and noise in gene expression. As it lies at the intersection of so many important evolutionary decisions, perhaps it is not surprising to see it pop up so prominently in the pursuit of the questions in this thesis. At present, this expression motif has been studied using simple models and limited experimentation that belie its seemingly more central role. This is an obvious area for future work, and noise mapping provides a powerful new tool to probe more deeply into this issue.

Finally, figure 5.15 shows a very intriguing finding. *E. coli* and *S. cerevesiae*, two organisms that diverged 2B years ago and have since followed two very different

evolutionary paths, have settled on a strikingly similar coupling between noise and plasticity. Is this just coincidence, or does it speak to such a coupling being an optimized response to the similar fluctuating environments experience by these organisms? While the work presented here can uncover this relationship and demonstrate how it may be related to 2-state transcriptional bursting, it is a question for future work to explore this relationship further. All research work should generate new questions, but must itself end before all these questions can be answered. And so it is that this thesis ends here.

LIST OF REFERENCES

1. Csete, M.E. and J.C. Doyle, *Reverse engineering of biological complexity*. Science, 2002. **295**(5560): p. 1664-1669.
2. Doktycz, M.J. and M.L. Simpson, *Nano-enabled synthetic biology*. Molecular Systems Biology, 2007. **3**.
3. Simpson, M.L., G.S. Saylor, J.T. Fleming and B. Applegate, *Whole-cell biocomputing*. Trends in Biotechnology, 2001. **19**(8): p. 317-323.
4. Barabasi, A.L. and R. Albert, *Emergence of scaling in random networks*. Science, 1999. **286**(5439): p. 509-512.
5. Jeong, H., B. Tombor, R. Albert, Z.N. Oltvai, and A.L. Barabasi, *The large-scale organization of metabolic networks*. Nature, 2000. **407**(6804): p. 651-654.
6. Jeong, H., S.P. Mason, A.L. Barabasi and Z.N. Oltvai, *Lethality and centrality in protein networks*. Nature, 2001. **411**(6833): p. 41-42.
7. Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, et al., *Network motifs: Simple building blocks of complex networks*. Science, 2002. **298**(5594): p. 824-827.
8. Shen-Orr, S.S., R. Milo, S. Mangan and U. Alon, *Network motifs in the transcriptional regulation network of Escherichia coli*. Nature Genetics, 2002. **31**(1): p. 64-68.
9. Simpson, M.L., C.D. Cox and G.S. Saylor, *Frequency domain analysis of noise in autoregulated gene circuits*. Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(8): p. 4551-4556.
10. Simpson, M.L., C.D. Cox and G.S. Saylor, *Frequency domain chemical Langevin analysis of stochasticity in gene transcriptional regulation*. Journal of Theoretical Biology, 2004. **229**(3): p. 383-394.
11. Singh, A. and L.S. Weinberger, *Stochastic gene expression as a molecular switch for viral latency*. Current Opinion in Microbiology, 2009. **12**(4): p. 460-466.
12. Weinberger, L.S., J.C. Burnett, J.E. Toettcher, A.P. Arkin, and D.V. Schaffer, *Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity*. Cell, 2005. **122**(2): p. 169-182.
13. Weinberger, L.S. and T. Shenk, *An HIV feedback resistor: Auto-regulatory circuit deactivator and noise buffer*. Plos Biology, 2007. **5**(1): p. 67-81.
14. Singh, A., B. Razooky, C.D. Cox, M.L. Simpson, and L.S. Weinberger, *Transcriptional Bursting from the HIV-1 Promoter Is a Significant Source of Stochastic Noise in HIV-1 Gene Expression*. Biophysical Journal, 2010. **98**(8): p. L32-L34.
15. Weinberger, L.S., R.D. Dar and M.L. Simpson, *Transient-mediated fate determination in a transcriptional circuit of HIV*. Nature Genetics, 2008. **40**(4): p. 466-470.
16. Alberts, B., *Molecular Biology of the Cell*. 2002.
17. Patashne, M., *GENETIC SWITCH PHAGE LAMBDA*. 1992.
18. Patashne, M., *GENES AND SIGNALS*. 2001.
19. Alon, U., *An introduction to systems biology : design principles of biological circuits*. 2007, Boca Raton, FL: Chapman & Hall/CRC. xvi, 301 p., [4] p. of plates.

20. Simpson, M.L., C.D. Cox, G.D. Peterson and G.S. Saylor, *Engineering in the biological substrate: Information processing in genetic circuits*. Proceedings of the Ieee, 2004. **92**(5): p. 848-863.
21. Delbruck, M., *Statistical fluctuations in autocatalytic reactions*. Journal of Chemical Physics, 1940. **8**(1): p. 120-124.
22. Delbruck, M., *The burst size distribution in the growth of bacterial viruses (bacteriophages)*. Journal of Bacteriology, 1945. **50**: p. 131-135.
23. Novick, A. and M. Weiner, *Enzyme induction as an all-or-none phenomenon*. Proceedings of the National Academy of Sciences of the United States of America, 1957. **43**: p. 553-566.
24. Ptashne, M., *A genetic switch: gene control and phage lambda*. 1986, Cambridge, MA: Cell Press. 128.
25. Arkin, A., J. Ross and H.H. McAdams, *Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells*. Genetics, 1998. **149**(4): p. 1633-1648.
26. Spudich, J.L. and D.E. Koshland, Jr., *Non-genetic individuality: chance in the single cell*. Nature, 1976. **262**: p. 467-471.
27. Elowitz, M.B., A.J. Levine, E.D. Siggia and P.S. Swain, *Stochastic gene expression in a single cell*. Science, 2002. **297**(5584): p. 1183-1186.
28. Swain, P.S., M.B. Elowitz and E.D. Siggia, *Intrinsic and extrinsic contributions to stochasticity in gene expression*. Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(20): p. 12795-12800.
29. Raser, J.M. and E.K. O'Shea, *Control of stochasticity in eukaryotic gene expression*. Science, 2004. **304**(5678): p. 1811-1814.
30. Colman-Lerner, A., A. Gordon, E. Serra, T. Chin, et al., *Regulated cell-to-cell variation in a cell-fate decision system*. Nature, 2005. **437**(7059): p. 699-706.
31. Simpson, P., *Notch signalling in development: on equivalence groups and asymmetric developmental potential*. Curr Opin Genet Dev, 1997. **7**(4): p. 537-42.
32. Taniguchi, Y., P.J. Choi, G.W. Li, H.Y. Chen, et al., *Quantifying E-coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells*. Science, 2010. **329**(5991): p. 533-538.
33. Newman, J.R.S., S. Ghaemmaghami, J. Ihmels, D.K. Breslow, et al., *Single-cell proteomic analysis of S-cerevisiae reveals the architecture of biological noise*. Nature, 2006. **441**(7095): p. 840-846.
34. Bar-Even, A., J. Paulsson, N. Maheshri, M. Carmi, et al., *Noise in protein expression scales with natural protein abundance*. Nature Genetics, 2006. **38**(6): p. 636-643.
35. Cohen, A.A., N. Geva-Zatorsky, E. Eden, M. Frenkel-Morgenstern, et al., *Dynamic proteomics of individual cancer cells in response to a drug*. Science, 2008. **322**(5907): p. 1511-6.
36. Kussell, E. and S. Leibler, *Phenotypic diversity, population growth, and information in fluctuating environments*. Science, 2005. **309**(5743): p. 2075-2078.
37. Thattai, M. and A. van Oudenaarden, *Stochastic gene expression in fluctuating environments*. Genetics, 2004. **167**(1): p. 523-530.

38. Acar, M., J.T. Mettetal and A. van Oudenaarden, *Stochastic switching as a survival strategy in fluctuating environments*. Nature Genetics, 2008. **40**(4): p. 471-475.
39. Kaern, M., T.C. Elston, W.J. Blake and J.J. Collins, *Stochasticity in gene expression: From theories to phenotypes*. Nature Reviews Genetics, 2005. **6**(6): p. 451-464.
40. Longo, D. and J. Hasty, *Imaging gene expression: tiny signals make a big noise*. Nature Chemical Biology, 2006. **2**(4): p. 181-182.
41. Kaufmann, B.B. and A. van Oudenaarden, *Stochastic gene expression: from single molecules to the proteome*. Current Opinion in Genetics & Development, 2007. **17**(2): p. 107-112.
42. Shahrezaei, V. and P.S. Swain, *The stochastic nature of biochemical networks*. Current Opinion in Biotechnology, 2008. **19**(4): p. 369-374.
43. Larson, D.R., R.H. Singer and D. Zenklusen, *A single molecule view of gene expression*. Trends in Cell Biology, 2009. **19**(11): p. 630-637.
44. Simpson, M.L., C.D. Cox, M.S. Allen, J.M. McCollum, et al., *Noise in biological circuits*. Wiley Interdisciplinary Reviews-Nanomedicine and Nanobiotechnology, 2009. **1**(2): p. 214-225.
45. Rosenfeld, N., J.W. Young, U. Alon, P.S. Swain, and M.B. Elowitz, *Gene Regulation at the Single-Cell Level*. Science, 2005. **307**(5717): p. 1962-1965.
46. Cox, C.D., J.M. McCollum, D.W. Austin, M.S. Allen, et al., *Frequency domain analysis of noise in simple gene circuits*. Chaos, 2006. **16**(2).
47. Cox, C.D., J.M. McCollum, M.S. Allen, R.D. Dar, and M.L. Simpson, *Using noise to probe and characterize gene circuits*. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(31): p. 10809-10814.
48. Mcquarri, D.A., *Stochastic Approach to Chemical Kinetics*. Journal of Applied Probability, 1967. **4**(3): p. 413-&.
49. Gillespie, D.T., *The chemical Langevin equation*. Journal of Chemical Physics, 2000. **113**(1): p. 297-306.
50. Simpson, M.L., Cox, C. D., Saylor, G. S., *Frequency Domain Analysis of Noise in Autoregulated Gene Circuits*. Proceedings of the National Academy of Sciences, 2003. **100**: p. 4551-4556.
51. Cox, C.D., G.D. Peterson, M.S. Allen, L.J. M., et al., *Analysis of Noise in Quorum Sensing*. Omics, 2003.
52. Gillespie, D.T., *Exact Stochastic Simulation of Coupled Chemical-Reactions*. Journal of Physical Chemistry, 1977. **81**(25): p. 2340-2361.
53. Prasher, D.C., V.K. Eckenrode, W.W. Ward, F.G. Prendergast, and M.J. Cormier, *Primary Structure of the Aequorea-Victoria Green-Fluorescent Protein*. Gene, 1992. **111**(2): p. 229-233.
54. Chalfie, M., Y. Tu, G. Euskirchen, W.W. Ward, and D.C. Prasher, *Green Fluorescent Protein as a Marker for Gene-Expression*. Science, 1994. **263**(5148): p. 802-805.

55. Ormo, M., A.B. Cubitt, K. Kallio, L.A. Gross, et al., *Crystal structure of the Aequorea victoria green fluorescent protein*. *Science*, 1996. **273**(5280): p. 1392-1395.
56. Yang, F., L.G. Moss and G.N. Phillips, *The molecular structure of green fluorescent protein*. *Nature Biotechnology*, 1996. **14**(10): p. 1246-1251.
57. Tsien, R.Y., *The green fluorescent protein*. *Annual Review of Biochemistry*, 1998. **67**: p. 509-544.
58. Heim, R., D.C. Prasher and R.Y. Tsien, *Wavelength Mutations and Posttranslational Autoxidation of Green Fluorescent Protein*. *Proceedings of the National Academy of Sciences of the United States of America*, 1994. **91**(26): p. 12501-12504.
59. Huh, W.K., J.V. Falvo, L.C. Gerke, A.S. Carroll, et al., *Global analysis of protein localization in budding yeast*. *Nature*, 2003. **425**(6959): p. 686-691.
60. Jaroszeski, M.J. and G. Radcliff, *Fundamentals of flow cytometry*. *Molecular Biotechnology*, 1999. **11**(1): p. 37-53.
61. Ibrahim, S.F. and G. van den Engh, *High-speed cell sorting: fundamentals and recent advances*. *Current Opinion in Biotechnology*, 2003. **14**(1): p. 5-12.
62. Elson, E.L. and D. Magde, *Fluorescence Correlation Spectroscopy .I. Conceptual Basis and Theory*. *Biopolymers*, 1974. **13**(1): p. 1-27.
63. Magde, D., E.L. Elson and W.W. Webb, *Fluorescence Correlation Spectroscopy .2. Experimental Realization*. *Biopolymers*, 1974. **13**(1): p. 29-61.
64. Magde, D., W.W. Webb and E. Elson, *Thermodynamic Fluctuations in a Reacting System - Measurement by Fluorescence Correlation Spectroscopy*. *Physical Review Letters*, 1972. **29**(11): p. 705-&.
65. Hess, S.T., S.H. Huang, A.A. Heikal and W.W. Webb, *Biological and chemical applications of fluorescence correlation spectroscopy: A review*. *Biochemistry*, 2002. **41**(3): p. 697-705.
66. Petersen, N.O., C. Brown, A. Kaminski, J. Rocheleau, et al., *Analysis of membrane protein cluster densities and sizes in situ by image correlation spectroscopy*. *Faraday Discussions*, 1998(111): p. 289-305.
67. Wiseman, P.W., F. Capani, J.A. Squier and M.E. Martone, *Counting dendritic spines in rat cerebellum neurons by image correlation spectroscopy*. *Biophysical Journal*, 2001. **80**(1): p. 504a-505a.
68. Wiseman, P.W. and N.O. Petersen, *Image correlation spectroscopy. II. Optimization for ultrasensitive detection of preexisting platelet-derived growth factor-beta receptor oligomers on intact cells*. *Biophysical Journal*, 1999. **76**(2): p. 963-977.
69. Sigal, A., R. Milo, A. Cohen, N. Geva-Zatorsky, et al., *Variability and memory of protein levels in human cells*. *Nature*, 2006. **444**(7119): p. 643-646.
70. Austin, D.W., M.S. Allen, J.M. McCollum, R.D. Dar, et al., *Gene network shaping of inherent noise spectra*. *Nature*, 2006. **439**(7076): p. 608-611.
71. Locke, J.C.W. and M.B. Elowitz, *Using movies to analyse gene circuit dynamics in single cells*. *Nature Reviews Microbiology*, 2009. **7**(5): p. 383-392.

72. Cox, C.D., G.D. Peterson, M.S. Allen, J.M. Lancaster, et al., *Analysis of noise in quorum sensing*. Omics a Journal of Integrative Biology, 2003. **7**(3): p. 317-34.
73. Gibson, M.A. and J. Bruck, *Efficient exact stochastic simulation of chemical systems with many species and many channels*. Journal of Physical Chemistry A, 2000. **104**(9): p. 1876-1889.
74. Kepler, T.B. and T.C. Elston, *Stochasticity in transcriptional regulation: Origins, consequences, and mathematical representations*. Biophysical Journal, 2001. **81**(6): p. 3116-3136.
75. Cox, C.D., G.D. Peterson, M.S. Allen, J.M. Lancaster, et al., *Analysis of noise in quorum sensing*. OMICS, 2003. **7**(3): p. 317-34.
76. Thieffry, D., A.M. Huerta, E. Perez-Rueda and J. Collado-Vides, *From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli*. Bioessays, 1998. **20**(5): p. 433-440.
77. Dunlap, J.C., *Molecular bases for circadian clocks*. Cell, 1999. **96**(2): p. 271-290.
78. Young, M.W. and S.A. Kay, *Time zones: A comparative genetics of circadian clocks*. Nature Reviews Genetics, 2001. **2**(9): p. 702-715.
79. Weinberger, L.S. and T. Shenk, *An HIV feedback resistor: auto-regulatory circuit deactivator and noise buffer*. PLoS Biol, 2007. **5**(1): p. e9.
80. Lutz, R. and H. Bujard, *Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/II-I2 regulatory elements*. Nucleic Acids Res, 1997. **25**(6): p. 1203-10.
81. Rasmussen, B., H.F. Noller, G. Daubresse, B. Oliva, et al., *Molecular basis of tetracycline action: identification of analogs whose primary target is not the bacterial ribosome*. Antimicrob Agents Chemother, 1991. **35**(11): p. 2306-11.
82. Oliva, B., G. Gordon, P. McNicholas, G. Ellestad, and I. Chopra, *Evidence that tetracycline analogs whose primary target is not the bacterial ribosome cause lysis of Escherichia coli*. Antimicrob Agents Chemother, 1992. **36**(5): p. 913-9.
83. Degenkolb, J., M. Takahashi, G.A. Ellestad and W. Hillen, *Structural requirements of tetracycline-Tet repressor interaction: determination of equilibrium binding constants for tetracycline analogs with the Tet repressor*. Antimicrob Agents Chemother, 1991. **35**(8): p. 1591-5.
84. Becskei, A., B. Seraphin and L. Serrano, *Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion*. Embo Journal, 2001. **20**(10): p. 2528-35.
85. Isaacs, F.J., J. Hastay, C.R. Cantor and J.J. Collins, *Prediction and measurement of an autoregulatory genetic module*. Proc Natl Acad Sci U S A, 2003. **100**(13): p. 7714-9.
86. Finzi, D., M. Hermankova, T. Pierson, L.M. Carruth, et al., *Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy*. Science, 1997. **278**(5341): p. 1295-300.
87. Chun, T.W., L. Stuyver, S.B. Mizell, L.A. Ehler, et al., *Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy*. Proc Natl Acad Sci U S A, 1997. **94**(24): p. 13193-7.

88. Pierson, T., J. McArthur and R.F. Siliciano, *Reservoirs for HIV-1: mechanisms for viral persistence in the presence of antiviral immune responses and antiretroviral therapy*. *Annu Rev Immunol*, 2000. **18**: p. 665-708.
89. Jordan, A., D. Bisgrove and E. Verdin, *HIV reproducibly establishes a latent infection after acute infection of T cells in vitro*. *Embo J*, 2003. **22**(8): p. 1868-77.
90. Jordan, A., P. Defechereux and E. Verdin, *The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation*. *Embo J*, 2001. **20**(7): p. 1726-38.
91. Lin, X., D. Irwin, S. Kanazawa, L. Huang, et al., *Transcriptional profiles of latent human immunodeficiency virus in infected individuals: effects of Tat on the host and reservoir*. *J Virol*, 2003. **77**(15): p. 8227-36.
92. Han, Y., M. Wind-Rotolo, H.C. Yang, J.D. Siliciano, and R.F. Siliciano, *Experimental approaches to the study of HIV-1 latency*. *Nat Rev Microbiol*, 2007. **5**(2): p. 95-106.
93. Lassen, K., Y. Han, Y. Zhou, J. Siliciano, and R.F. Siliciano, *The multifactorial nature of HIV-1 latency*. *Trends Mol Med*, 2004. **10**(11): p. 525-31.
94. Cullen, B.R., *Nuclear mRNA export: insights from virology*. *Trends Biochem Sci*, 2003. **28**(8): p. 419-24.
95. Pagans, S., A. Pedal, B.J. North, K. Kaehlcke, et al., *SIRT1 Regulates HIV Transcription via Tat Deacetylation*. *PLoS Biol*, 2005. **3**(2): p. e41.
96. Perelson, A.S., A.U. Neumann, M. Markowitz, J.M. Leonard, and D.D. Ho, *HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time*. *Science*, 1996. **271**(5255): p. 1582-6.
97. Klotman, M.E., S. Kim, A. Buchbinder, A. DeRossi, et al., *Kinetics of expression of multiply spliced RNA in early human immunodeficiency virus type 1 infection of lymphocytes and monocytes*. *Proc Natl Acad Sci U S A*, 1991. **88**(11): p. 5011-5.
98. Hooshangi, S. and R. Weiss, *The effect of negative feedback on noise propagation in transcriptional gene networks*. *Chaos*, 2006. **16**(2): p. -.
99. Golding, I., J. Paulsson, S.M. Zawilski and E.C. Cox, *Real-time kinetics of gene activity in individual bacteria*. *Cell*, 2005. **123**(6): p. 1025-36.
100. Pedraza, J.M. and J. Paulsson, *Effects of molecular memory and bursting on fluctuations in gene expression*. *Science*, 2008. **319**(5861): p. 339-43.
101. Cai, L., N. Friedman and X.S. Xie, *Stochastic protein expression in individual cells at the single molecule level*. *Nature*, 2006. **440**(7082): p. 358-62.
102. Newman, J.R., S. Ghaemmaghami, J. Ihmels, D.K. Breslow, et al., *Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise*. *Nature*, 2006. **441**(7095): p. 840-6.
103. Boettiger, A.N. and M. Levine, *Synchronous and stochastic patterns of gene activation in the Drosophila embryo*. *Science*, 2009. **325**(5939): p. 471-3.
104. Degenhardt, T., K.N. Rybakova, A. Tomaszewska, M.J. Mone, et al., *Population-level transcription cycles derive from stochastic timing of single-cell transcription*. *Cell*, 2009. **138**(3): p. 489-501.

105. Raj, A., C.S. Peskin, D. Tranchina, D.Y. Vargas, and S. Tyagi, *Stochastic mRNA synthesis in mammalian cells*. PLoS Biol, 2006. **4**(10): p. e309.
106. Cohen, A.A., T. Kalisky, A. Mayo, N. Geva-Zatorsky, et al., *Protein dynamics in individual human cells: experiment and theory*. PLoS ONE, 2009. **4**(4): p. e4901.
107. Blake, W.J., G. Balazsi, M.A. Kohanski, F.J. Isaacs, et al., *Phenotypic consequences of promoter-mediated transcriptional noise*. Mol Cell, 2006. **24**(6): p. 853-65.
108. Suel, G.M., R.P. Kulkarni, J. Dworkin, J. Garcia-Ojalvo, and M.B. Elowitz, *Tunability and noise dependence in differentiation dynamics*. Science, 2007. **315**(5819): p. 1716-1719.
109. Cai, L., C.K. Dalal and M.B. Elowitz, *Frequency-modulated nuclear localization bursts coordinate gene regulation*. Nature, 2008. **455**(7212): p. 485-90.
110. Bar-Even, A., J. Paulsson, N. Maheshri, M. Carmi, et al., *Noise in protein expression scales with natural protein abundance*. Nat Genet, 2006. **38**(6): p. 636-43.
111. Mitchell, R.S., B.F. Beitzel, A.R. Schroder, P. Shinn, et al., *Retroviral DNA Integration: ASLV, HIV, and MLV Show Distinct Target Site Preferences*. PLoS Biol, 2004. **2**(8): p. E234.
112. Schroder, A.R., P. Shinn, H. Chen, C. Berry, et al., *HIV-1 integration in the human genome favors active genes and local hotspots*. Cell, 2002. **110**(4): p. 521-9.
113. Burnett, J.C., K. Miller-Jensen, P.S. Shah, A.P. Arkin, and D.V. Schaffer, *Control of stochastic gene expression by host factors at the HIV promoter*. PLoS Pathog, 2009. **5**(1): p. e1000260.
114. Kim, D.W., T. Uetsuki, Y. Kaziro, N. Yamaguchi, and S. Sugano, *Use of the human elongation factor 1 alpha promoter as a versatile and efficient expression system*. Gene, 1990. **91**(2): p. 217-23.
115. Ramezani, A., T.S. Hawley and R.G. Hawley, *Lentiviral vectors for enhanced gene expression in human hematopoietic cells*. Mol Ther, 2000. **2**(5): p. 458-69.
116. Weinberger, L.S., R.D. Dar and M.L. Simpson, *Transient-mediated fate determination in a transcriptional circuit of HIV*. Nat Genet, 2008. **40**(4): p. 466-70.
117. Li, X.Q., X.N. Zhao, Y. Fang, X. Jiang, et al., *Generation of destabilized green fluorescent protein transcription reporter*. Journal of Biological Chemistry, 1998. **273**(52): p. 34970-34975.
118. Gilchrist, D.A., G. Dos Santos, D.C. Fargo, B. Xie, et al., *Pausing of RNA Polymerase II Disrupts DNA-Specified Nucleosome Organization to Enable Precise Gene Regulation*. Cell, 2010. **143**(4): p. 540-551.
119. Vallabhapurapu, S. and M. Karin, *Regulation and function of NF-kappaB transcription factors in the immune system*. Annu Rev Immunol, 2009. **27**: p. 693-733.
120. Epstein, C.B. and R.A. Butow, *Microarray technology - enhanced versatility, persistent challenge*. Curr Opin Biotechnol, 2000. **11**(1): p. 36-41.

121. Schulze, A. and J. Downward, *Navigating gene expression using microarrays--a technology review*. Nat Cell Biol, 2001. **3**(8): p. E190-5.
122. Heller, M.J., *DNA microarray technology: devices, systems, and applications*. Annu Rev Biomed Eng, 2002. **4**: p. 129-53.
123. Gasch, A.P., P.T. Spellman, C.M. Kao, O. Carmel-Harel, et al., *Genomic expression programs in the response of yeast cells to environmental changes*. Molecular Biology of the Cell, 2000. **11**(12): p. 4241-4257.
124. Basehoar, A.D., S.J. Zanton and B.F. Pugh, *Identification and distinct regulation of yeast TATA box-containing genes*. Cell, 2004. **116**(5): p. 699-709.
125. Yang, C.H., E. Bolotin, T. Jiang, F.M. Sladek, and E. Martinez, *Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters*. Gene, 2007. **389**(1): p. 52-65.
126. Blake, W.J., G. Balazsi, M.A. Kohanski, F.J. Isaacs, et al., *Phenotypic consequences of promoter-mediated transcriptional noise*. Molecular Cell, 2006. **24**(6): p. 853-865.
127. Lee, W., D. Tillo, N. Bray, R.H. Morse, et al., *A high-resolution atlas of nucleosome occupancy in yeast*. Nature Genetics, 2007. **39**(10): p. 1235-1244.
128. Huisinga, K.L. and B.F. Pugh, *A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in Saccharomyces cerevisiae*. Molecular Cell, 2004. **13**(4): p. 573-585.
129. Lopez-Maury, L., S. Marguerat and J. Bahler, *Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation*. Nature Reviews Genetics, 2008. **9**(8): p. 583-593.
130. Landry, C.R., B. Lemos, S.A. Rifkin, W.J. Dickinson, and D.L. Hartl, *Genetic properties influencing the evolvability of gene expression*. Science, 2007. **317**(5834): p. 118-121.
131. Tirosh, I. and N. Barkai, *Two strategies for gene regulation by promoter nucleosomes*. Genome Research, 2008. **18**(7): p. 1084-1091.
132. Chang, D.E., D.J. Smalley and T. Conway, *Gene expression profiling of Escherichia coli growth transitions: an expanded stringent response model*. Molecular Microbiology, 2002. **45**(2): p. 289-306.
133. White-Ziegler, C.A., S. Um, N.M. Prez, A.L. Berns, et al., *Low temperature (23 degrees C) increases expression of biofilm-, cold-shock- and RpoS-dependent genes in Escherichia coli K-12*. Microbiology-Sgm, 2008. **154**: p. 148-166.
134. Khil, P.P. and R.D. Camerini-Otero, *Over 1000 genes are involved in the DNA damage response of Escherichia coli*. Molecular Microbiology, 2002. **44**(1): p. 89-105.
135. Traxler, M.F., S.M. Summers, H.T. Nguyen, V.M. Zacharia, et al., *The global, ppGpp-mediated stringent response to amino acid starvation in Escherichia coli*. Molecular Microbiology, 2008. **68**(5): p. 1128-1148.
136. Franchini, A.G. and T. Egli, *Global gene expression in Escherichia coli K-12 during short-term and long-term adaptation to glucose-limited continuous culture conditions*. Microbiology-Sgm, 2006. **152**: p. 2111-2127.

137. Tani, T.H., A. Khodursky, R.M. Blumenthal, P.O. Brown, and R.G. Matthews, *Adaptation to famine: A family of stationary-phase genes revealed by microarray analysis*. Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(21): p. 13471-13476.
138. Hayes, E.T., J.C. Wilks, P. Sanfilippo, E. Yohannes, et al., *Oxygen limitation modulates pH regulation of catabolism and hydrogenases, multidrug transporters, and envelope composition in Escherichia coli K-12*. BMC Microbiology, 2006. **6**: p. -.
139. Sangurdekar, D.P., F. Sreenc and A.B. Khodursky, *A classification based framework for quantitative description of large-scale microarray data*. Genome Biology, 2006. **7**(4): p. -.
140. Kershaw, C.J., N.L. Brown, C. Constantinidou, M.D. Patel, and J.L. Hobman, *The expression profile of Escherichia coli K-12 in response to minimal, optimal and excess copper concentrations*. Microbiology-Sgm, 2005. **151**: p. 1187-1198.
141. Eisen, M.B., P.T. Spellman, P.O. Brown and D. Botstein, *Cluster analysis and display of genome-wide expression patterns*. Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.
142. Grissom, J., Conway T. , <http://genexpdb.ou.edu> 2011.
143. Freyre-Gonzalez, J.A., J.A. Alonso-Pavon, L.G. Trevino-Quintanilla and J. Collado-Vides, *Functional architecture of Escherichia coli: new insights provided by a natural decomposition approach*. Genome Biology, 2008. **9**(10): p. R154.
144. Murray, J.I., M.L. Whitfield, N.D. Trinklein, R.M. Myers, et al., *Diverse and specific gene expression responses to stresses in cultured human cells*. Mol Biol Cell, 2004. **15**(5): p. 2361-74.
145. Dar, R.D., D.K. Karig, J.F. Cooke, C.D. Cox, and M.L. Simpson, *Distribution and regulation of stochasticity and plasticity in Saccharomyces cerevisiae*. Chaos, 2010. **20**(3): p. -.
146. Lehner, B., *Conflict between Noise and Plasticity in Yeast*. Plos Genetics, 2010. **6**(11): p. -.
147. Gama-Castro, S., H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, et al., *RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units)*. Nucleic Acids Res, 2011. **39**(Database issue): p. D98-105.
148. Rosenfeld, N., T.J. Perkins, U. Alon, M.B. Elowitz, and P.S. Swain, *A fluctuation method to quantify in vivo fluorescence data*. Biophysical Journal, 2006. **91**(2): p. 759-766.
149. Ghaemmaghani, S., W. Huh, K. Bower, R.W. Howson, et al., *Global analysis of protein expression in yeast*. Nature, 2003. **425**(6959): p. 737-741.
150. McCollum, J.M., G.D. Peterson, C.D. Cox, M.L. Simpson, and N.F. Samatova, *The sorting direct method for stochastic simulation of biochemical systems with varying reaction execution behavior*. Comput Biol Chem, 2006. **30**(1): p. 39-49.

151. Gillespie, D.T., *General Method for Numerically Simulating Stochastic Time Evolution of Coupled Chemical-Reactions*. *Journal of Computational Physics*, 1976. **22**(4): p. 403-434.
152. Raj, A., C.S. Peskin, D. Tranchina, D.Y. Vargas, and S. Tyagi, *Stochastic mRNA synthesis in mammalian cells*. *Plos Biology*, 2006. **4**(10): p. 1707-1719.

APPENDIX

7.1 Fundamentals and Methodology

7.1.1 Biased versus Unbiased Autocorrelation

For the case of limited signal acquisition the normalized biased (B) and unbiased (UB) composite autocorrelation functions (CACF's) differ in their scaling factors as follows:

$$\Phi_B(\tau) = \frac{\sum_{m=1}^M \sum_{n=0}^N X(nT_s)X(nT_s + \tau)}{\sum_{m=1}^M \sum_{n=0}^N X^2(nT_s)} \quad , \quad \Phi_{UB}(\tau) = \frac{\sum_{m=1}^M \sum_{n=0}^N \frac{X(nT_s)X(nT_s + \tau)}{(N \cdot T_s - \tau)}}{\sum_{m=1}^M \sum_{n=0}^N \frac{X^2(nT_s)}{(N \cdot T_s)}} \quad (7.1)$$

Where $X(nT_s)$ is the noise of cell m at time nT_s , T_s is the time sampling or interval, $n = 1, \dots, N$ are the number of imaging intervals in the experiment, and finally M is the total number of cells collected in the experiment.

Here the unbiased ACF is more accurate for the average AC of a particular time lag (τ), but less accurate with larger variations appearing at larger lag values. On the other hand the biased ACF has smaller variation over the time-lag domain having been scaled by the total signal duration and not part of it. The larger lag values are suppressed and tied down at the experiment duration. This difference between the biased and unbiased ACF becomes more significant as the number of data points approaches the lag number. Ultimately due to a predetermined under sampling of the measured noise signal the biased autocorrelation was preferred due to reduced error at smaller lag values, in particular the half-correlation time region ($\tau_{1/2}$).

7.1.2 Automated tracking of an adhered slow-growing cell monolayer

An alternative automated tracking scheme was applied for human T-cell experiments. The T-cell sample consists of an inverted imaging setup where the sample is imaged from below looking upwards. T-cells are adhered to a glass substrate with liquid media on top, under temperature and environmental control (37°C, CO₂, and humidity). For examples of a fluorescent T-cell image taken with the inverted Olympus spinning disc confocal microscope see Figures 2.14 and 3.13 (more details on T-cell imaging can be found in Chapter 3). Since mammalian cells are slow growing, in this case ~14 hour doubling time, the experiment objective was to only track adult T-cell expression. This is accomplished by dispersion of the un-adhered daughter cells into the liquid media environment and out of the imaging frame shortly after a doubling event occurs leaving the adhered adult cell behind in the same location as before. The single-cell segmentation and tracking is very similar to that previously described, but a tailored “*Segmentator*” program, depicted in figure 7.1, was needed to satisfy the following experiment demands: (1) the program must track all cells present and adhered for at least 4 hours (cells may blink and turn on/off, fly off, or adhere onto the substrate in the middle of the experiment) and (2) the program must account for unsynchronized expression behavior in which some cells turn on or off at different time points.

The resulting program (`segmentator_tifstablesweep.m`) included a novel pixel-logic approach for identifying and collecting cell trajectories (Fig. 7.1). A moving (or sweeping) 4 hour window (at 10 minute imaging intervals, 4 hours = 24 image window) was used across the experiment duration. For each window location single-cells were segmented and compared using a series of logic comparisons between the binary yes/no (1 or 0) cellular pixel array of each individually processed image. After a series of logic operations the program identified cells that are present throughout the 4 hour window. For the case of a 12 hour experiment the window sweep would stop and collect cells at hours 0-4, 2-6, 4-8, ..., and 8-12. The final step in the program was to compare cell centroid locations of all cells collected in all windows (Fig. 7.2). If cells were close enough to one another then those trajectories were combined and the group of identified

cells were considered the same cell identity. This aggregation of identical adhered cells assumes that no foreign cell occupies a cell's location before or after the cell is tracked.

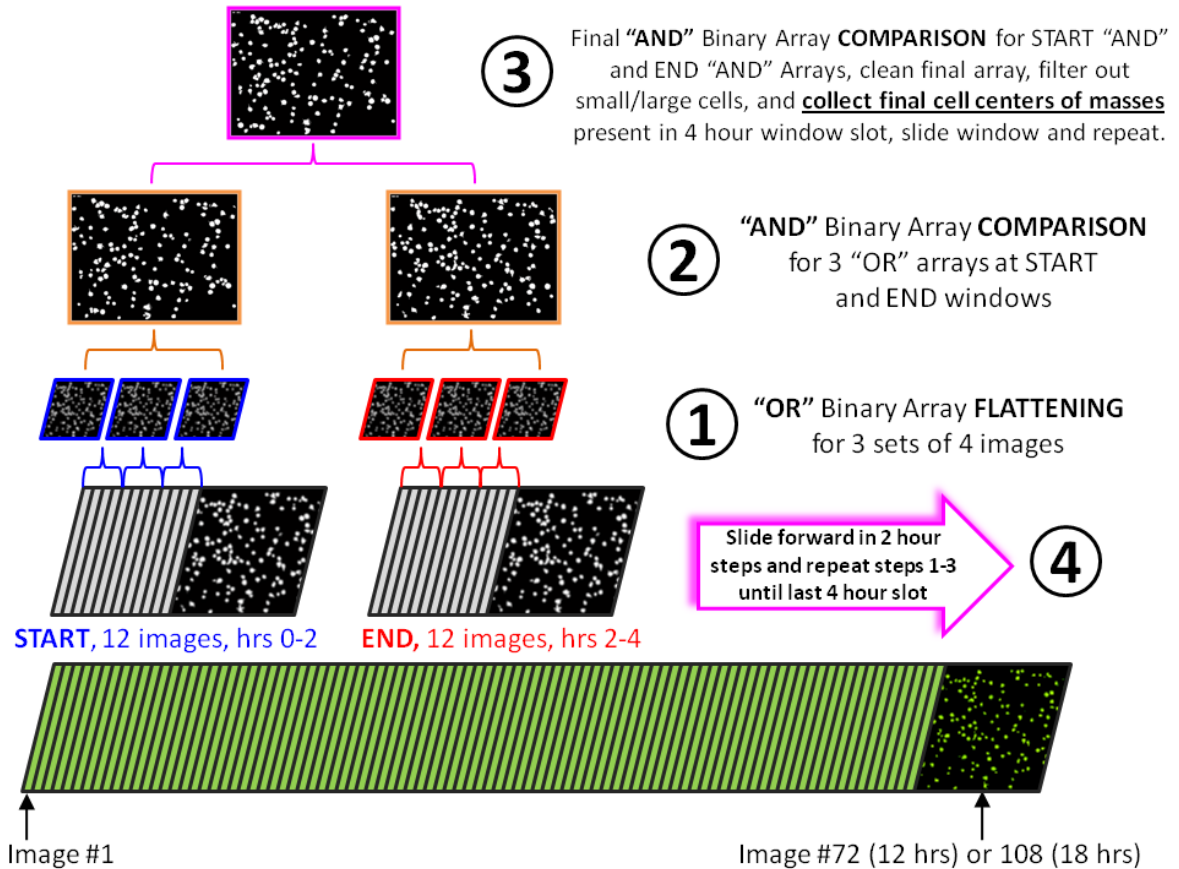


Figure 7.1 "Segmentator" program for segmentation and tracking of adhered T-cells. The program uses two 2 hour windows of binary labeled images and a series of logical comparisons (1-3) to identify and quantify cells present throughout an imaging experiment. After testing a 4-hour block (steps 1-3) the two comparison windows slide 2 hours forward and continue the routine until all 12 or 18 hours are processed (step 4).

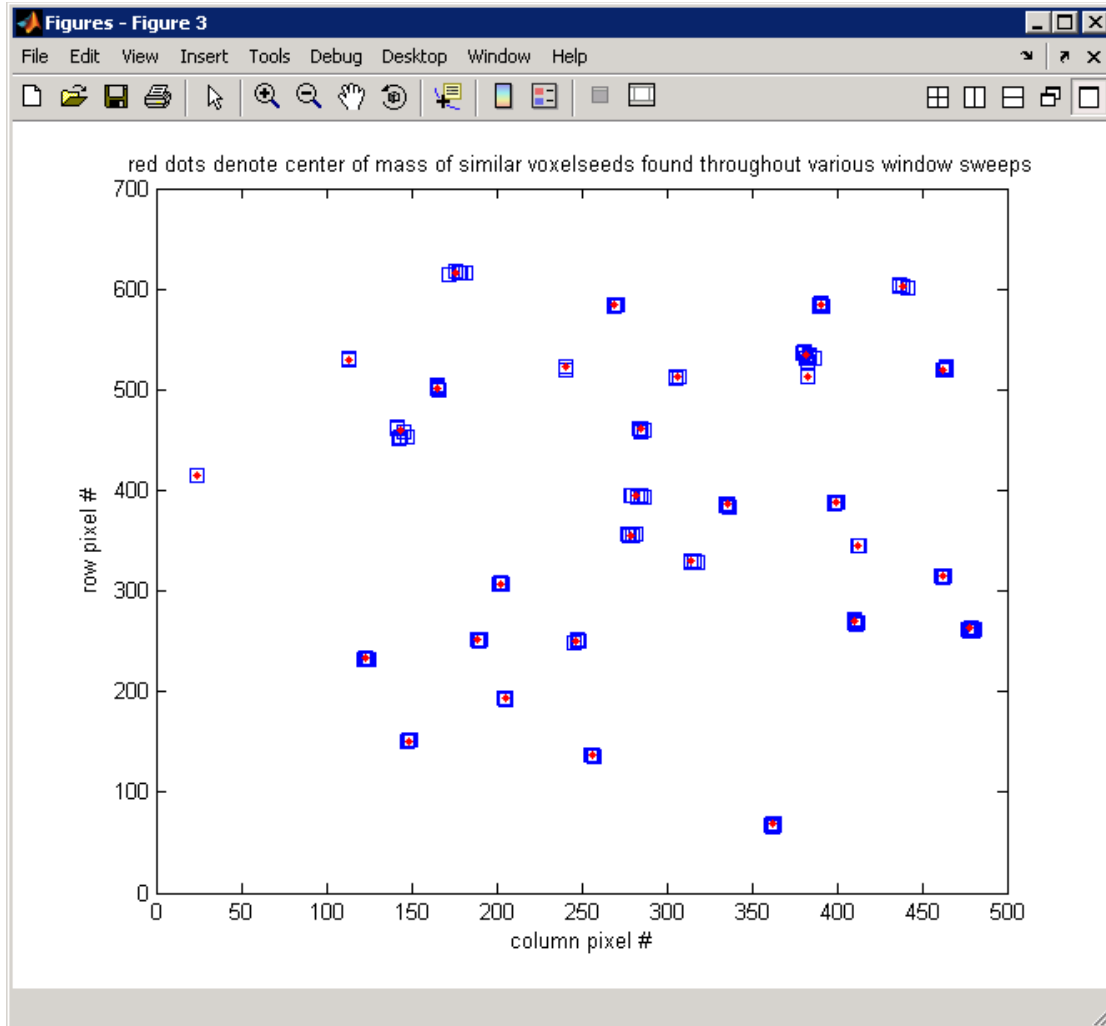


Figure 7.2 Clustering of oversampled cell seeds from automated tracking program. In the above individual blue squares represent a single cell center-of-mass detected in a specific 4 hour automated logic window. The program clusters cell seeds (blue squares) of identical cells and calculates a composite red seed which is used over the entire experiment duration.

7.1.3 Cellular Fluorescent Intensity and Fluorescent Protein Abundance are Correlated

One of the basic assumptions regarding quantitative fluorescence measurements is that within a linear range of measurement or detectability the number of fluorescent reporters (e.g. protein-GFP fusions) in the cell is directly proportional to the (arbitrary unit) fluorescence intensity quantified by image pixel intensity. The fluorescent intensity signal is considered a direct measure of protein levels inside the single cells. It is possible to convert between the two ($\langle FL \rangle$ and $\langle P \rangle$) by either using a binomial splitting calibration method based on cell doubling in *E. coli* [45, 148], single-molecule sensitivity in single cells using a YFP-protein fusion library in *E. coli* [32], or by system-wide protein abundance quantification using standard molecular biology approaches, [149]. A recent review article on methods and studies of single-molecule gene expression is reported by Larson et al, (2009) [43].

The correlation between fluorescence and abundance is clearly seen in figure 7.3 which was derived using online supplementary datasets supplied by Newman et al, (2006) [33] ($\langle FL \rangle$ from flow cytometry) and Ghaemmaghami et al, (2003) [149] (budding yeast protein abundance per cell). In addition to seeing the correlation between the two, these plots, which accounted for different sub-cellular protein localizations, yields information about the quantum yield of a single protein-GFP fusion as a function of its sub-cellular localization. Consistent in two types of media in the Newman study (upper plot is for rich YEPD and lower plot is using minimal SD+), small sub-cellular component volumes such as the bud neck, bud, mitochondrion, and nucleolus seem to consistently have lower quantum yields than the nucleus and cytoplasm localizations. In simple terms the data shows that two genes with the same protein abundance per cell can have drastically different average fluorescence values (and noise magnitude or CV) based on their sub-cellular localization. This quantum yield difference is heavily dependent on sub-cellular component volume and may be attributed to variations in 2D versus 3D mobility or even sequestered, non-fluorescent, or inactive protein-GFP fusions in a

confined cellular region. These issues may need to be considered when comparing protein-GFP noise of two genes localized in two different sub-cellular components.

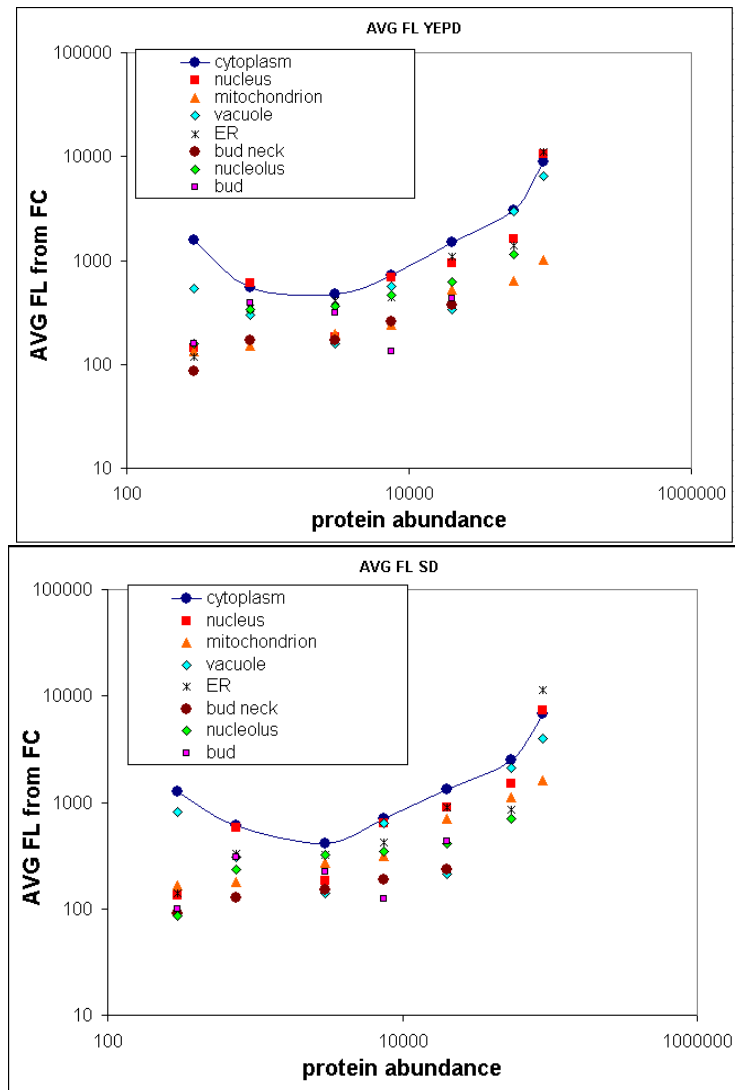


Figure 7.3 Fluorescence intensity and fluorescent protein abundance are correlated. The above plots were derived and calculated from the system-wide budding yeast flow cytometry measurements by Newman et al, [102], and the Ghaemmaghami et al system-wide yeast protein abundance study [149]. Trends of protein abundance are plotted versus their average fluorescence measured with flow cytometry for a specific protein abundance range and for various sub-cellular component localizations in both rich YEPD medium (upper) and minimal SD medium (lower). In both types of media the cytoplasm and nucleus trends are found to be higher with greater quantum efficiency, while small volume components like the bud neck, bud, nucleolus, and mitochondrion are found consistently towards the bottom with lower efficiencies.

7.1.4 Stochastic Simulation and Gillespie's Algorithm [52]

7.1.4.1 Basic concept

1. Assume a simple transcription-translation gene expression model such as:

Table 7.1 Simulation of a basic gene expression model

Chemical Reaction	Rate Constant
$G \rightarrow G + M$	k_m
$M \rightarrow M + P$	k_p
$M \rightarrow *$	γ_m
$P \rightarrow *$	γ_p

2. Each molecular species has an initial population (G_0, M_0, P_0).
3. The algorithm starts by choosing two random numbers from $[0,1]$; one to select which reaction to perform, and is weighted by (rate constant) * (molecular species population or concentration), e.g. $[G]*k_m$ for transcribing mRNA, $[M]*k_p$ for translating protein P, etc., and the second is for the time at which to perform the reaction (Δt) and is sampled from an exponential distribution.
4. After the reaction is performed all of the molecular populations are updated, the time is advanced by $t + \Delta t$ and the algorithm repeats itself iteratively.

This Monte Carlo simulation of coupled chemical reactions is equivalent to the chemical master equation, is exact for both low and high numbers of molecules, and is computationally expensive such that simulations should be planned carefully.

7.1.4.2 Considerations when simulating

In simulating any genetic model it is important to understand what processes and steps can be ignored, simplified, or merged together. At least with regards to noise correlations, these are usually short lived processes that don't modify the primary time-constants (and Frequency-Domain poles) of the system. There is a motivation to simplify the model as it minimizes the number of unknown parameters and the number of reactions in the simulation. For some of the most studied systems in biology (e.g. the Lac Operon), many parameters have simply not been measured, yet at times it is still possible to constrain parameter value ranges using experimental data (e.g. see ATc-ribosome extrinsic noise simulation in Chapter 3 from Austin *et al*, 2006 [70], or stochastic simulation of quorum sensing by Cox *et al*, 2003 [72]). In addition, it is possible to run a careful sensitivity analysis for high dimensional models with large numbers of parameters and start to narrow in on which parameters dominate the system's dynamics the most.

7.1.4.3 Biospreadsheet: A User-Friendly Simulator

Biospreadsheet (available for download at <http://biocomp.ece.utk.edu> or by request from Mike McCollum) is a Java based exact stochastic simulation (ESS) program designed to automate simulations and was created by Mike McCollum (currently at VCU). It implements an optimized version [150] of the Gillespie algorithm[52].

Biospreadsheet has a user friendly interface. It starts with an "Information" tab (Fig. 7.4) which allows the author to name the program, identify themselves as the author, and use the additional description space to insert details regarding the simulation. The second tab is a "Species" tab in which the user inputs the names or symbol of which chemical (or molecular) species is in the model, their initial populations, and finally which time-dependent species output is desired (Fig. 7.5). By checking additional boxes the program can output all of the species selected, they are added as additional columns of data in the output tab-delimited file. Third is the "Reactions" tab which details which species interact with one another, how they interact, how species are produced, and how they are degraded (Fig 7.6). Next to the reactions column are the rates at which each

reaction occurs. Finally, to check and run the simulation the user clicks under “Model” in the taskbar and selects “Check” or “Simulate”.

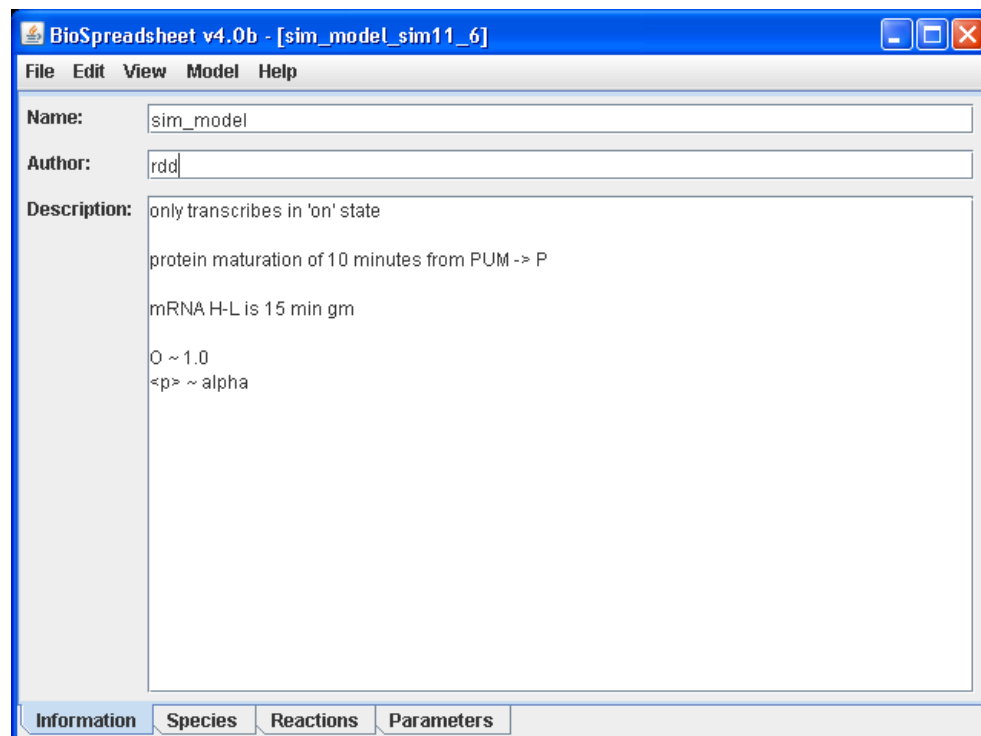


Figure 7.4 Information tab of BioSpreadsheet Simulator.

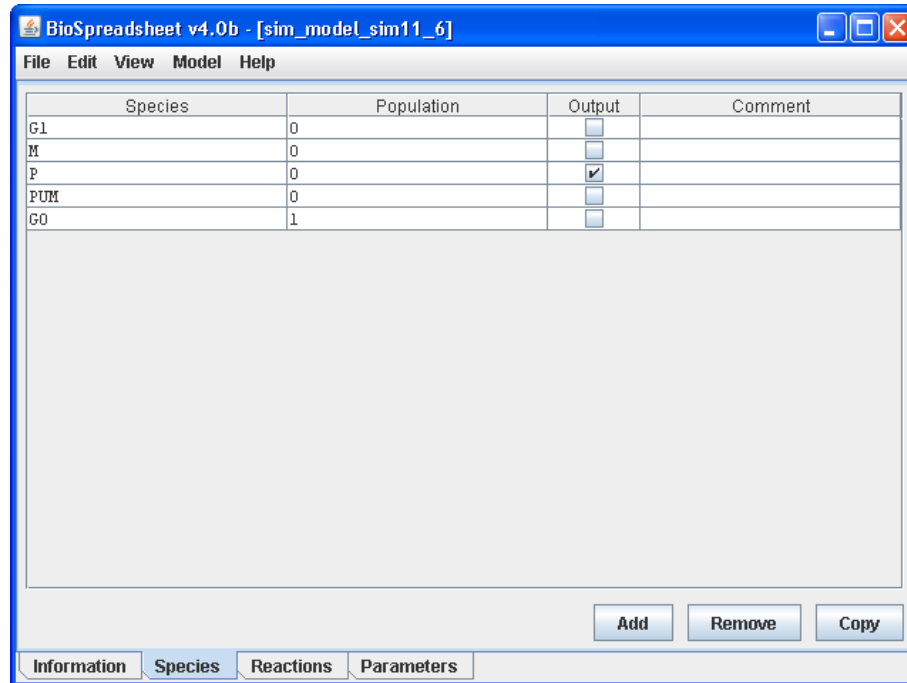


Figure 7.5 Species tab of BioSpreadsheet Simulator.

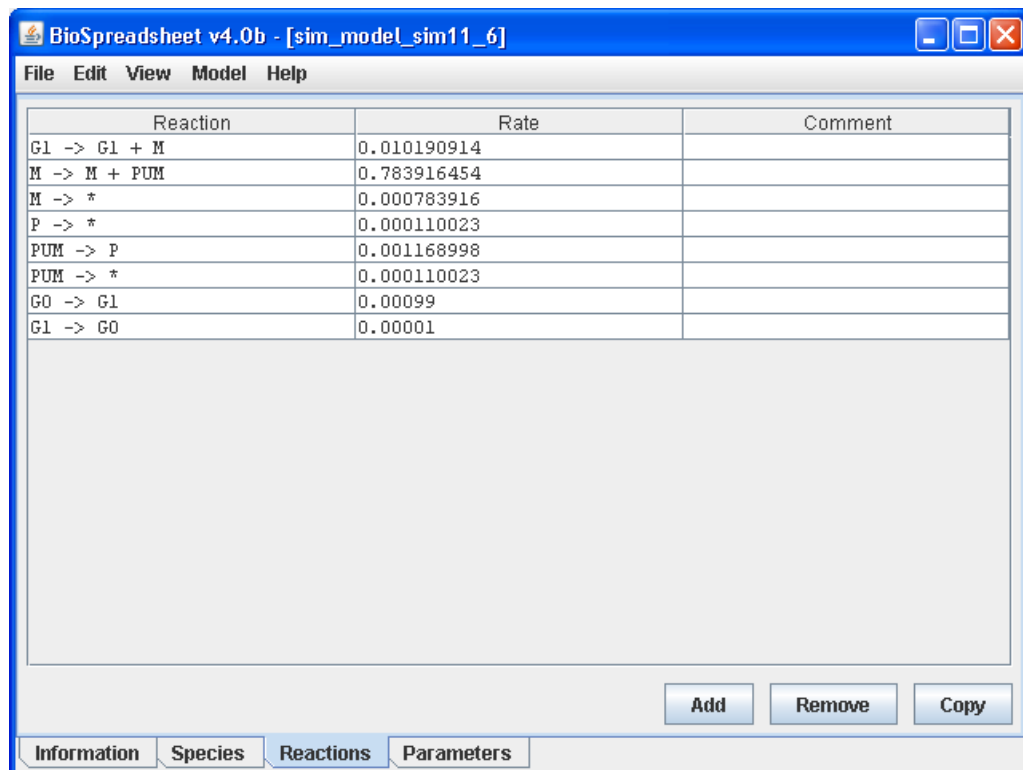


Figure 7.6 Reactions tab of BioSpreadsheet Simulator.

The simulation window is shown in figure 7.7. Here the user defines the start, end, and interval times in seconds, as well as the file name convention and the number of files to output under different seeds (seeds to random number generation used by the program). Regarding the initial time to start recording the simulation it is important to note that if the initial population of the output protein is not at its steady-state value then there will be an initial transient behavior to the simulation. To avoid this it is worth setting the initial start time to many times the dominant time constant of the system being simulated. Another important detail is how long of a simulation is desired. If single cell simulations are desired it is possible to use the seed based file generation. Another option is to simulate a very long trajectory and then cut it up into individual single cell trajectories. To avoid correlation or similarity between trajectories it is important to simulate enough data to buffer independence between individual cell trajectories.

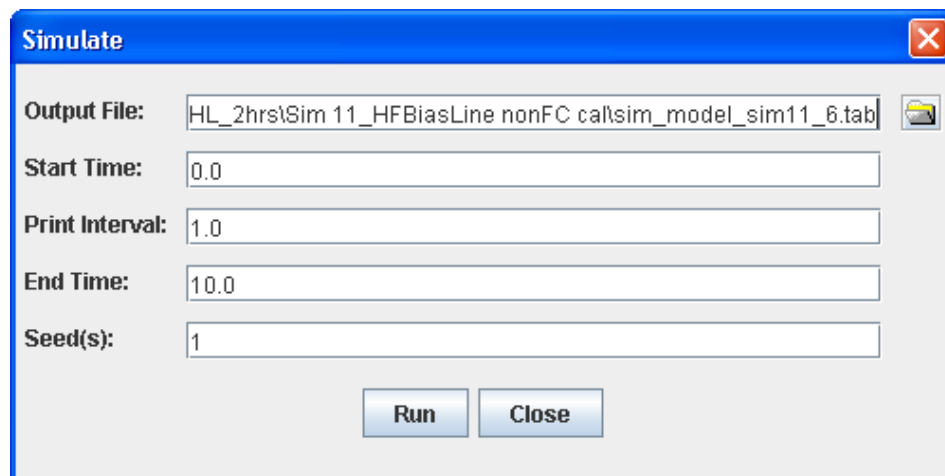


Figure 7.7 Simulation settings tab in BioSpreadsheet.

7.1.4.4 An Example of Stochastic Simulation

The snapshots of the previous “Reactions” simulation tab actually describe a working simulation model for 2-state transcriptional bursting. In this case the simulation is identical to table 7.1 in section 7.1.4.1, but has the additional switching between a non-expressing basal gene state (G0) and an elevated transcriptional state (G1) at a switching frequency of $k_{on} + k_{off}$. In addition, the GFP half-life (H-L) of the simulation was 2 hours, cell dilution was set to a 14 hour doubling time, the mRNA H-L was 15 minutes, the protein maturation time was 10 minutes (PUnMat \rightarrow PMat), and the simulation was run for two different average “on” times ($O = k_{on} / (k_{on} + k_{off})$). For constitutive expression the model was set as the reaction rates show in Fig 7.6, to a high on time, $O = 0.99$, i.e. constantly in the elevated transcription rate (Fig. 7.8, upper). For simulating transcriptionally bursty gene expression (Fig 7.8, lower) all of the rates were left the same, except for the on time which was set to $O=0.25$, and the elevated transcription rate was increased to yield an equivalent mean protein level to the first simulation ($\langle p \rangle = \sim 8.5k$, blue steady state line in Fig 7.8 upper and lower).

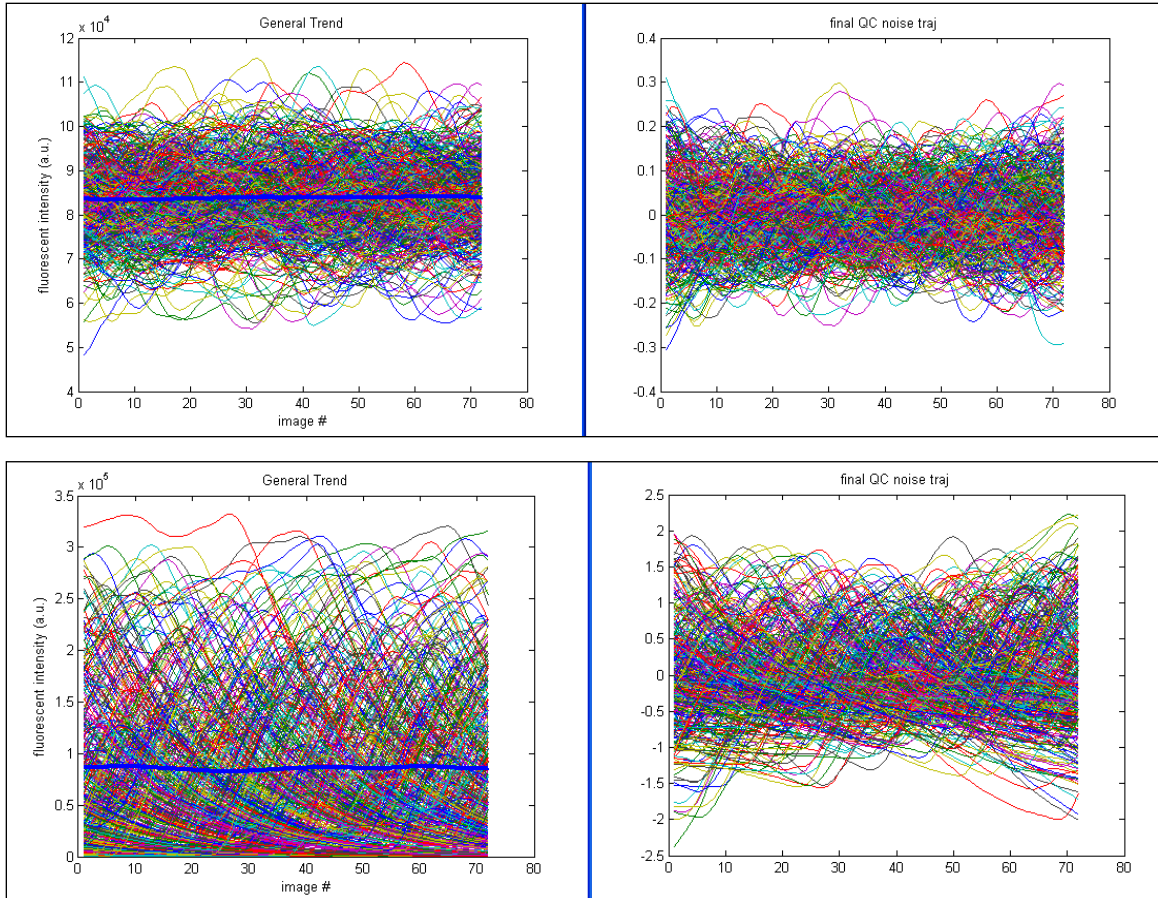


Figure 7.8 Stochastic simulation of single cell gene expression. (upper) Constitutive gene expression was simulated (upper-left) and noise was processed (upper-right) for 500 individual 12 hour single-cell trajectories (image interval of 10 minutes gives 72 total images). (lower) Simulation of transcriptional bursting with a two state gene expression model. The gene is in the elevated transcription state 25% of the time and the basal state is non-expressive.

7.1.5 Manual Quality Control of Acquired Fluorescence Signals

The detailed image processing protocol described above would not be complete without a critical manual quality control (by a person) of each single cell intensity trajectory collected to assure that all processed trajectories are of high quality. This is done by a fairly straightforward program that flashes each trajectory in front of the user and enables the trajectory to be marked as PASS, FAIL, or additional subgroups of interest. For example if there is a specific phenotypic signature of interest such as an intensity spike, rise, or drop in gene expression that needs to be analyzed separately, this is a convenient time to separate out such features into individual sub-groups for further analysis. If needed, these sorting tasks may also be automated. After finally acquiring the high quality raw fluorescence intensity trajectories, a multiple step signal processing protocol is implemented.

7.2 The Coupling of Gene Circuit and Noise Structures

7.2.1 Half-Correlation Time Error Bar Estimation

High frequency half-correlation time (HF- $\tau_{1/2}$) error bars were estimated using exact stochastic simulation (ESS) of the simplified 2-pole model with no feedback described below (Table 7.2). Principles of ESS can be found above in Appendix section 7.1.4. Stochastic simulation software (BioSpreadsheet; available for download at <http://biocomp.ece.utk.edu>) was used to generate time series data, and custom software was used to generate composite autocorrelation functions using different selected ensemble number (M) of single cell trajectories. From the simulations, many different collections of cells were created for each population size (M) and the high frequency $\tau_{1/2}$ for each of these collections was calculated. The collections of cells were selected from a simulated population of 3000 uncorrelated 12 hour single cell trajectories. The standard deviation in the HF- $\tau_{1/2}$ was calculated for all of the collections of each value of M resulting in a 1- σ half correlation time error of ~ 0.1 hours for cell ensembles greater than 30 single cells (Fig. 7.9 below). All stochastic simulations were based on variations of the Gillespie stochastic simulation algorithm [52, 73, 151].

Table 7.2 2-pole stochastic simulation model

Reaction	Rate
1. $G \rightarrow G + M$	k_M
2. $M \rightarrow M + P$	k_P
3. $M \rightarrow *$	γ_M
4. $P \rightarrow *$	γ_P

The lowest frequency pole (f_{GFP}) was set by reaction 4 to be $\gamma_P/2\pi$, using a protein half-life of 7 hours. The second pole (f_{mRNA}) was set by reaction 3 to be $\gamma_M/2\pi$, using an

mRNA half-life of 10 minutes. M and P production rates were set by using a burst of 100 and $\langle M \rangle = 10$ for all simulated experiments.

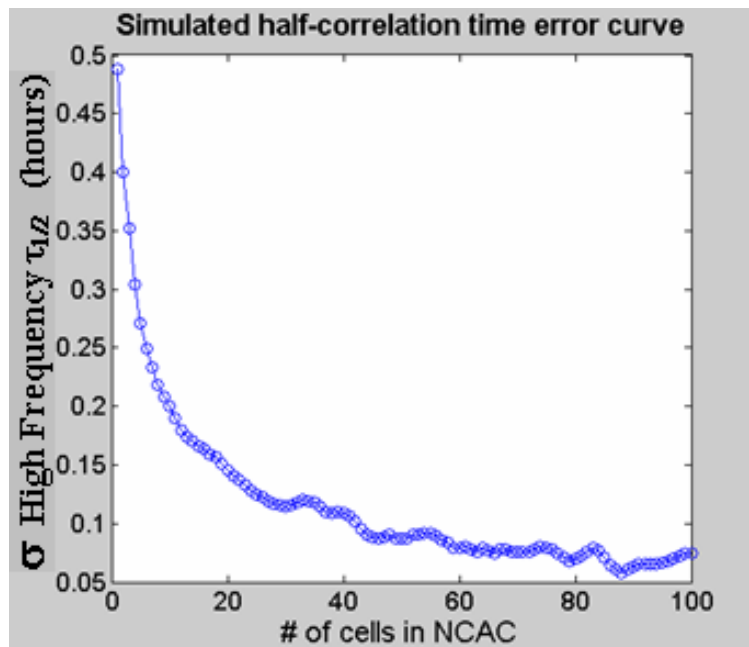


Figure 7.9 Simulated error in 12 hour HF-T50 as a function of number of cells in the collected ensemble. NCAC stands for normalized composite autocorrelation. After sampling more than 30 cells in a population the half-correlation error is reduced to less than ± 0.1 hours.

7.3 Noise Mapping

7.3.1 Advantages and Disadvantages of Polyclonal Noise Mapping

Measurement of single cell gene expression from a polyclonal sample has advantages and disadvantages.

The main disadvantages include:

1. Low sampling statistics as a unique integration site is represented by only one cell. Upon collecting thousands of genomic loci, due to the limited sampling, the emergent picture is ‘fuzzy’ and may be resolution limited when studying certain biological systems.
2. Currently there is no way to know the precise and characterized region in which the reporting vector has integrated. Is it within the control of a native promoter, NF κ B site, TATA-box, and abundant nucleosome occupancy? All of these would be resolved if it were possible to isolate, grow out, and sequence specific cells of interest (e.g. perhaps a future capability would enable the selection of a cell in real time that has a specific noise map signature for isolation, growth, and sequencing – presenting noise driven integration site selection which may compliment a synthetic biology toolbox).

The main advantages include:

1. Within a short time-frame of 1-2 weeks of imaging experiments fair genome-wide coverage may be acquired.
2. There is no need to create a genome-wide protein-GFP library such as those mentioned in Chapter 2, but instead only a lenti-viral vector delivered gene circuit.
3. The high-throughput genomic dynamics measurements may help establish a new research field of “Noise Omics”. Further contributing to systems biology understanding of the cell by lead to a large number of novel studies to research ‘information transport’ in complex systems. Including the integration of more complex circuitry such as feedback circuits, different cell types or cellular states,

and additional signaling molecules of interest. The current study described the basal integration site burst landscape on top of which all genetic circuitry must function.

7.3.2 Cell-cycle synchronization

Cell synchronization was performed to control for differences in cell cycle or state. The synchronized cell population (85%-90% synchronized in G1) was imaged using the same imaging parameters as the non-synchronized cells above to obtain noise maps and NPD maps (Fig. 7.10). The NPD maps for a G1-synchronized Ld2G population were then compared to NPD maps of an unsynchronized Ld2G population to create a NPD-difference map in order to check for any differences between synchronized and unsynchronized cells. No significant difference could be found between synchronized and unsynchronized Ld2G cells as evidenced by the interspersed NPD-difference map (Fig. 7.10).

To further check for differences between synchronized and unsynchronized cell populations, we compared the HF-T50 distributions (Fig. 7.11). Both synchronized and unsynchronized cells displayed a significant shift in HF-T50 compared to the constitutive model (Fig. 7.11 below) with the mean of the shift being the same for both synchronized and unsynchronized Ld2G populations.

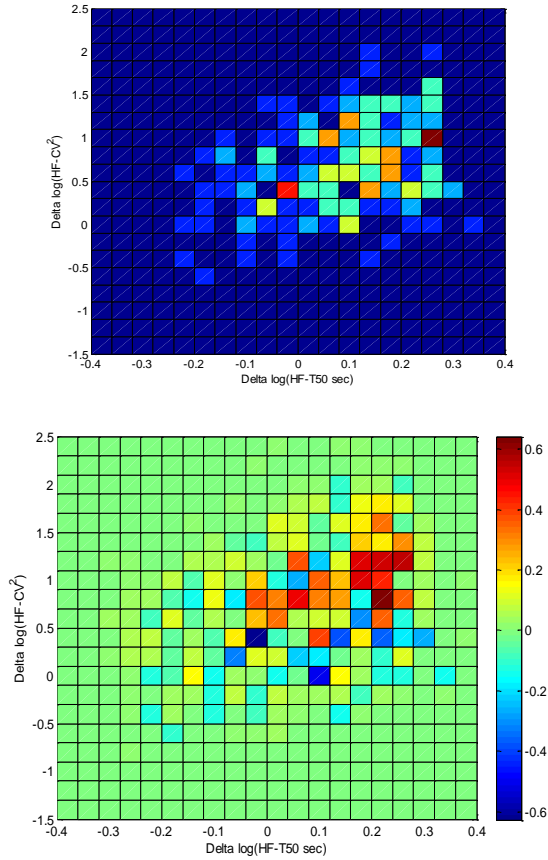


Figure 7.10 NPD and difference map for synchronized and unsynchronized cells. **Upper:** NPD map for a synchronized Ld2G populations of cells. **Lower:** NPD-difference map between a synchronized Ld2G population versus an unsynchronized Ld2G population (from Fig 2b-c, main text) demonstrating no significant difference between the two polyclonal populations. NPD comparison = $(\text{NPD_Ld2G_unsynchronized}) - (\text{NPD_Ld2G_synchronized})$.

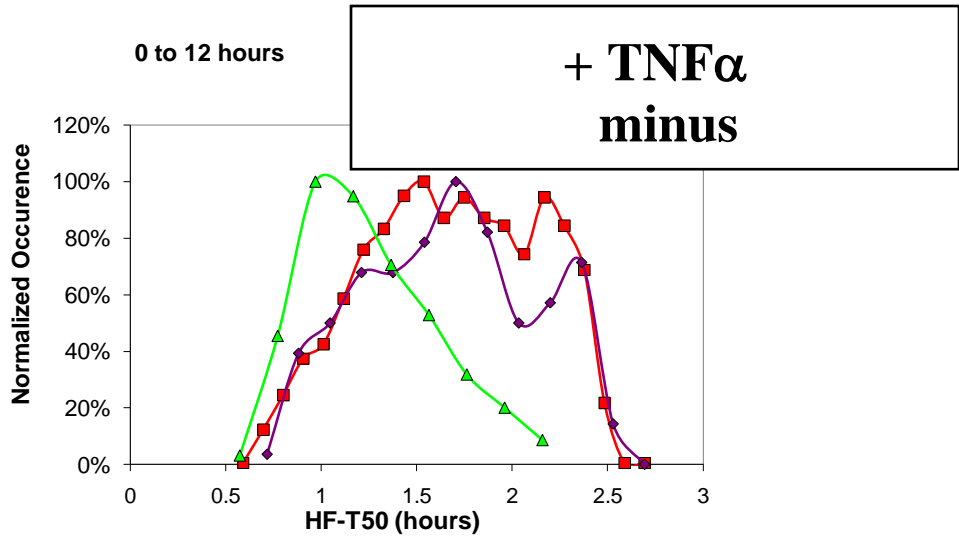


Figure 7.11 HF-T50 distribution comparison for synchronized and unsynchronized cells. Here the constitutive model (green) is compared to an unsynchronized Ld2G cell population (red) and a synchronized Ld2G cell population (purple). The constitutive distribution is the measured ‘most constitutive’ distribution described above for clones C32 and D36. No significant difference could be detected between the means or medians of the Ld2G synchronized vs unsynchronized populations; peaks in the synchronized Ld2G cell population are due to the small cell number compared to the unsynchronized population,.

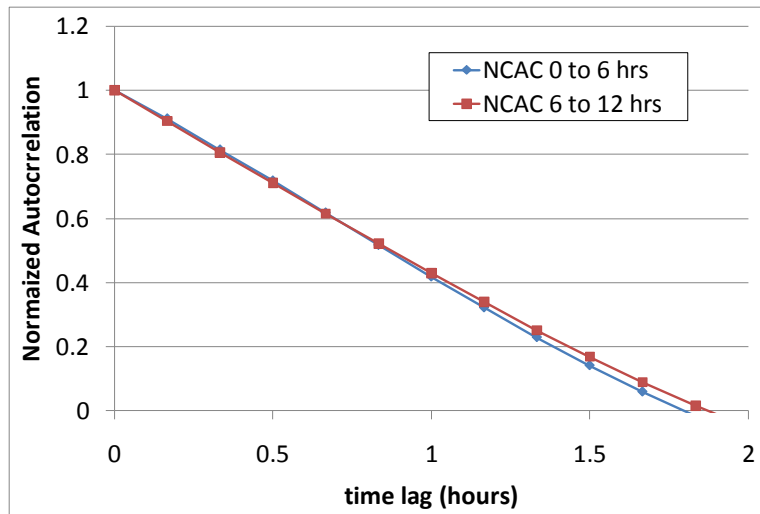


Figure 7.12 Correlation time is independent of cell cycle state. Here the first and second 6 hours of synchronized Ld2G population imaging were processed to confirm that the frequency is not changing during the cell cycle. Note that the 6 hour Normalized Composite Autocorrelation (NCAC) yields a shorter HF- $\tau_{1/2}$ than the 12 hour case.

7.3.3 Example of noise maps for 6 LTR-d2GFP Isoclonal experiments

In the earlier polyclonal noise mapping, fair genomic representation came at the cost of low sampling statistics as each cell represented one clone. Even with this low statistical sampling, significant noise map based inferences were made possible. Monoclonal LTR d2GFP experiments were used to define the constitutive gene expression model.

This section aims to demonstrate that greater cellular statistics for individual monoclonal experiments can yield various noise map signatures. The results shown here are unpublished and we are currently performing various monoclonal experiments with the ultimate goal of extracting individual O-K, 2-state parameters for each integration site. These integrations can be sequenced and further characterized to connect between functional noise map signatures and specific gene circuit structures.

Similar to polyclonal experiments performed and processed earlier, we first extract noise magnitude and correlation for the 6 monoclonal populations imaged (Fig. 7.13). The scatter is represented for each monoclonal population and the constitutive monoclonal model line is shown in green. Individual monoclonal noise maps are generated and can be further analyzed and modeled later (Fig. 7.14). Interestingly, if accounting for all the cells from all 6 monoclonal populations, the half correlation time distribution matches the distributions from the polyclonal experiments (Fig. 7.15).

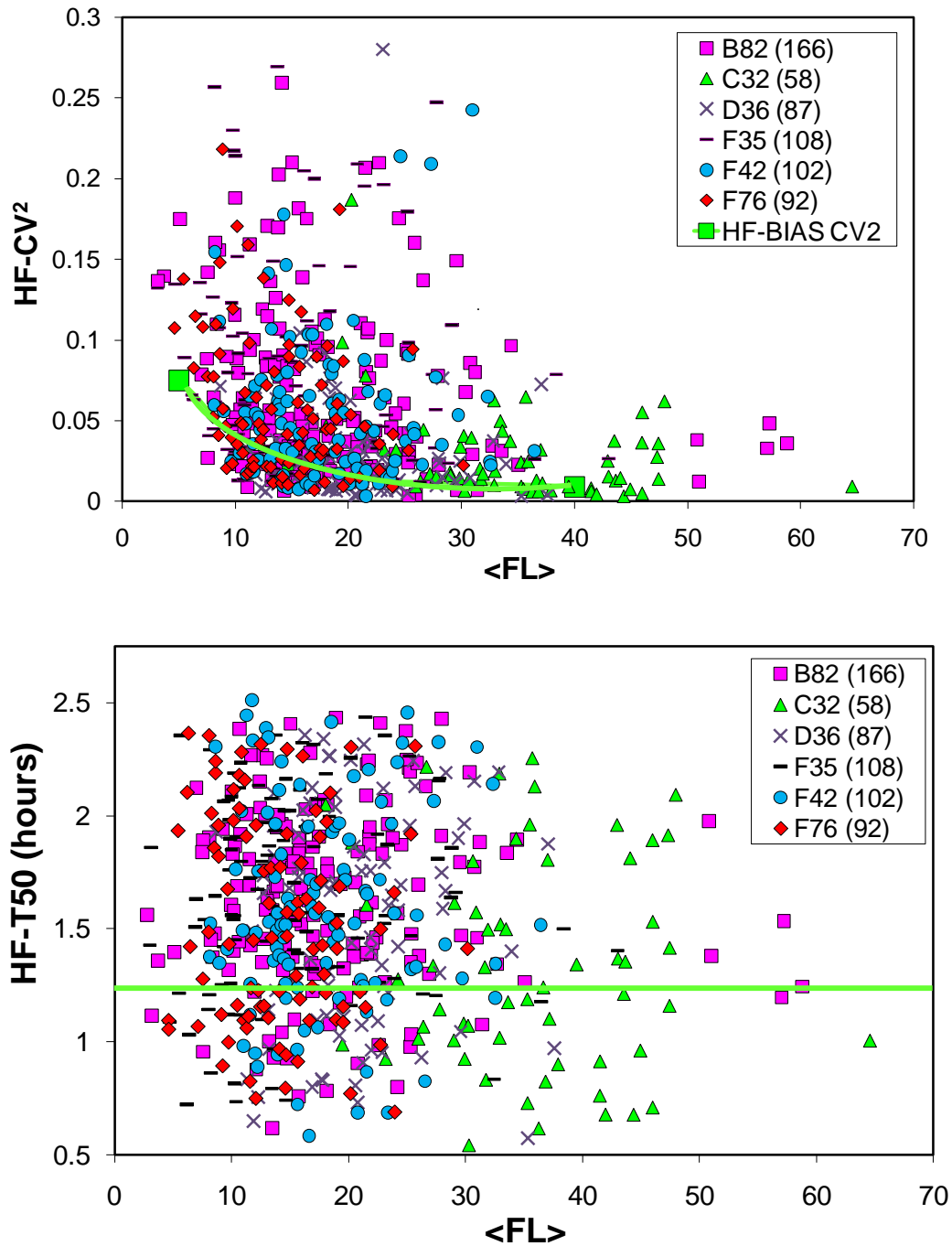


Figure 7.13 Noise magnitude and correlation for Ld2G monoclones. (**upper**) 12 hr HF-CV² versus $\langle FL \rangle$, (**lower**) 12 hr HF-T50 versus $\langle FL \rangle$. The previously described constitutive model line is shown in green.

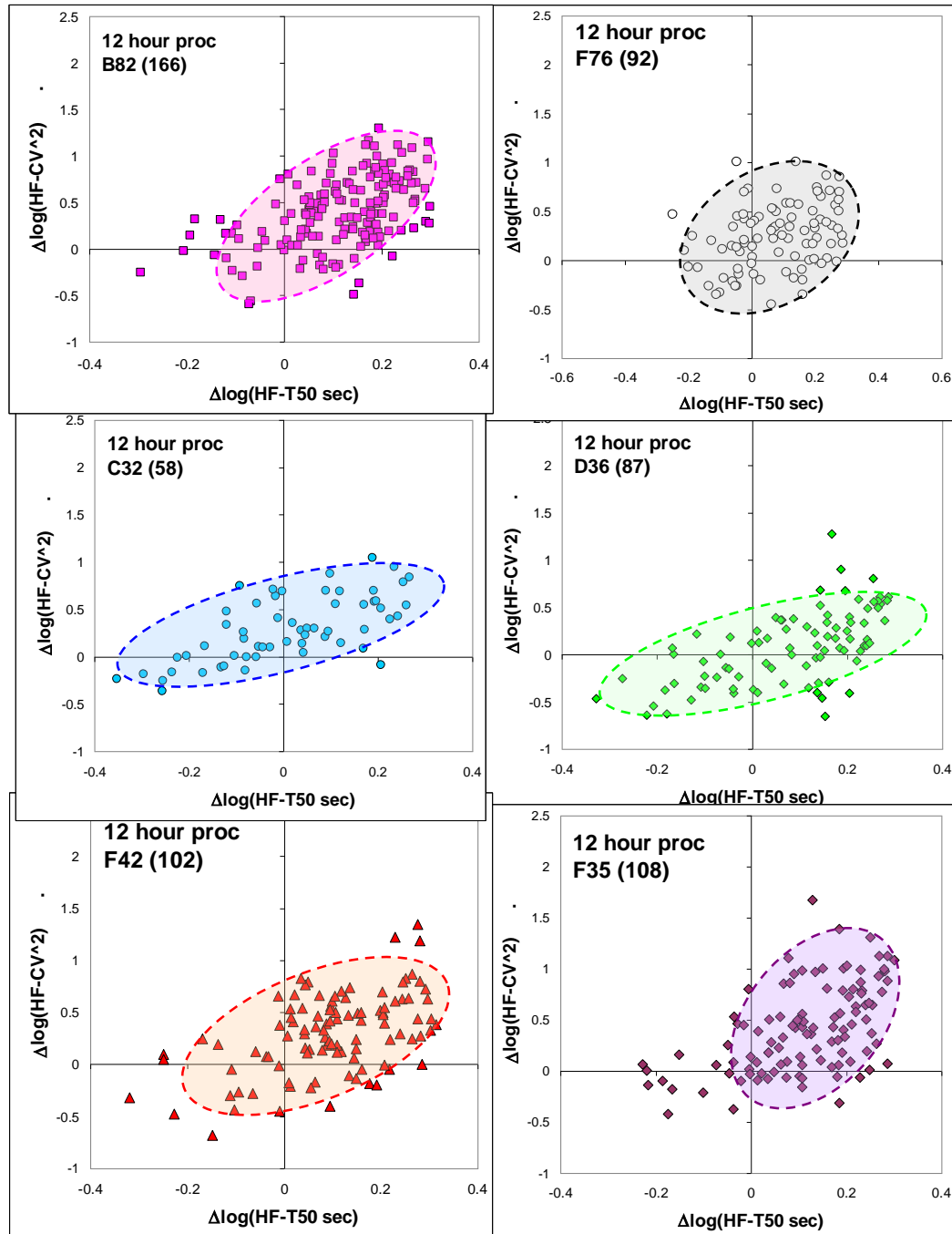


Figure 7.14 Individual noise map signatures for 6 Ld2G monoclones. Some isoclones look enriched in bursty, first quadrant noise map occupancy (e.g. B82, F42, and F35), while others are more constitutive (C32 and D36).

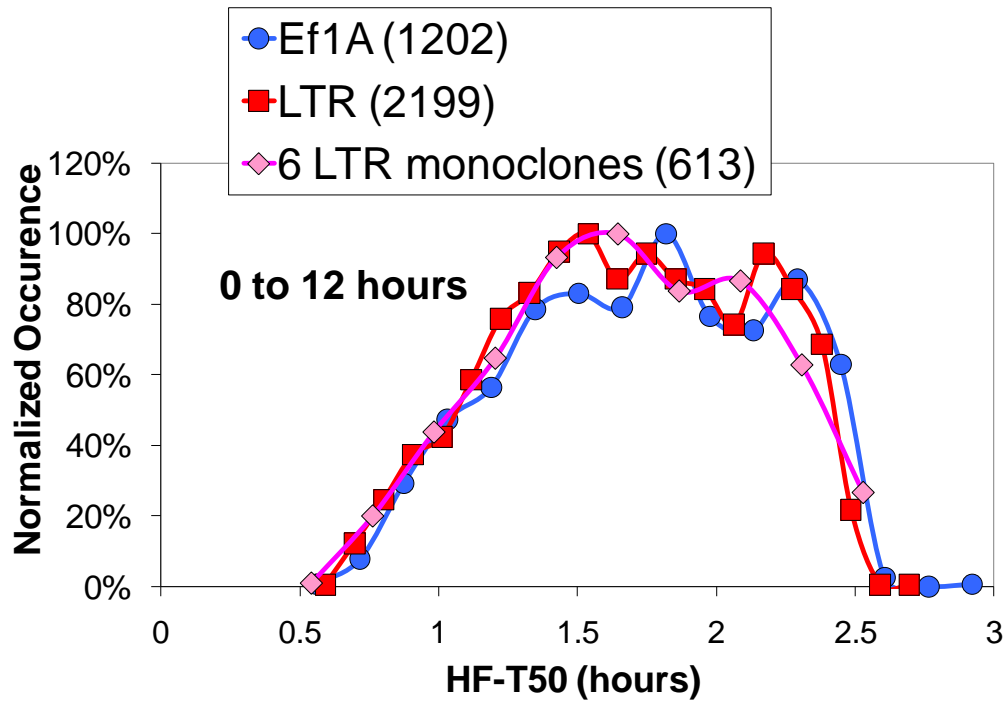


Figure 7.15 Monoclonal and polyclonal HF-T50 distributions are similar between Ef1A poly (blue circles), LTR poly (red squares) and all 6 LTR monoclonals (pink diamonds).

7.3.4 Longer cell recovery has little effect on noise map signature

To examine the influence of additional cellular recovery on gene expression, noise map coordinates were processed for the first and last 12 hours of 18 hour imaging experiments (i.e. 6 hours of additional recovery). This was done for both the Ld2G polyclonal and monoclonal experiments and usually consisted of a late plateau signature in the population general intensity trend (Fig. 7.16). The resulting noise maps demonstrate that the composite experimental coordinates with and without additional cell recovery do not deviate from one another (Fig. 7.17). This result is reminiscent of the +FB analysis in Weinberger, Dar, and Simpson (2008)[15] where expression dynamics in transient and steady-state were identical resulting in modeling of positive feedback strength in the transient response.

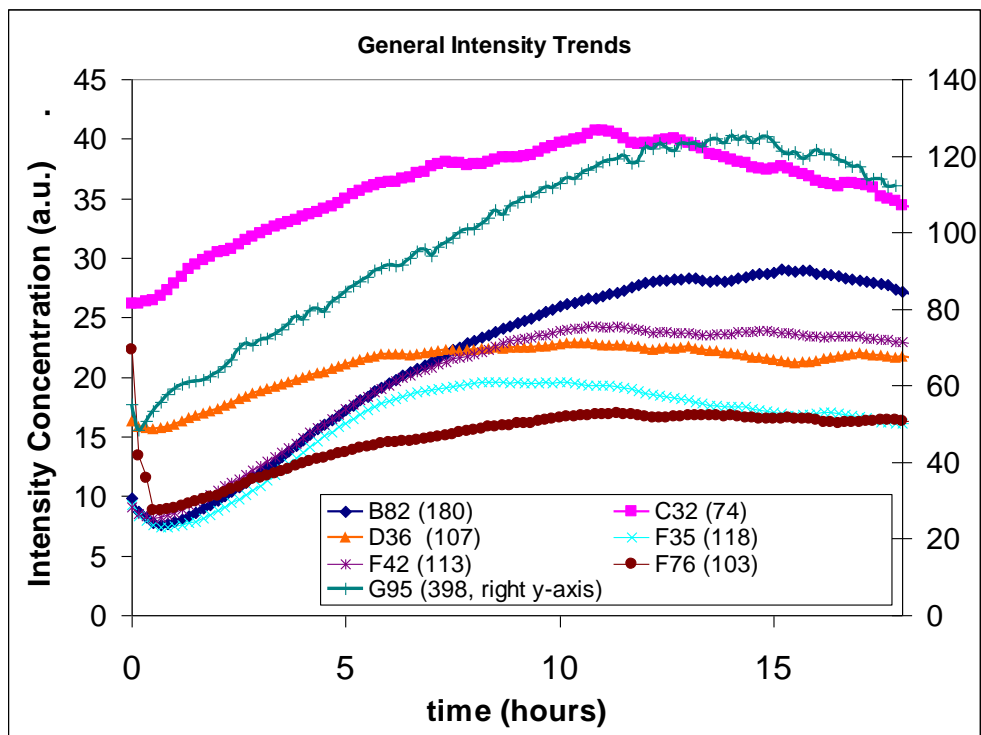


Figure 7.16 Plateau of Ld2G monoclonal general intensity trends. 7 separate isoclonal experiments sampling a total of over 1k cells show a typical rise and plateau over time.

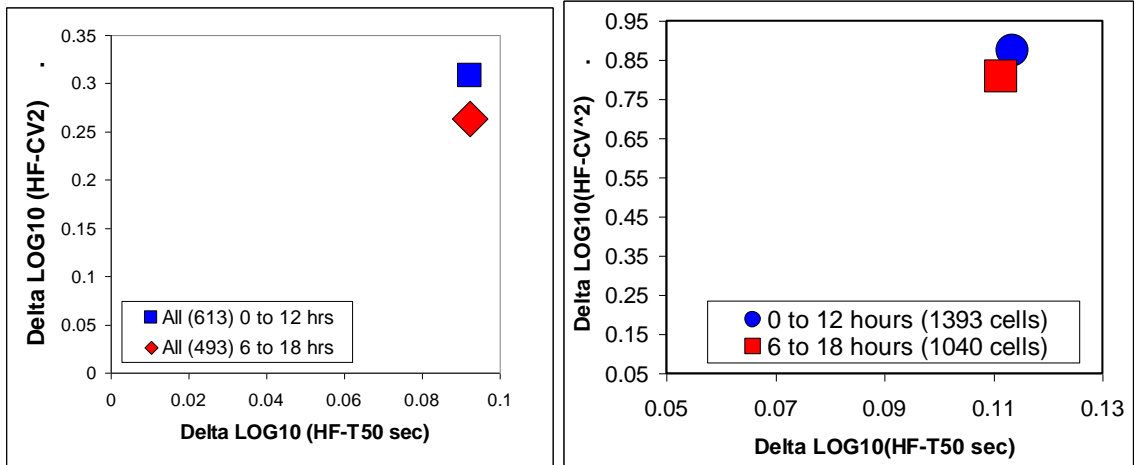


Figure 7.17 Longer sample recovery does not change the experimental results. An additional 6 hours of sample recovery doesn't change the resulting noise map coordinate. At **left** are the coordinates for the isoclinal experiments, and at **right** are the composite coordinates for over 1k cells of the Ld2G polyclonal experiment.

7.3.5 Noise map scatter dependence on experiment duration

To investigate the influence of experiment duration different than 12 hour single cell scatter in the noise map space, a wide range of experiment durations were simulated (Fig. 7.18). Shorter experiment durations down to 4 hours were simulated to see the effect on the biased autocorrelation and the HF-T50 cutoff observed in the 12 hour processing. The correlation cutoff was quantified and observed to increase linearly with experiment duration (Fig. 7.18).

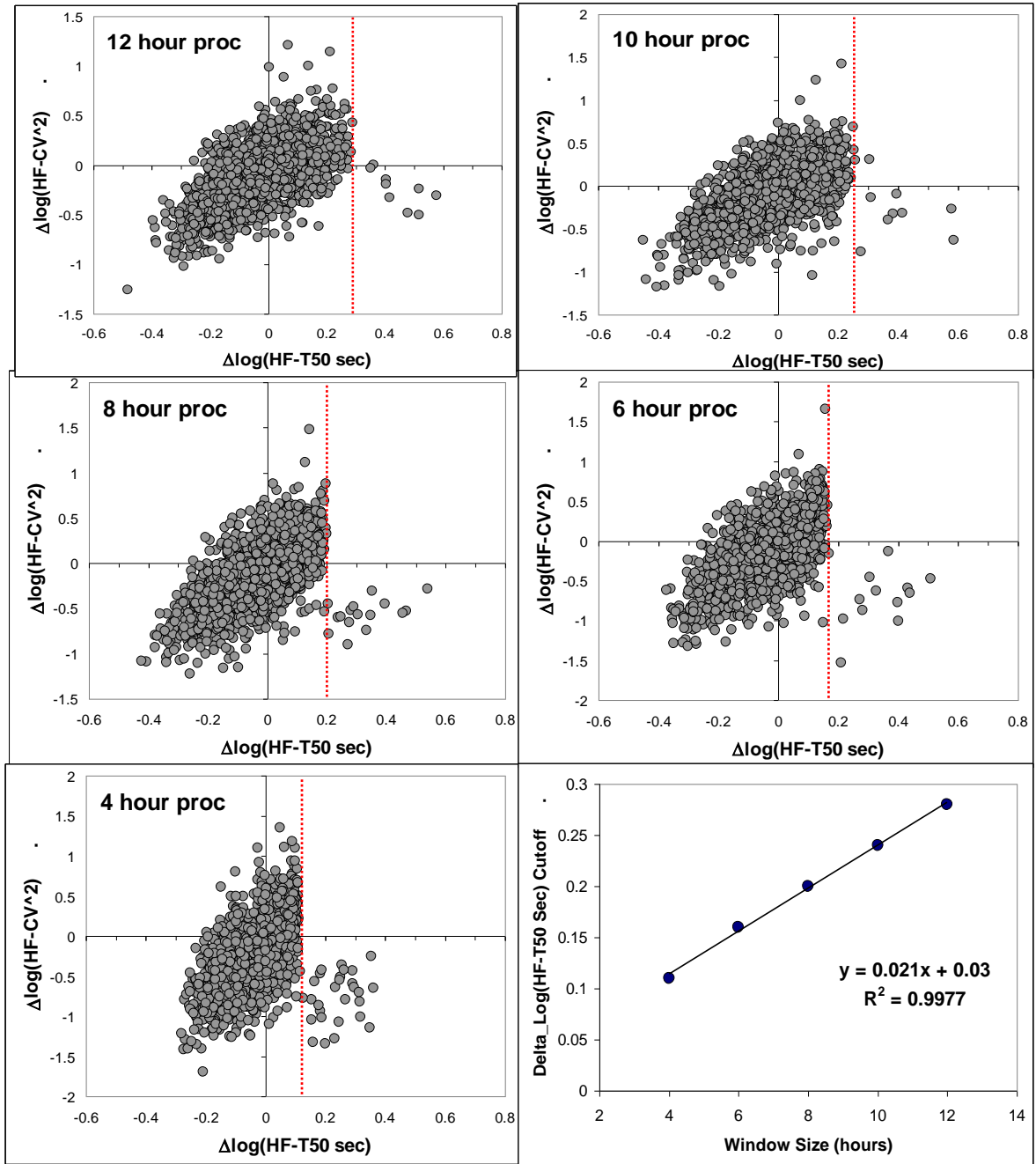


Figure 7.18 Noise map limited duration correlation cutoff. In the above, noise map scatters were simulated for 1600 independent single cell trajectories for experiment durations ranging from 4 to 12 hours. The $\Delta\log_{10}(\text{HF-T50 sec})$ axis cutoff increases linearly with experiment window size (lower right).

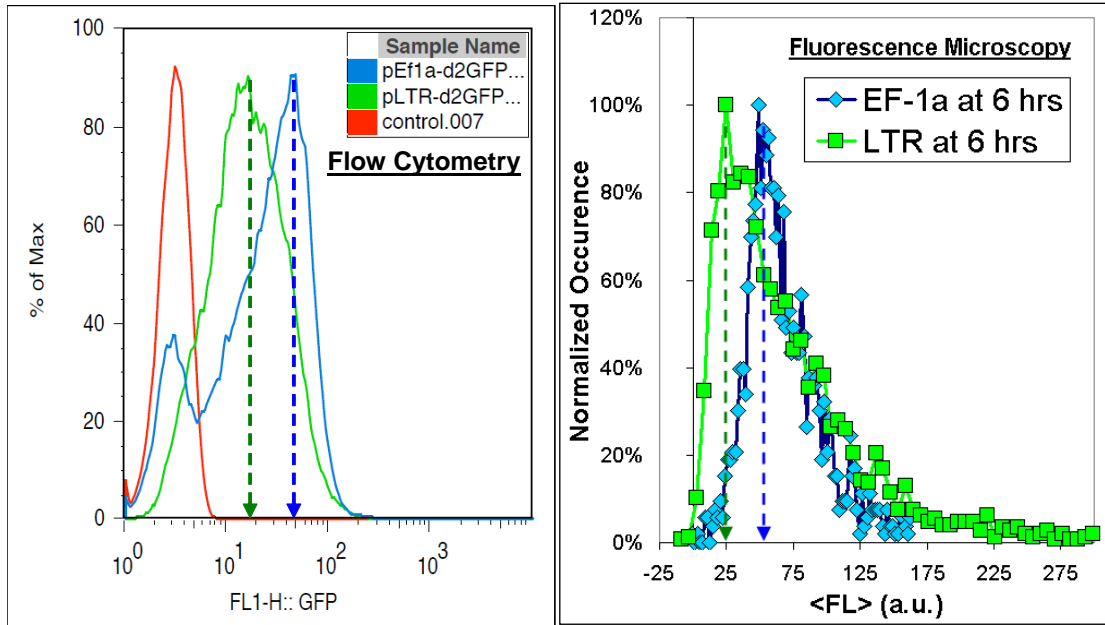


Figure 7.19 EF-1 α d2G and LTR d2G polyclonal fluorescence intensity distributions. Both flow cytometry (**left**) and fluorescence microscopy (**right**) are consistent with one another where the EF-1a d2G poly peak is $\sim 2\times$ the Ld2G poly peak. Here, as expected, CV_{LTR} is greater than CV_{EF1A} .

7.3.6 Experimental Methods Summary

Lentiviral vectors were cloned as previously described[12] and used to infect Jurkat cells at a multiplicity of infection < 0.1 . Cells were infected and then fluorescently imaged on glass-bottom dishes in RPMI 1640 with 10% fetal calf serum and 1% penicillin-streptomycin at 37°C and 5% CO₂ under humidified conditions for 12-24 hours under a 40X (1.2 NA) oil-immersion objective on an Olympus DSU™ microscope equipped with an automated linear-encoded X-Y stage, as previously described[79, 116]. Image processing and cell tracking was performed in Matlab™ using custom-written code and each experiment could generate up to 1000 trajectories for analysis. Noise mapping was performed as described in chapter 4. Stochastic simulations[151] of constitutive expression, generation of the calibration library, and fits were performed using a custom simulation program. The simulations utilized a reported and optimized accelerated version of the Gillespie algorithm[150].

7.3.7 Resampling algorithm for converting polyclonal noise maps to O-k probability landscapes

The objective is to use observations of noise in the $\Delta\tau_{1/2} - \Delta CV^2$ space to make inferences about the underlying values of O and K in the gene circuit model. Theoretically, each O-K pair maps to a precise location in the $\Delta\tau_{1/2} - \Delta CV^2$ map. However, in the present case, this mapping is complicated by at least three factors. First, the semi-random integration of the GFP reporter circuit thorough out the genome means that there is a distribution of O and K represented in the experimental noise data, rather than a single pair of values. Therefore, experimental observations of noise in the $\Delta\tau_{1/2} - \Delta CV^2$ will be diffusely distributed. Secondly, the noise data is represented by a time series of finite duration, which contributes significant uncertainty to the observation of $\Delta\tau_{1/2} - \Delta CV^2$ of a given integration site. In other words, a noise time series having a particular value of O and K maps to a single point in the $\Delta\tau_{1/2} - \Delta CV^2$ space when the time series is infinitely long; however, for practical situations involving time series of finite duration, a single OK value maps to a distribution of $\Delta\tau_{1/2} - \Delta CV^2$ values. Third, the distribution of noise from each O-K value is broad and therefore overlaps other nearby values of O-K. The combined effect of these factors can be seen in Figure 7.20, in which the $\Delta\tau_{1/2} - \Delta CV^2$ mapping for individual simulations of various values of O and K can be observed. As a result of these three factors, a single noise observation in $\Delta\tau_{1/2} - \Delta CV^2$ may potentially be assigned to a very broad spectrum of O-K values. These challenges confounded attempts to determine a unique distribution of O and K values that could describe the observed noise data.

Instead, a resampling approach was developed that attempts to answer the following question: “Of the N_{exp} experimental observations of noise in the $\Delta\tau_{1/2} - \Delta CV^2$ space, what is the maximum number n of those observations that can be described by a given O – K pair?” Formally, a null hypothesis was specified such that n experimentally observed data points could be described by a given value of O and K. The null hypothesis was to be rejected when the experimental data set had fewer data points in certain regions of the $\Delta\tau_{1/2} - \Delta CV^2$ space than would be expected based on the distribution of noise for a

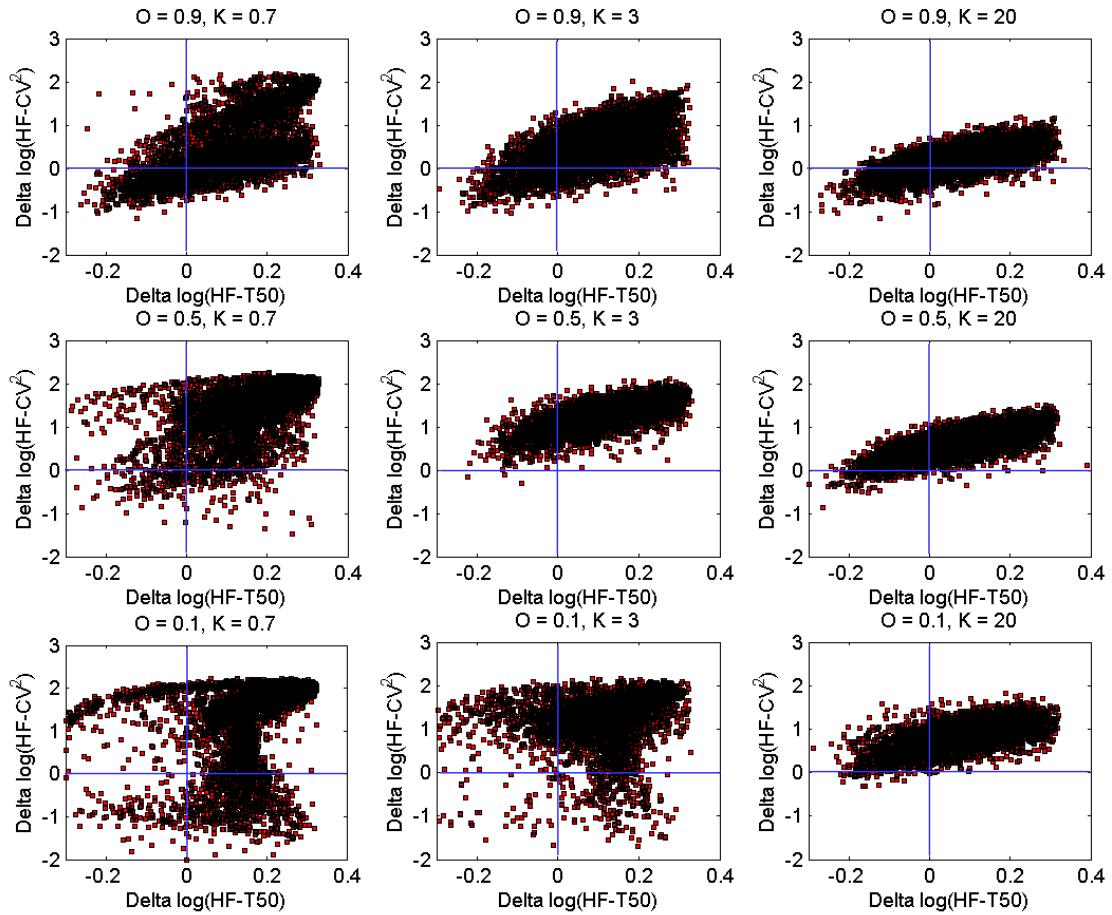


Figure 7.20 Examples of two-state noise map simulations. Simulations ranging from low (left) to high (right) burst kinetic rates, and low (lower) to high (upper) On fractions are displayed. High k and high O displays a constitutive signature. The noise map spread is attributed to single cell 12 hour snapshots of the process. As k and O are lowered the signature moves into the 1st noise map quadrant.

particular value of O and K given by the simulation library. Next the minimum value of n at which the null hypothesis could be rejected at the 95% confidence level was found. This value of n represents the maximum number of experimental observations that can be assigned to a given value of O and K. This procedure was repeated over a wide range of O and K values such that a distribution was obtained.

The detailed resampling algorithm is described as follows. Each O-K coordinate considered was represented by 5000 simulations in the simulation library. While each simulation maps to a single point in the $\Delta\tau_{1/2} - \Delta CV^2$, collectively the 5000 simulations define the distribution of noise for a given value of O and K. The analysis was confined to physiologically relevant values of K ranging from 0.5 to $30\gamma_p$ and values of O ranging from 0.1 to 0.98. For each value of O and K we established a 15 by 15 rectangular grid in the $\Delta\tau_{1/2} - \Delta CV^2$ space that encompassed the range of values observed in the 5000 simulations. In order to increase the resolution of the most characteristic noise features of a given O-K pair, a light filtering was applied to the simulation library to remove outlier simulations. The filtering was based on the Mahalanobis distance from the center of mass of the simulation cloud. The cutoff Mahalanobis distance was initially set at 10 and increased as necessary such that no more than 0.5% of the simulations for a given O-K pair were removed. Removed simulations were replaced with simulations randomly sampled from the remaining simulations in the library to return the total number of simulations to 5000.

Increasing values of n ($\Delta n \approx 0.05N_{exp}$) were incrementally tested to determine the lowest value that caused the hypothesis to be rejected. For each pixel i in the 15x15 grid the number of simulations n_i that fell into that pixel was determined such that:

$$\sum_1^{225} n_i = n$$

The cumulative normal distribution of each value of n_i was then determined by resampling n simulations from the library 1000 times, keeping O and K constant. After

determining the statistical distribution of n_i each of the 1000 simulations was reanalyzed by quantifying its lower-least-probable-distribution statistic DS_{lp} which we define as:

$$DS_{lp} = \left[\prod_{i=1}^{225} f(n_i) \right]^{m_{0.5}}$$

where

$$\begin{aligned} f_i(n_i) &= cdf_i(n_i) & \text{for } cdf_i(n_i) \leq 0.5 \\ f_i(n_i) &= 1 & \text{for } cdf_i(n_i) > 0.5 \end{aligned}$$

$cdf_i()$ is the cumulative distribution function for pixel i , and $m_{0.5}$ is the number of pixels for which $cdf_i(n_i) \leq 0.5$. The distribution of DS_{lp} was then characterized based on the 1000 samples as $N(\mu_{DS}, \sigma_{DS})$ where N is the normal distribution with mean μ and standard deviation σ . Finally the null hypothesis was rejected for

$$cdf\left(\frac{DS_{lp,exp} - \mu_P}{\sigma_P}\right) < 0.05 .$$

Where $DS_{lp,exp}$ is the lower-least-probable-distribution statistic evaluated for the experimental data set. $DS_{lp,exp}$ is calculated using the $cdf(n_i)$ that were developed by resampling the simulation library.

In the above procedure, hypothesis testing was conducted over the entire grid at once instead of pixel-by-pixel to prevent over-representation of type I errors. The lower-least-probable distribution statistic has the desired property in that it prevents over-representation of type II errors that arise due to $N_{exp} > n$ as illustrated in Figure 7.21. Therefore, the developed procedure has the desired feature of controlling for both type I and type II errors.

The above procedure was applied for each value of O-K in the simulation library. The result is that for N_{exp} observations of noise in the $\Delta\tau_{1/2} - \Delta CV^2$ the maximum number (n) of experimental data points that could have originated from our model with particular parameters O and K can be inferred. Finally, the distribution of probable O-K values are represented using a heat map.

Several tests of the resampling algorithm are performed by generating artificial data sets from the simulation library. In each test, the artificial data set was generated by randomly sampling 1000 points in the $\Delta\tau_{1/2} - \Delta CV^2$ noise space for known values of O and K or a known mixture of O and K. The $\Delta\tau_{1/2} - \Delta CV^2$ points were then treated as experimental data and the ability to resolve the artificial data back to the original O and K values was determined. First each of the nine values of O and K used in Figure 7.21 were tested. The results are shown in Figure 7.21. In each case, the resampling algorithm correctly identifies the original O and K values by indicating that all 1000 data points could have potentially originated from that particular O-K pair. In all cases, the resampling algorithm indicates that some of the data points could have potentially come from values of O and K that were not used to generate the artificial data set. In fact for some values of O and K, the resampling algorithm concludes that all of the artificial data points could have come from an incorrect value of O and K. This demonstrates the limitations of the ability of finite duration time series for uniquely resolving the underlying mechanisms and parameter values giving rise to the observed noise. However, this method is capable of identifying the general range of O and K values with the greatest potential to generate the observed noise profiles.

Next, the ability of the resampling algorithm to resolve mixtures of O and K values was determined. A mixture of 1000 $\Delta\tau_{1/2} - \Delta CV^2$ values were sampled equally from two well-separated O-K pairs: O = 0.1, K = $1\gamma_P$ and O=0.92, K= $6\gamma_P$. The results (Figure 7.22, left) shows two well-resolved peaks centered at or near the original O and K values. The analysis indicated that up to 700 of the 1000 points could have originated

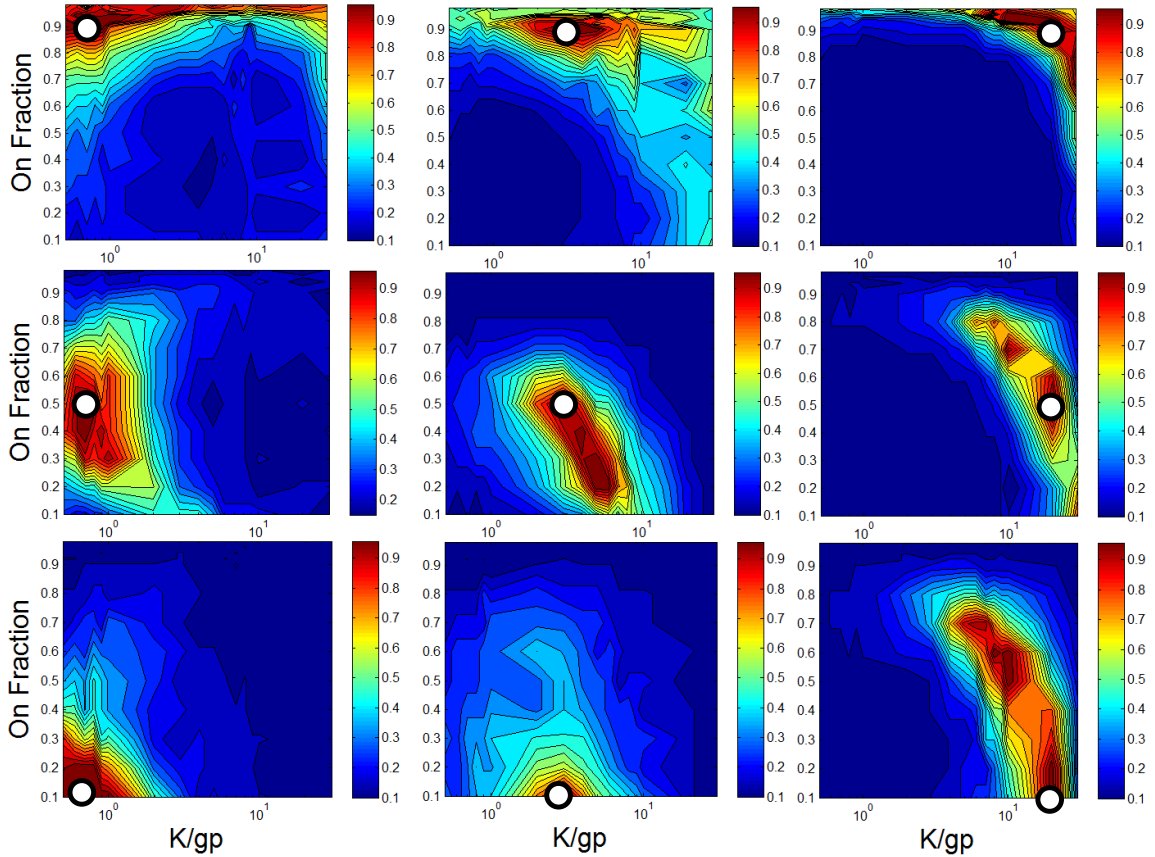


Figure 7.21 Probability point spread functions of known O-k simulations. Resampling algorithm was applied to the O-k simulations from the previous figure. True simulated O-k coordinate used is labeled in each probability landscape with a white dot.

from the peak O-K coordinates, compared to 500 points sampled from each of the two locations. Next, a sampling from the same two O-K pairs was used, except that the mixture included 900 simulations from one point ($O=0.92$, $K=6\gamma_P$) and only 100 simulations from the other ($O=0.1$, $K=1\gamma_P$). The resampling algorithm was still able to resolve both peaks (Figure 7.22, middle) since the small bump near $O=0.1$, $K=1\gamma_P$ is not observed when points are only sampled from $O=0.92$, $K=6\gamma_P$. Next, the resolution in how well the two peaks could be resolved as the O-K values were moved closer together was determined. For this test a mixture of 500 samples each from $O=0.5$, $K=5\gamma_P$ and $O=0.8$, $K=5\gamma_P$ were used. In this case, it was not possible to resolve the two peaks (Figure 7.22, right). However, the fact that a mixture was likely to be present was indicated by the fact that none of the discrete values of O and K in the grid were able to account for all of the simulations (a maximum of 950 simulations could have originated from $O=0.7$ and $K=4\gamma_P$). In conclusion, the resampling algorithm is capable of identifying the maximum number of simulations that could come from specific values of O and K. The resulting heat map is indicative of the possible distribution of O and K values that underlie the observed noise data.

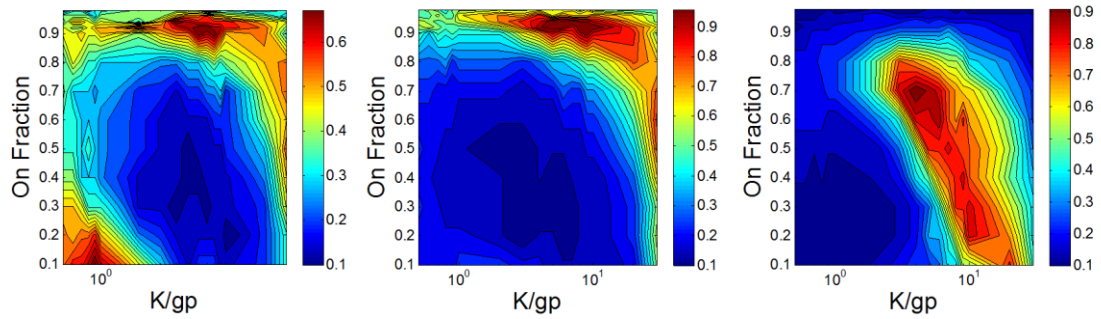


Figure 7.22 Resolution in the burst probability landscape. (left) resolution in the probability landscape of two unique O-k simulations contributing equally to the noise map (center) A 90-10 split between the O-k pair in the left plot. The lower 10% peak is seen as a slight increase in light blue (right) two O-k from similar regions do not separate out and are hard to resolve a double peak.

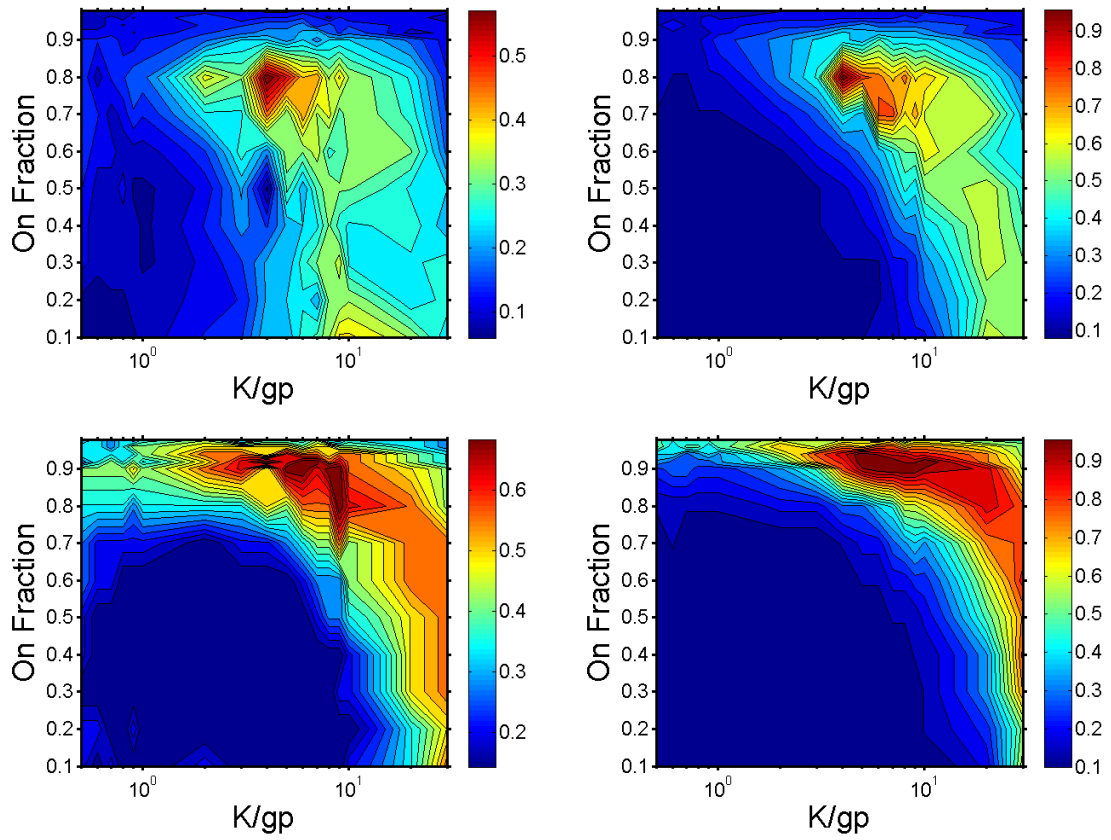


Figure 7.23 Best simulation match to experiments. (left) Experiments LTR with nothing (upper-left) and LTR + TNF (lower left) and their corresponding best single O-k simulation match to their right.

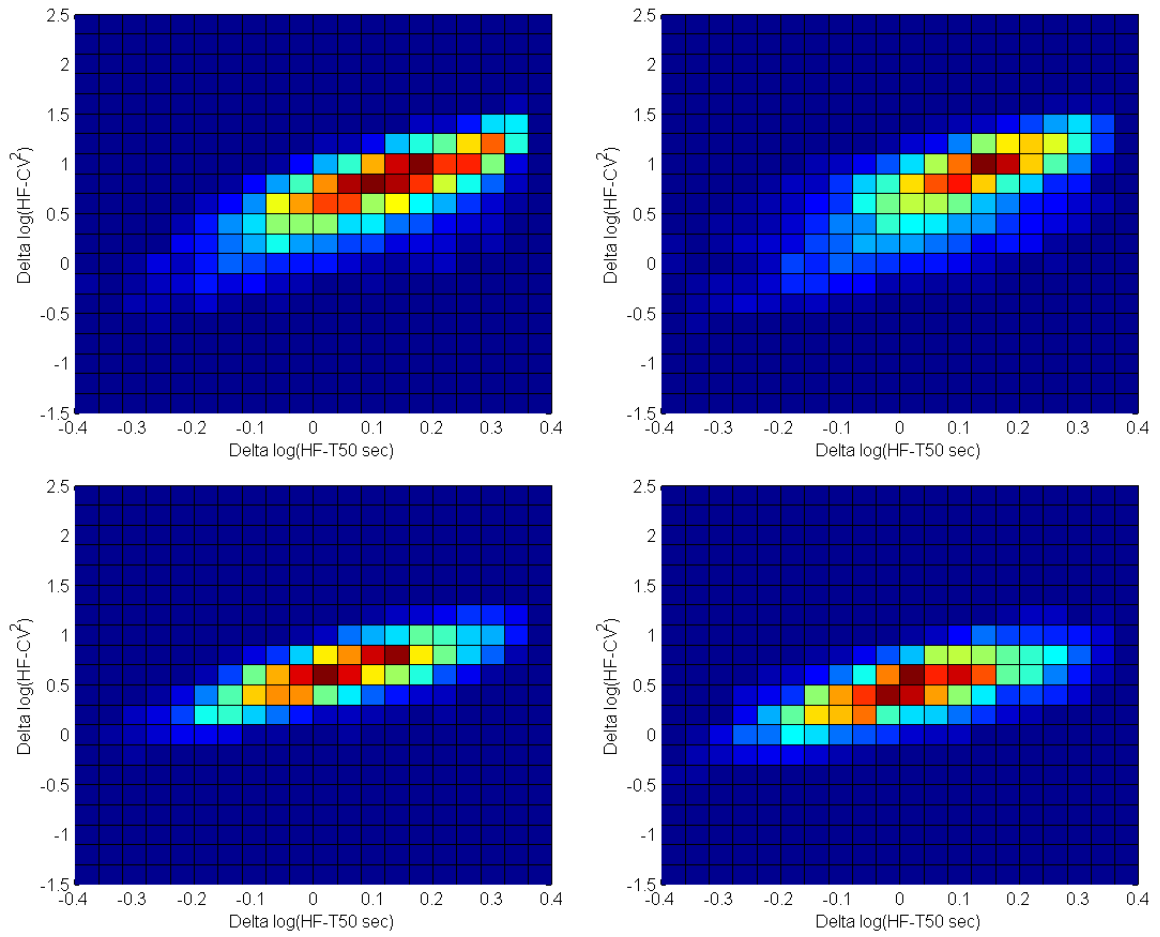


Figure 7.24 Samples of simulated NPD maps. The O-k parameters used are $O = 0.7$ and $k = 0.7 \text{ Hr}^{-1}$ (upper left), $O = 0.2$ and $k = 0.7 \text{ Hr}^{-1}$ (upper right), $O = 0.5$ and $k = 2.1 \text{ Hr}^{-1}$ (lower left), and finally $O = 0.2$ and $k = 3.5 \text{ Hr}^{-1}$ (lower right).

Table 7.3 Parameters for the 2-state model simulation library

<u>Simulation Parameter</u>	<u>Biological Interpretation</u>	<u>Value</u>	<u>Rationale</u>
α_0	Basal expression rate (0%, 30% and 50% basal expression)	0, $0.3*\alpha_1$, $0.5*\alpha_1 \text{ sec}^{-1}$	Chosen to create a basis set of NPD maps with the best match to measured noise behavior (see Basal expression section below).
α_1	Additional expression rate during a burst	0.0132 sec^{-1}	Chosen to populate the upper right quadrant of the noise map in agreement with measured data
k_p	Translation rate	$0.07839165 \text{ sec}^{-1}$	Chosen in combination with transcription rate to match the measured CV^2 bias vector component
g_m	mRNA decay rate	$0.0007839 \text{ sec}^{-1}$	Chosen to match the measured HF-T50 bias vector component

g_p	GFP decay rate	0.00011 sec^{-1}	Literature value[117 , 152] Applies to both immature and mature GFP.
k_{mat}	GFP maturation rate	$0.000385 \text{ sec}^{-1}$	Chosen by matching to experimental constitutive monoclonal noise maps
k	Burst kinetics parameter = $k_{\text{on}}+k_{\text{off}}$	$0.00035 - 173.3 \text{ Hr}^{-1}$ (various sampling in range)	Swept across large range to create the set of basis NPDs
O	On fraction (fraction of time spent in the high expression state)	0.1, 0.2, ..., 0.9, 0.92, 0.94, ..., 0.98	Swept across large range to create the set of basis NPDs

7.3.8 Determination of the basal transcription level

The following plots show the calculated composite autocorrelation noise map coordinates for all simulations in each of the 0% (left), 30% (right), and 50% (lower) basal expression libraries. Also included are the mean noise map values for LTR d2G poly + nothing and Ef1A d2G poly + nothing. The 0% basal expression library best covers the measured experimental noise map coordinates and based on the composite map predicts an $O=0.8$ and $O=0.95$ for each respectively.

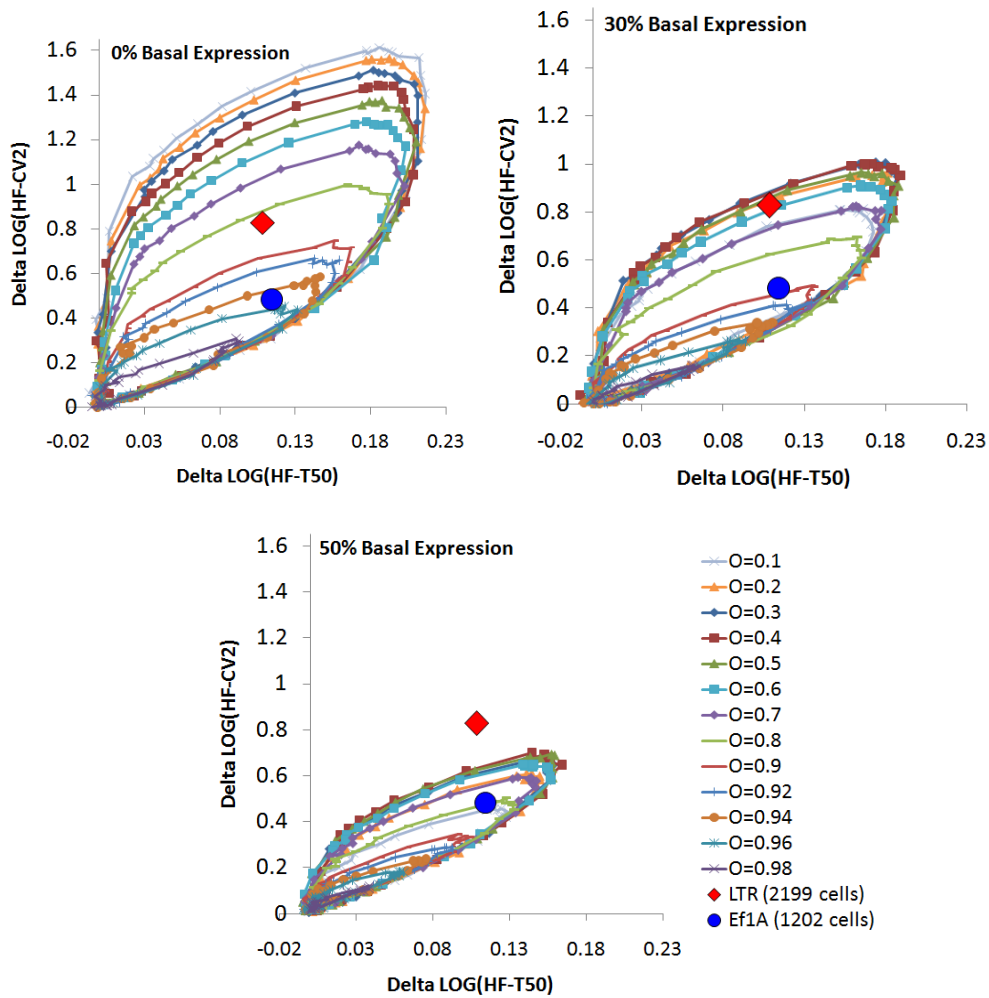


Figure 7.25 Composite noise map for an array of O and k values. The lines and small points show the composite noise map locations for various simulated values of O and k in each of the three basal (0%, 30%, and 50% of burst rate). The large data points show the composite noise map location for the LTR and EF-1 α experiments.

7.3.9 Multiple General Trends for Polyclonal or 2-Reporter Experiments

An additional modification to the noise processing algorithm needs to be considered in experimental cases that have more than one underlying gene circuit architecture driving fluorescent reporter expression. The immediate and simplest example is a two reporter system that has non-overlapping emission spectra. In this case two separate general population intensity trends, $A_1(t)$ and $A_2(t)$, need to be calculated and then used to process noise for each single cell fluorescence channel relating to those reporters separately. $A_1(t)$ and $A_2(t)$ may be related or influence each other's dynamics, but separating the noise processing is still required.

A more challenging situation arises with the lentiviral integrated polyclonal T-cell experiments described in detail in chapter 3. In this case each cell has a GFP reporting vector integrated in different and unique genomic loci. This means that each single cell has its own unique underlying genetic architecture and would optimally have its own general population trend ($A_i(t)$ for all $i=1,\dots,N$ cells in the experiment). To properly generate the general trends and characterize each integration site, cells would need to be isolated and grown out into thousands of individual isoclonal populations which would completely defeat the high throughput (low statistics) polyclonal method's objective. So since a single cell trajectory cannot represent an individual general trend, and it isn't feasible to grow out each individual isoclone, we seek to compromise between the two extremes (Fig. 7.26). We group cell trajectories according to their 12th hour intensity value; 20% highest, 40% medium, and 40% lowest, and generate a general population trend based on each of these subgroups and then process each polyclonal experiment, 3 times, according to each of these 3 subgroups. The 3 general trend regions are a compromise to cover the wide range of deterministic behaviors present in the underlying polyclonal integration site landscape.

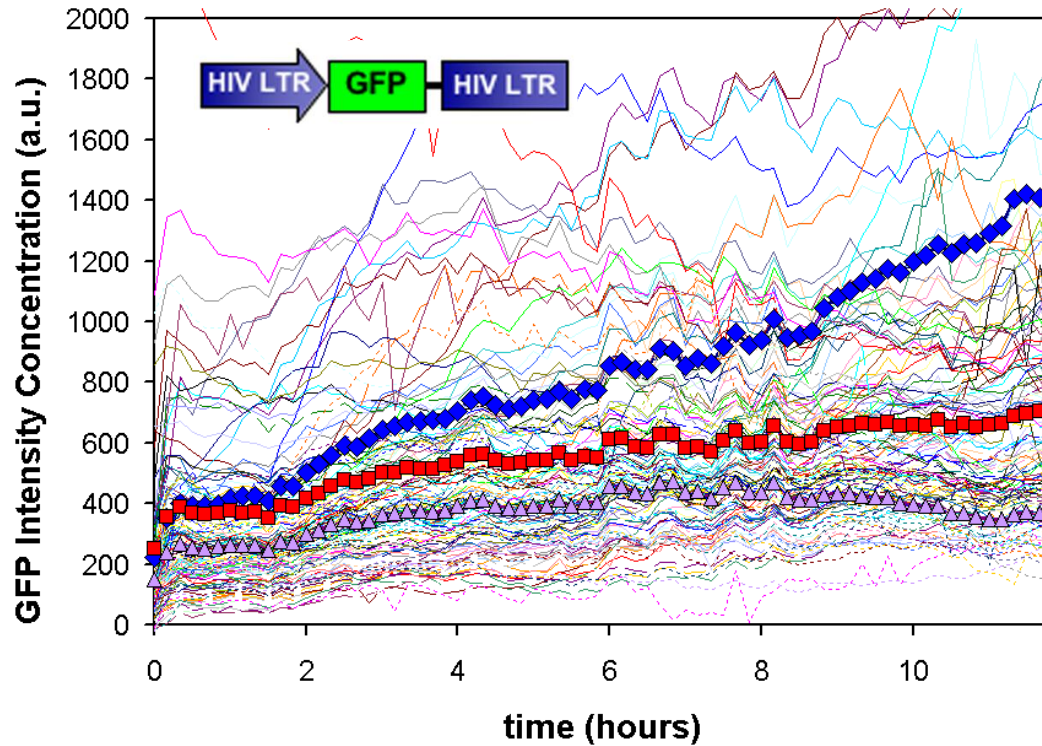


Figure 7.26 Multiple deterministic trends for analyzing polyclonal experiments. As described above and previously in Austin et al, *Nature* (2006)[70] and Weinberger et al, *Nat. Genet.* (2008)[15] a main step in extracting stochastic fluctuations in gene expression is determining the general intensity or deterministic trend of the underlying gene circuit. In these previous cases all the cells collected in a population had identical gene circuits resulting in calculating a time dependent average deterministic trend over the whole population. On the other hand, in the current study every cell in the polyclonal population has a different gene circuit structure and unique chromosomal integration site. To process the polyclonal experiments three representative deterministic trends were calculated using the 20% highest intensity trajectories (based on the 12th hour time point) shown in **blue** above, the 40% medium intensities (**red**), and finally the 40% lowest intensities (**purple**). This representative set of deterministic trends yields an approximation of general polyclonal trends and was used consistently with all performed polyclonal experiments.

7.3.10 Are noise map shifts due to extrinsic noise?

In addition to transcriptional bursting, extrinsic noise[27] could be responsible for the measured noise map shifts to the upper right quadrant. However, a principle advantage of HF processing is that it focuses on high frequency intrinsic noise which is directly modulated by gene circuit structure and function while de-emphasizing lower frequency extrinsic noise. To examine extrinsic noise-mediated shifts in HF noise maps, constitutive gene expression was simulated using various levels of extrinsic noise using the noise simulation model described in the Supplementary Information of Austin *et al*, (2006)[70]. Figure 7.27 below shows the unfiltered and HF shifts in the average noise map locations for extrinsic noise levels of 9%, 39%, and 56% of total noise. Although the unfiltered noise map locations show considerable movement away from the origin with the addition of extrinsic noise, the HF-processed points remain contained in a region near the origin. As a result, the HF-NPD map shifts shown in chapter 4 cannot be accounted for by assuming large amounts of extrinsic noise.

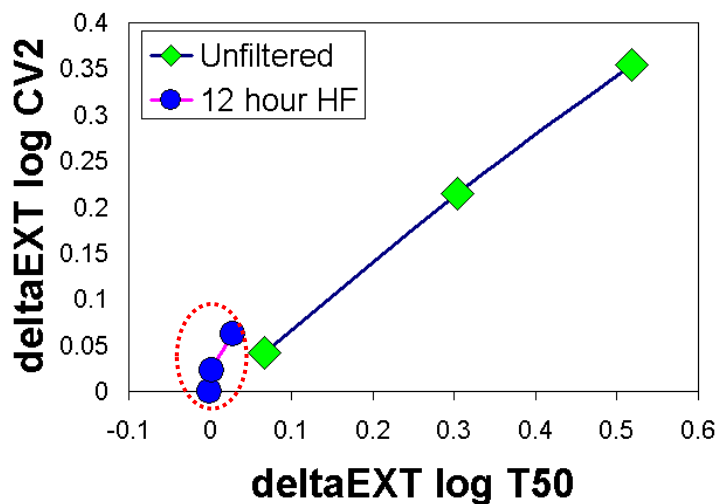


Figure 7.27 Simulated HF extrinsic noise cannot explain reported HF-NPD map shifts.

7.4 The Coupling of Stochastic and Plastic Response

7.4.1 Derivation of the relationship between excess noise and PL

Using the assumption that $b \gg 1$ and ACF functions derived in [47] and [10]:

$$\Phi(\tau) \approx \frac{\alpha O b}{\gamma_d} b e^{(-\gamma_p \tau)} + \left(\frac{\alpha O b}{\gamma_d} \right)^2 \frac{(1-O)}{O k} \left(\frac{\gamma_d}{\left[1 - \left(\frac{\gamma_d}{k} \right)^2 \right]} e^{(-\gamma_p \tau)} + \frac{k}{\left[1 - \left(\frac{k}{\gamma_d} \right)^2 \right]} e^{-k \tau} \right)$$

$$\Phi(0) \approx \frac{\alpha O b}{\gamma_d} b + \left(\frac{\alpha O b}{\gamma_d} \right)^2 \frac{(1-O)}{O k} \left(\frac{\gamma_d}{\left[1 - \left(\frac{\gamma_d}{k} \right)^2 \right]} + \frac{k}{\left[1 - \left(\frac{k}{\gamma_d} \right)^2 \right]} \right)$$

$$C_k = \left(\frac{\gamma_d/k}{\left[1 - \left(\frac{\gamma_d}{k} \right)^2 \right]} + \frac{1}{\left[1 - \left(\frac{k}{\gamma_d} \right)^2 \right]} \right) = \left(\frac{R}{[1-R^2]} + \frac{R^2}{[R^2-1]} \right) = \left(\frac{R^2 - R}{[R^2 - 1]} \right)$$

$$R = \gamma_d/k$$

$$CV = \sqrt{\frac{\Phi(0)}{\langle p \rangle^2}} \approx \sqrt{\frac{1}{\langle p \rangle} b + C_k \frac{(1-O)}{O}}$$

$$\begin{aligned}
\Phi(0) &\approx \frac{\alpha O b^2}{\gamma_d} + \left(\frac{\alpha O b}{\gamma_d} \right)^2 C_k \frac{(1-O)}{O} = n_{shot} + n_{burst} = n_{shot}(1 + N_{B/S}) \\
N_{B/S} &= \frac{n_{burst}}{n_{shot}} = \frac{\left(\frac{\alpha O b}{\gamma_d} \right)^2 C_k \frac{(1-O)}{O}}{\frac{\alpha O b^2}{\gamma_d}} = \frac{\alpha(1-O)}{\gamma_d} C_k = \frac{\alpha(1-O)}{\gamma_d} \left(\frac{(\gamma_d/k)^2 - (\gamma_d/k)}{(\gamma_d/k)^2 - 1} \right) = \\
&\frac{\alpha(1-O)}{k} \left(\frac{(\gamma_d/k) - 1}{(\gamma_d/k)^2 - 1} \right) = \frac{\alpha(1-O)}{k} \left(\frac{(\gamma_d/k) - 1}{((\gamma_d/k) - 1)((\gamma_d/k) + 1)} \right) = \frac{\alpha(1-O)}{k} \left(\frac{1}{((\gamma_d/k) + 1)} \right) = \\
&\alpha(1-O) \left(\frac{1}{k((\gamma_d/k) + 1)} \right) = \frac{\alpha(1-O)}{(\gamma_d + k)} \\
CV &= \sqrt{\frac{n_{shot}(1 + N_{B/S})}{\langle p \rangle^2}} = \sqrt{\frac{b(1 + N_{B/S})}{\langle p \rangle}} = \sqrt{\frac{b \left(1 + \frac{\alpha(1-O)}{(\gamma_d + k)} \right)}{\langle p \rangle}} \\
DM &= CV - CV_{shot} = \sqrt{\frac{n_{shot}(1 + N_{B/S})}{\langle p \rangle^2}} - \sqrt{\frac{n_{shot}}{\langle p \rangle^2}} = \sqrt{(1 + N_{B/S})} \sqrt{\frac{n_{shot}}{\langle p \rangle^2}} - \sqrt{\frac{n_{shot}}{\langle p \rangle^2}} = \\
&(\sqrt{(1 + N_{B/S})} - 1) \sqrt{\frac{n_{shot}}{\langle p \rangle^2}} = \left(\sqrt{\left(1 + \frac{\alpha(1-O)}{(\gamma_d + k)} \right)} - 1 \right) \sqrt{\frac{n_{shot}}{\langle p \rangle^2}} = \left(\sqrt{\left(1 + \frac{\alpha(1-O)}{(\gamma_d + k)} \right)} - 1 \right) \sqrt{\frac{b}{\langle p \rangle}}
\end{aligned}$$

Define plasticity (Pl) as:

$$\begin{aligned}
Pl &= \frac{\langle P \rangle_{\max}}{\langle P \rangle_{\min}} \\
&= \frac{\alpha O_{\max} b}{\gamma_d} \\
Pl &= \frac{\gamma_d}{\alpha O_{\min} b} = \frac{O_{\max}}{O_{\min}} \\
&= \frac{\gamma_d}{\alpha O_{\min} b}
\end{aligned}$$

For maximum plasticity, $O_{\max} \rightarrow 1$; then

$$Pl \approx \frac{1}{O_{\min}}$$

And

$$DM = \left(\sqrt{\left(1 + \frac{\alpha(1-O_{\min})}{(\gamma_d+k)} \right)} - 1 \right) \sqrt{\frac{b}{\langle p \rangle}} = \left(\sqrt{\left(1 + \frac{\alpha(Pl-1)}{Pl(\gamma_d+k)} \right)} - 1 \right) \sqrt{\frac{b}{\langle p \rangle}} = \left(\sqrt{\left(1 + \frac{\alpha(Pl-1)}{Pl(\gamma_d+k)} \right)} - 1 \right) \sqrt{\frac{b\gamma_d}{\alpha O_{\min} b}} =$$

$$\left(\sqrt{\left(1 + \frac{\alpha(Pl-1)}{Pl(\gamma_d+k)} \right)} - 1 \right) \sqrt{\frac{\gamma_d}{\alpha} Pl} = \left(\sqrt{\left(1 + \frac{\alpha(Pl-1)}{Pl(\gamma_d+k)} \right) \frac{\gamma_d}{\alpha}} - \sqrt{\frac{\gamma_d}{\alpha}} \right) \sqrt{Pl} = \left(\sqrt{\left(1 + \frac{(Pl-1)}{Pl(\gamma_d+k)} \right) \frac{\gamma_d}{\alpha}} - \sqrt{\frac{\gamma_d}{\alpha}} \right) \sqrt{Pl}$$

The following additional assumptions are applied:

1. $\sqrt{\left(1 + \frac{\alpha(1-O_{\min})}{(\gamma_d+k)} \right)} \gg 1$
2. $\frac{\alpha(1-O_{\min})}{(\gamma_d+k)} \gg 1$
3. $\gamma_d \gg k$

Then:

$$DM = \left(\sqrt{\left(1 + \frac{\alpha(1-O_{\min})}{(\gamma_d+k)} \right)} - 1 \right) \sqrt{\frac{n_{shot}}{\langle p \rangle^2}} \approx \left(\sqrt{\left(1 + \frac{\alpha(1-O_{\min})}{(\gamma_d+k)} \right) \frac{b}{\langle p \rangle}} \right) = \left(\sqrt{\left(1 + \frac{\alpha(1-O_{\min})}{(\gamma_d+k)} \right) \frac{b}{\alpha O}} \right) =$$

$$\sqrt{\left(1 + \frac{\alpha(1-O_{\min})}{(\gamma_d+k)} \right) \frac{\gamma_d}{\alpha O}} \approx \sqrt{\left(\frac{\alpha(1-O_{\min})}{(\gamma_d+k)} \right) \frac{\gamma_d}{\alpha O}} = \sqrt{\left(\frac{(1-O_{\min})}{(\gamma_d+k)} \right) \frac{\gamma_d}{O}} \approx \sqrt{\left(1 - \frac{1}{Pl} \right) Pl} = \sqrt{(Pl-1)}$$

$$\boxed{DM \approx \sqrt{(Pl-1)}}$$

7.4.2 Additional Representation of Noise-Plasticity Coupling in Yeast

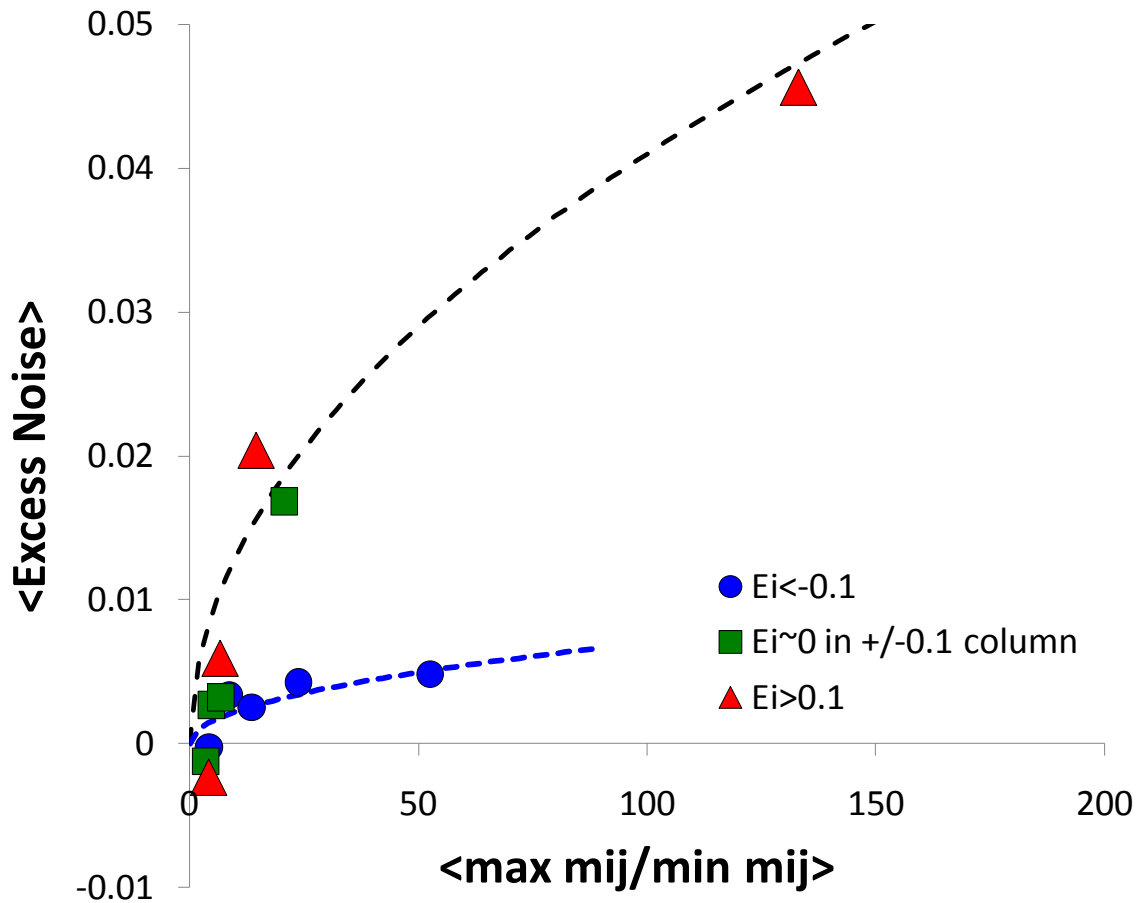


Figure 7.28 Response dependent noise-plasticity coupling in yeast. All three regions of response (induction, neutral, and repression) have a noise-plasticity coupling. Blue genes are deterministically expressed in healthy conditions and repressed under stress. Their coupling trend is much lower than the other response clustering (red and green). Here the model line is $DM \sim C \cdot \text{Sqrt}(PL)$. For the stress induced and neutral response the coefficient was 0.0041. For the blue growth repressed genes $C = 0.0007$.

7.4.3 Additional Representation of Noise-Plasticity Coupling in *E.coli*

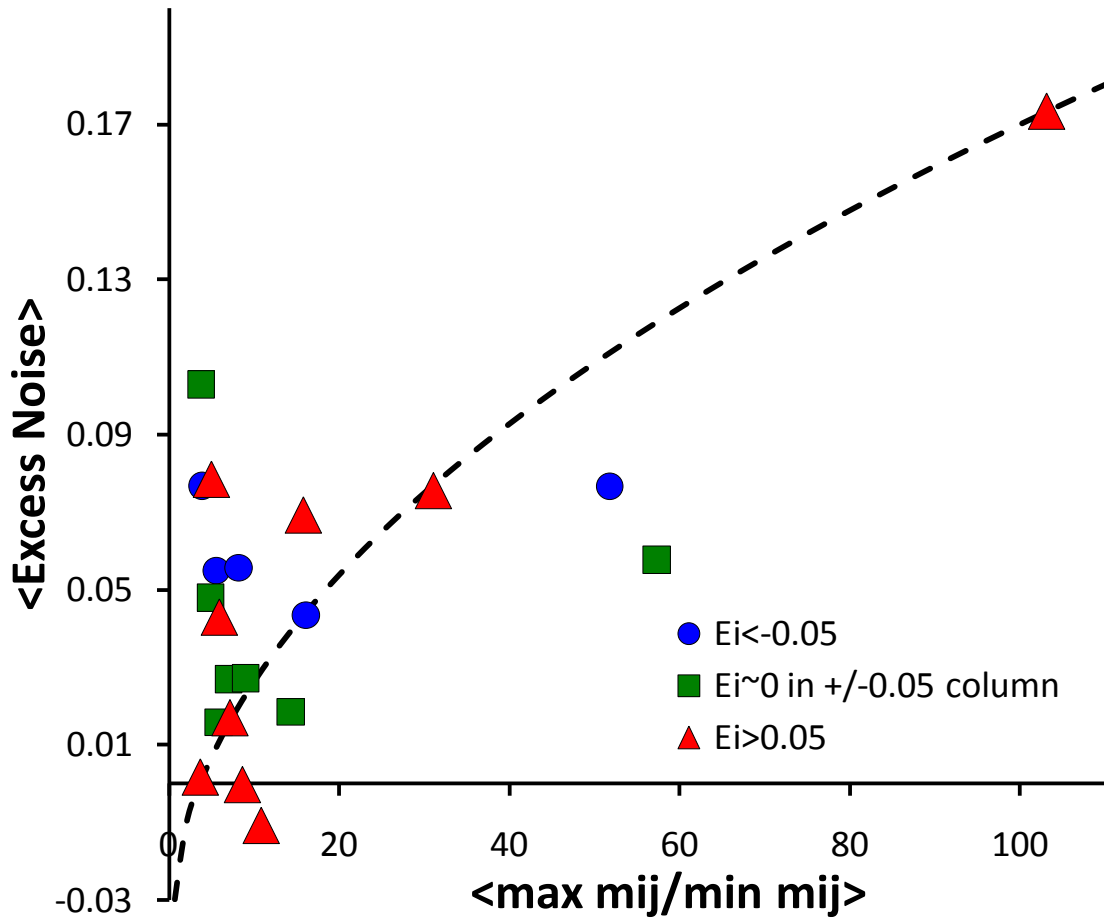


Figure 7.29 Response dependent noise-plasticity coupling in *E. coli*. All three regions of response (induction, neutral, and repression) seem to have a noise-plasticity coupling, but in addition all three have a notable number of high noise-low plasticity genes. E_i is the average transcriptional response as defined in Dar et al, Chaos 2010, red indicates up-regulation to stress, green is a small average response, and blue is down-regulation to stress. The model line (dashed black line) here is $DM = 0.021 * \text{SQRT}(PL) - 0.04$

VITA

Roy Dar was born in April of 1980 in Singapore and moved 3 months later to Manhattan, New York. In 1987 he moved to the California Bay Area and graduated from Los Altos High School (LAHS) in 1997. At LAHS Roy was named to the all-league water polo team and earned MVP, coaches, and John Felix Awards. After high school, Roy and his family moved to Tel Aviv, Israel where he participated in a year of the Exact Sciences Trend in the *Mechina* university preparatory program at the University of Tel Aviv (UTA). At 18, Roy served 3 years in the Israeli Defense Forces (IDF) after which he went directly into the Bachelors of Science Physics and Mathematics program at the Hebrew University of Jerusalem (HUJI), graduating in 2004. After graduation, Roy worked several jobs including a student research assistant position at HUJI in Prof. Dan Davidov's experimental condensed matter physics group studying electron paramagnetic resonance and magnetic nanoparticle absorption of near-field microwave radiation with Fadi Sakran. Roy also excelled in a sales department in a Tel Aviv telemarketing firm. In 2005, Roy joined Dr. Michael Simpson's Molecular-Scale Engineering and Nanoscale Technologies (MENT) research group at Oak Ridge National Laboratory (ORNL) as part of a DOE, Office of Science, Science Undergraduate Laboratory Internship (SULI through ORISE). The internship lasted a total of 8 months working with PhD student Derek Austin on the "Modeling and experimentation of stochastic processes in bacterial cells". The following year, under the direction of Dr. Michael L. Simpson, he enrolled full-time in a doctoral program in Physics at the University of Tennessee, Knoxville, but only after taking off a semester to backpack the whole of Central America. Presently, Roy is excited to continue his scientific development with a postdoctoral appointment in Leor Weinberger's Laboratory at the Gladstone Institute of Virology and Immunology, and Biophysics department at the University of California, San Francisco (UCSF). In his spare time, Roy enjoys nature, art, music, volunteering, and sports.

PUBLICATIONS

1. D. W. Austin, M. S. Allen, J. M. McCollum, **R. D. Dar**, J. R. Wilgus, G. S. Sayler, N. F. Samatova, C. D. Cox, and M. L. Simpson, "Gene Network Shaping of Inherent Noise Spectra," *Nature* **439**, (2006). [link](#)
2. C. D. Cox, J. M. McCollum, D. W. Austin, M. S. Allen, **R. D. Dar**, and M. L. Simpson, "Frequency domain analysis of noise in simple gene circuits," *Chaos* **16**, (2006). [link](#)
3. L. S. Weinberger*, **R. D. Dar***, & M. L. Simpson, "Transient-mediated fate determination in a transcriptional circuit of HIV," *Nature Genetics* **40**, (2008), * -- Equal Contribution. [link](#)
4. C. D. Cox, J. M. McCollum, M. S. Allen, **R. D. Dar**, and M. L. Simpson, "Using noise to probe and characterize gene circuits," *Proc. Nat. Acad. Sci. USA*, **105**(31), (2008). [link](#)
5. M. L. Simpson, C. D. Cox, M. S. Allen, J. M. McCollum, **R. D. Dar**, D. K. Karig, and J. F. Cooke, "Noise in Biological Circuits," John Wiley & Sons, Inc. *WIREs Nanomed Nanobiotechnol* **1** (2009). [link](#)
6. **R. D. Dar**, D. K. Karig, J. F. Cooke, C. D. Cox, and M. L. Simpson, "Distribution and regulation of stochasticity and plasticity in *Saccharomyces cerevisiae*," *Chaos*, (2010). [link](#)
7. **R. D. Dar***, B. Razoogy*, A. Singh, T. Trimeloni, J. M. McCollum, D. K. Karig, J. F. Cooke, C. D. Cox, L. S. Weinberger, and M. L. Simpson, "Detection of an Intrinsic Transcriptional Frequency Across the Human Genome," in preparation, (2011), * -- Equal Contribution.
8. **R. D. Dar**, D. K. Karig, J. F. Cooke, C. D. Cox, and M. L. Simpson, "A Novel Noise-Plasticity Coupling across Unicellular Organism Genomes," in preparation, (2011).