Title
Systematic Dissection of the Determinants of HIV Expression Noise

Author
Foley, Jonathan Emmett

Publication Date
2013

Peer reviewed|Thesis/dissertation

Systematic Analysis of the Determinants of HIV Expression Noise

By

Jonathan Emmett Foley

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Joint Doctor of Philosophy
with University of California, San Francisco

In
Bioengineering
in the
Graduate Division
of the

University of California, Berkeley

Committee in charge:

Professor David V. Schaffer, Co-Chair
Professor Adam P. Arkin, Co-Chair
Professor Jasper Rine
Professor Hana El-Samad

Spring 2013

Abstract

Systematic Analysis of the Determinants of HIV Expression Noise

By
Jonathan Emmett Foley

Joint Doctor of Philosophy in Bioengineering
University of California, Berkeley with
University of California, San Francisco

Professor David V. Schaffer, Co-Chair

Professor Adam P. Arkin, Co-Chair

Despite over three decades of active research, Human Immunodeficiency Virus Type I (HIV-1) represents a pressing and ongoing challenge to our ability to understand complex pathogens that threaten human health at global scale. This stems in part from the highly dynamic and multi-scale lifecycle of the virus, which is reflected in the necessity for multi-drug cocktails to suppress viral replication. A fundamental aspect of the HIV lifecycle is the insertion or integration across a plethora of sites in the human genome. From these sites of integration, the viral promoter coordinates multiple inputs from the cellular signaling environment, the local genetic and epigenetic context, and a virally encoded feedback circuit. Through integration of these inputs via a poorly understood transfer function, the viral promoter regulates the temporal probability of the expression of viral genes and ultimately the kinetics of the viral lifecycle. The HIV promoter, the long terminal repeat (LTR), is composed of a core TATA promoter and a large *cis* acting enhancer. Furthermore, two well-positioned nucleosomes form the basic epigenetic regulatory structures that regulate LTR output. These features are highly similar to endogenous mammalian promoters. Therefore, studies of the LTR can reveal insights relevant to HIV biology as well as mammalian gene control.

HIV represents a highly attractive model system to apply large-scale experimental and computational analysis to systematically study the role of promoter architecture and genomic context in transcriptional regulation. Previous work demonstrated that the LTR operates in a stochastic regime, which results in large cell-to-cell differences in gene expression within clonal populations. In HIV model systems and natural circuits, it has been shown that such heterogeneity or noise can have dramatic phenotypic consequences. However, how such noise is related to the features of promoters, the local genomic context, and their interaction with genetic circuits and cell biology, is poorly understood.

Here using combined large-scale experimental and computational analysis we develop systems to isolate and study the effects of viral integration positions, promoter *cis* acting architecture, transcriptional feedback and post-transcriptional processes on viral gene expression noise. Specifically, by using systematic experimental analysis of single integration clones coupled with computational modeling and statistical inference, we are able to link LTR gene expression noise to dynamic nucleosome occupancy. Furthermore, we develop a method to generate highly diverse combinatoric promoters that will enable systematic study of the LTR structure-function relationship. Additionally, we develop quantitative strategies to cluster highly heterogeneous feedback distributions and reveal characteristic moment scaling relationships. Lastly, using simulations and analysis of mRNA localization using RNA fluorescent *in situ* hybridization, we demonstrate that slow mRNA export can function as a low-pass filter to buffer the cytoplasm from the effects of noisy transcription. Together, these efforts reveal molecular and cellular processes underlying noisy gene expression. Furthermore, we develop novel systems level approaches and tools for the study of complexity in eukaryotic gene control.

Table of Contents

Appendices and Tables

“Everything we see hides another thing, we always want to see what is hidden by what we see. There is an interest in that which is hidden and which the visible does not show us. This interest can take the form of a quite intense feeling, a sort of conflict, one might say, between the visible that is hidden and the visible that is present.”
-René Magritte.

“The clearest way into the Universe is through a forest wilderness.”
― John Muir

Acknowledgements

I would like to express gratitude to several people with whom I have interacted and worked with and those who have assisted me in several ways during my journey of personal and scientific discovery.

I am deeply indebted to my wife, Louisa Roberts, for her support and encouragement through all the highs and lows of graduate studies. I am deeply grateful to my friend and mentor, Robert Beatty, for the constant guidance and encouragement that he has always provided.

During the past seven years I have had the chance to interact with several people in the Schaffer and Arkin laboratories who have helped me with various aspects of these projects. I would like to thank Misha Samoilov, Ron Skupsky, and Morgan Price from the Arkin laboratory with whom I have had countless discussions about stochastic systems analysis and genomics. I would like to thank Siddarth Dey who I collaborated with on the work presented in Chapter 2. Furthermore, I would like to thank John Burnett, Kathryn Miller-Jensen and Priya Shah for their technical expertise and knowledge of HIV and cellular biology. I would also like to thank other members of the Schaffer lab for their camaraderie and Wanichaya Ramey for administrative assistance.

Finally, I would like to thank Professor David Schaffer and Professor Adam Arkin for their terrific scientific mentoring and for being incredibly patient, insightful, and supportive people to work with.

Chapter 1: Background and Motivation

1.1 Gene Expression and Biological Noise

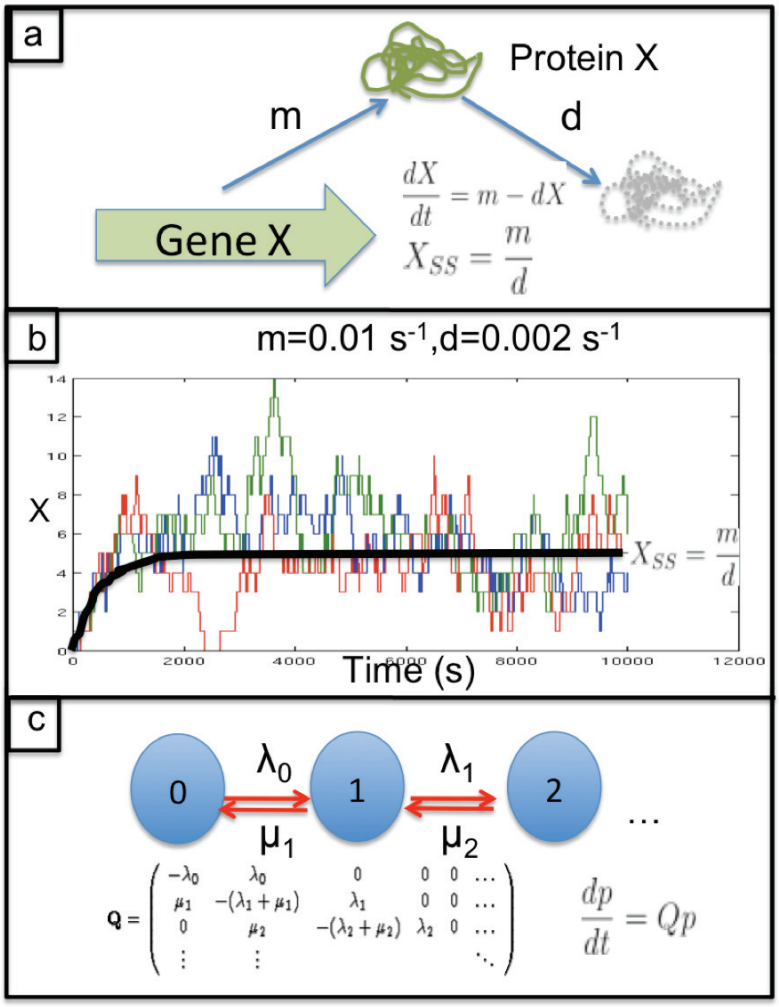


Figure 1.1|The 'birth-death' model of stochastic gene expression
Non-genetic individuality in clonal populations under homogenous environmental conditions can arise from the stochastic expression of genes. The simplest model of gene expression is the 'birth-death' model. Framed as a classical chemical kinetic model (a), the system is predicted to follow a deterministic approach to steady-state (b, black). However, slow-reaction rates and finite molecule effects can produce large excursions from this deterministic behavior. Individual trajectories (b, red, blue, green) of a stochastic stimulation evolve according to state transition probabilities of a continuous-time Markov process (c). The birth-death process is a Markov process in which only transitions that lead to an increase or decrease of the system size by 1 or 0 in each time step are allowed. The time evolution of state probabilities obeys the forward equation consisting of the matrix product of the current state probability with the transition matrix, Q. Averaging over 1000 stochastic realizations approaches the deterministic behavior predicted by classical kinetics with μ_X=4.94 and σ_X^2=4.79 at steady-state.

Reactions exhibiting slow kinetics or low numbers of reagent molecules are susceptible to stochastic effects[1]. Such stochastic effects can produce 'deviant' non-deterministic behavior[2] and trajectories with large excursions from the average time evolution predicted by a system of ordinary differential equations (ODE) that describe the reactions[3]. In the late 1970's, Berg [4] and Rigney[5] demonstrated that the stochastic expression of genes could generate non-genetic individuality[6] in homogenous environments. These studies analyzed a simple 'birth-death' process, which is the most intuitive model of gene expression (Fig.1.1a) The birth-death process is a continuous-time Markov process where the only state transitions are production and degradation (Fig.1.1c). Represented as a kinetic model, production and degradation occur as zeroth and first order reactions, respectively, with rates *m* and *d.* Fluctuations in the number of proteins in a single cell over time, X(t), comprise a stochastic realization (Fig.1.1b) of the system. The system trajectory results from the time evolution of the stochastic birth-death process with the number of proteins X per cell following a Poisson distribution with expectation E(X)=Var(X)=m/d at steady-state. The waiting time between production events is an Exp(1/m) random variable. The average system view of the deterministic ODE model (Fig. 1.1b) represents the average over many stochastic realizations.

The 'noise' in this process is a measure of the relative spread in the distribution of the number of proteins across many cells, and is defined as the variance divided by the square of the mean, $\eta^2 = \frac{\sigma^2}{\mu^2}$ [7].For a simple Poisson gene expression process, where $\sigma^2 = \mu$ and $\eta^2 = \frac{1}{\sqrt{\mu}}$, relative noise decreases with increasing mean expression. The 'noise-strength' is often represented by the Fano factor[8], $\eta = \frac{\sigma^2}{\mu}$, which is equal to 1 for a Poisson process. Deviations above 1 indicate an augmentation of noise and dynamics not captured by a simple Poisson process. This noise can generate significant cell-to-cell heterogeneity[9] of expression levels in clonal populations.

In the context of genetic circuits, such heterogeneity can affect cell fate decisions and result in dramatic phenotypic differences within isogenic populations[10,11] Recent studies in prokaryotic[12], yeast[13] and mammalian[14] systems pioneered the direct measurement of expression levels of protein or mRNA at the level of single cells. These studies demonstrate that transcription is a noisy stochastic process with variance in the distribution in the number of RNA or protein molecules per cell significantly greater than a Poisson distribution. The models proposed by these studies suggest that this noise arises from intrinsic stochastic transitions in promoters between 'OFF' and 'ON' states (Fig.1.2b) causing transcription to occur in infrequent bursts. These studies pioneered techniques to directly quantitate noise in biological systems; however, the processes regulating promoter transitions are not well understood. Specifically, the molecular features that may either enhance or buffer transcriptional noise are poorly understood. Only recently has the role of promoter architecture and transcription factors in regulating noise and quantitative promoter output[15,16]been investigated. These studies used inducible synthetic promoters that lack the combinatoric complexity and regulatory schemes of endogenous eukaryotic promoters; such promoters

function as computational machines integrating multiple inputs from cell and gene states to dynamically regulate expression. Thus, a quantitative understanding of promoter complexity and function would further our ability to model and engineer biological systems.

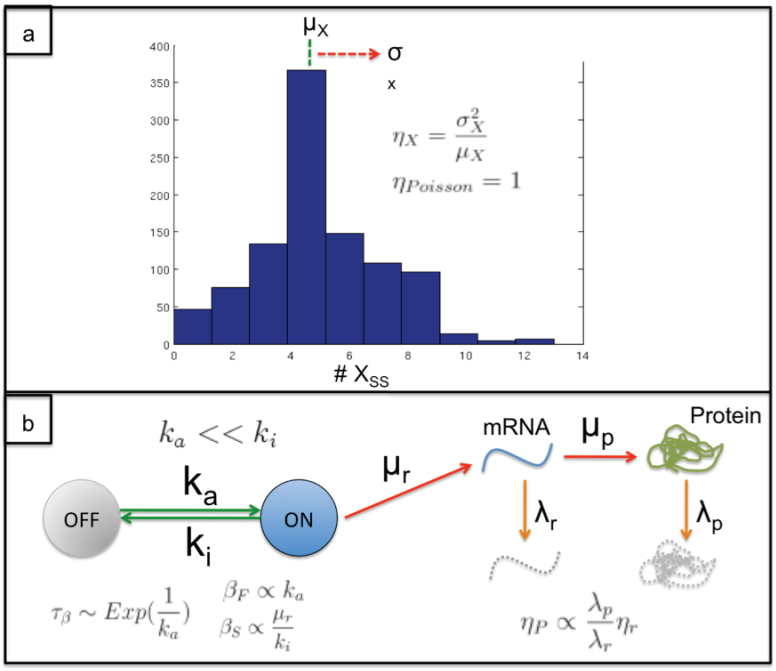


Figure 1.2|Noise in stochastic processes and a model of transcriptional bursting
(a) Noise in the birth-death process is a measure of the relative spread in the distribution of the number of protein molecules per cell at steady-state (X_{SS}) about the sample mean. For a general stochastic process, the sample variance divided by the sample mean (Fano factor,η) is commonly used to measure deviation from a simple Poisson process where $\mu/\sigma^2=1$. **(b)** Studies in prokaryotic[26], eukaryotic[27] and mammalian[28] systems demonstrate that transcription is an intrinsically noisy process characterized by wide distributions in mRNA per cell. The degree to which mRNA noise is transmitted to protein is dependent on their relative half-lives (λ_p/λ_r). These studies suggest that this noise results from infrequent stochastic transitions in promoters from an 'OFF' to an 'ON' state. If the transcription rate in the ON state is large relative to that of the OFF state (0 or $<<\mu_r$), then transcription occurs in infrequent bursts with burst size β_S and frequency β_F). Similarly to the birth-death process, intra-burst times τ_β are assumed to be Exp($1/k_a$) distributed.

1.2 Epigenetic Regulation of Gene Expression

Eukaryotic genomes are packaged into higher order structures via association with octomeric nucleosome complexes; each one is composed of four core histone proteins. Nucleosomes in association with DNA form a 'beads on a string' structure known as chromatin. The association between nucleosomes and DNA is regulated by post-translational modification of the N-terminal tails of histone proteins, the basic unit of the nucleosome. These modifications form a combinatoric repertoire termed the 'histone code' that is 'read' by cellular factors to regulate gene expression[17]. By controlling the access of transcription factors and the basal

transcription apparatus to the genome, these modifications serve to configure regions of the genome abstractly in either an 'open' or 'closed' configuration. These configurations respectively allow or repress information flow from the genome. Chromatin modifications can be highly localized and are inherently dynamic[18], regulated in a gene-specific manner, and undergo transitions in response to cellular and gene states. In this manner, the epigenetic regulatory level is similar to transistor gates in an integrated circuit or valves in micro-fluidic systems that dynamically control the flow of information through complex system architectures.

1.3 HIV Transcriptional Regulation

Hallmarks of HIV, a member of the primate lentivirus family of retroviruses, include error-prone replication and the stable integration of the viral genome in the human genome[19]. Through error-prone reverse transcription of the two genomic viral RNAs packaged in each virus into a single dsDNA, point mutations[20] as well as insertions or deletions of fragments[21] are introduced. Viral factors then mediate the integration of this DNA copy of the viral genome at semi-random sites[22] within the human genome, with a bias toward actively transcribed regions of the genome. Each integration position has different regulatory properties determined by the local chromatin environment. While the chromatin environment of actively transcribed regions is biased towards the 'open' configuration[23], local differences in the histone code may regulate access to the HIV genome and affect the local recruitment of actors involved in activating HIV gene expression[24]. Once integrated, the HIV genome is transcribed by the canonical eukaryotic mRNA transcription apparatus organized around the RNA polymerase II (RNAPII) molecular machine.

The HIV genome encodes a viral promoter, the Long-Terminal-Repeat (LTR), and an autoregulatory positive feedback loop (Fig.1.3a) that interact to dramatically upregulate viral gene expression and virus production. In the absence of Tat, the viral positive feedback protein, basal LTR mediated transcription is weak, producing small numbers of full-length transcripts[25]. This basal expression activity is characterized by the preponderance of abortive 50-100mer transcripts, and stalled RNAPII transcription complexes[26] adjacent to the transcriptional start site (TSS). Furthermore, the two well-positioned nucleosomes that straddle the TSS are nominally hypoacetylated[27] and serve to repress active transcription[28]. The 59mer TAR is produced prior to RNAPII complex stalling and remains tethered with the stalled complex[29]. While most eukaryotic transcription factors enhance transcription through interaction with DNA sequence elements, Tat is non-canonically recruited to TAR rather than to a DNA element in the LTR. This directed recruitment of Tat to TAR situates Tat in close proximity to the stalled RNAPII complex and the nucleosomes adjacent to the transcriptional start site. From this position, Tat orchestrates[30] the recruitment of chromatin remodeling as well as co-activator complexes that result in local chromatin 'opening' and catalyze activation of the stalled RNAPII transcriptional complex.

1.4 Stochastic Tat feedback

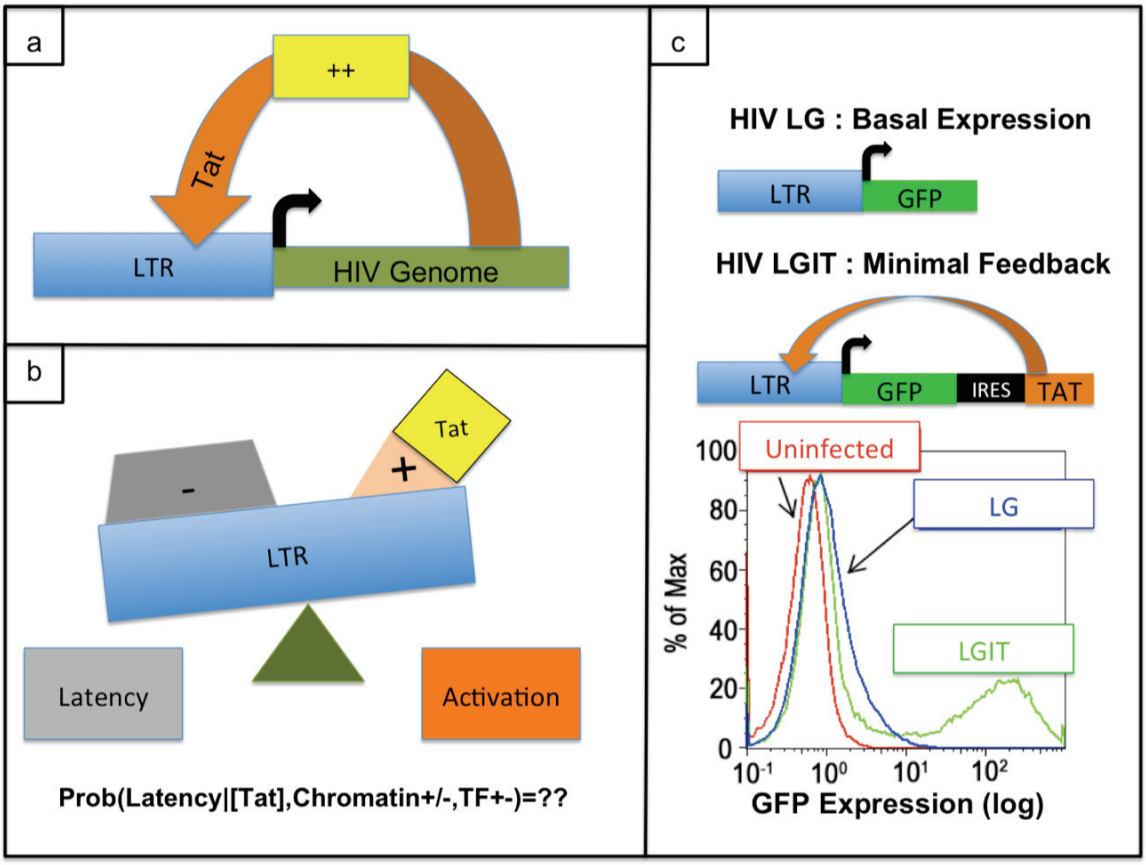


Figure 1.3| HIV LTR and Tat Positive Feedback function as probabilistic switch
(a) HIV encodes a viral promoter, the Long-Terminal-Repeat (LTR), and an autoregulatory positive feedback loop mediated by the viral factor Tat. **(b)** We hypothesize that this feedback loop functions as a probabilistic switch with the probability of latency conditioned on the concentration of Tat, local chromatin environment and transcription factor occupancy. Chromatin and transcription factor inputs can either be activating (+) or repressive (-)**(c)** Basal LTR mediated expression (LG) is weak and highly skewed whereas minimal Tat feedback (LGIT) supports bimodal expression representing dynamic switching between OFF and ON states.

Importantly, both chromatin remodeling and Tat's molecular functions are regulated by factors recruited to *cis* acting sequence elements in the LTR[31]. Furthermore, initial production of Tat is wholly dependent on the recruitment of factors to the LTR that permit low-level productive transcription in the absence of Tat. We hypothesize that because of the slow kinetics of basal HIV transcription and the small molecule numbers of viral and host cell factors (starting with ~1 genome/cell), HIV transcriptional activation is susceptible to stochastic effects. In this manner, operation of the positive feedback circuit functions not as a

deterministic latch to activate viral production, but rather a probabilistic switch between an active and quiescent state. In this context, a positive feedback loop functions as a noise amplifier. The 'gain' of this amplifier modulates the propensity of the switch for the OFF state and phenotypic heterogeneity.

Previous work[32] demonstrates that a minimal HIV Tat positive feedback circuit driving the expression of GFP in a dividing T cell line can support bimodal gene expression in clonal populations. This study proposed a computational model that hypothesizes stochastic fluctuations in the concentration of Tat generate the observed phenotypic bimodality. While this work represents a significant contribution to the field, it is not understood how these fluctuations and the regulation of HIV transcriptional noise are modulated by viral promoter genotype, transcription factor logic, and chromatin dynamics. Furthermore, this model does not generally explain the wide variety of expression phenotypes observed.

1.5 HIV LTR structure and evolution

Interestingly, the error-prone nature of HIV reverse transcription often introduces mutations in the LTR by disabling, deleting, forming, or adding additional copies of *cis* acting elements[33]. This process contributes to a high-degree of sequence diversity in single patients and has contributed to the worldwide evolution of distinct HIV species that differ in the composition and arrangement of *cis* acting elements within the LTR (LTR architecture). However, the functional and phenotypic consequences of these architectural differences and their impact on the dynamics of the Tat feedback circuit are not understood. Furthermore, the modularity and robustness of the LTR to mutation and rearrangement have not been systematically studied.

Importantly, the general structure of the LTR, with an enhancer region upstream of basal transcription initiation elements such as the TATA box, is similar to endogenous mammalian promoters[34]. Furthermore, the chromatin structure of the LTR, with two well-positioned nucleosomes situated about a core enhancer and the transcriptional start site, is also similar to endogenous eukaryotic promoters. This enhancer region as well as many other *cis* acting elements can recruit either repressive or activated complexes depending on cellular state or the stochastic recruitment of interaction partners[35].

The interaction between LTR *cis* acting elements, cellular transcription factors, and the Tat positive feedback circuit forms a transcription factor 'logic' with individual factors and complexes potentially having differential effects on stochastic promoter fluctuations, mean expression level, and noise. Furthermore, inherent heterogeneity in the response of cell signaling to time varying phenomena[36] may bias this logic toward a particular expression pattern. In this view, the LTR is a dynamic state machine with inputs including LTR genotype and transcription factor logic, Tat feedback, and the local chromatin environment that map to a phenotypic output of via an undetermined transfer function P(x).

1.6 Goals and Organization

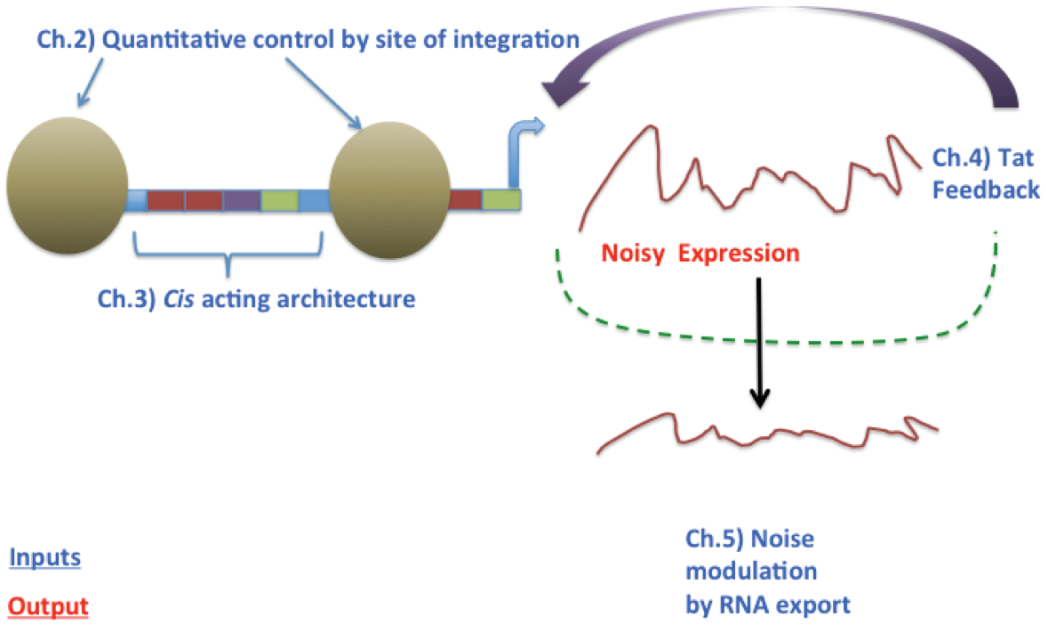


Figure 1.4|Systematic Dissection of the Determinants of Noisy HIV expression
Transcription occurs through the dynamic interaction between DNA sequence elements, the local epigenetic environment of the gene and the physiological state of the cell. Toward the aim of unraveling this complexity, research projects were undertaken that systematically target and study the effects of genomic position and local epigenetic context (symbolized as brown spherical nucleosomes), promoter *cis* acting architecture, transcriptional feedback, and noise modulation by RNA export.

Human Immunodeficiency Virus (HIV), the causative agent of Acquired Immune Deficiency Syndrome (AIDS), currently infects an estimated 33 million individuals globally with 2.7 million new infections per year[37]. HIV is a global pandemic that incurs significant economic, political and social costs[38,39]. Elucidating the mechanisms and dynamics of HIV interactions with its host are crucial to the rational design of targeted therapeutics, with the ultimate goal of mitigating and ultimately eradicating this global scourge. Moreover, the multi-component and time dependent nature[40] of this interaction invites a quantitative systems level investigation of the viral and host processes that contribute to the establishment of clinical disease states. Furthermore, HIV represents a highly attractive model system to apply large-scale experimental and computational analysis to systematically study the role of promoter architecture and genomic context in transcriptional regulation.

Therefore, my research projects have focused on systematically isolating and studying critical inputs that may underlie noisy HIV expression and contribute to expression heterogeneity. In Chapter 2, we apply large scale analysis of single cell clones with flow cytometry, single molecule RNA FISH, and chromatin accessibility measurements to understand how the site of integration quantitatively modulates gene expression noise. In particular, we are able to link chromatin accessibility over specific LTR sites to kinetic parameters in a stochastic model of transcription. During reverse transcription, recombination between the two RNA molecules packaged in each HIV virion, as well as the insertion and deletion of short DNA elements (indels), are fundamental processes in retroviral evolution. These processes contribute to the differences in motif composition and multiplicity observed in subtype LTR[41]. However, the effects of such architectural differences have not been systemically studied. In Chapter 3, we describe a system for generating large combinatoric libraries of synthetic promoters that can be used to study viral evolution, to facilitate control of gene expression output in synthetic systems and in gene therapy applications. In Chapter 4, we quantitatively analyze the highly heterogeneous expression phenotypes generated by a minimal Tat feedback circuit and develop a system for relating basal expression noise to Tat feedback dynamics. Lastly, in Chapter 5 we report on the novel observation of the modulation of expression noise by RNA export from the nucleus and describe an experimental system to demonstrate a functional and biologically significant role for this phenomenon. Together, these efforts have revealed new insight into the mechanisms underlying stochastic gene expression and have generated experimental and computational tools to systematically study noisy gene expression at large scale in single cells.

1.7 References

1. Maheshri, N. & O'Shea, E. K. Living with noisy genes: how cells function reliably with inherent variability in gene expression. *Annual review of biophysics and biomolecular structure* **36**, 413–434 (2007).
2. Samoilov, M. S. & Arkin, A. P. Deviant effects in molecular reaction pathways. *Nature biotechnology* **24**, 1235–1240 (2006).
3. Kaern, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics* **6**, 451–464 (2005).
4. Berg, O. G. A model for the statistical fluctuations of protein numbers in a microbial population. *Journal of theoretical biology* **71**, 587–603 (1978).
5. Rigney, D. R. Stochastic model of constitutive protein levels in growing and dividing bacterial cells. *Journal of theoretical biology* **76**, 453–480 (1979).
6. Spudich, J. L. & Koshland, D. E. Non-genetic individuality: chance in the single cell. *Nature* **262**, 467–471 (1976).
7. Paulsson, J. Models of stochastic gene expression. *Physics of Life Reviews* **2**, 157–175 (2005).

8. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & Van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nature genetics* **31**, 69–73 (2002).
9. Colman-Lerner, A. *et al.* Regulated cell-to-cell variation in a cell-fate decision system. *Nature* **437**, 699–706 (2005).
10. Pedraza, J. M. & Van Oudenaarden, A. Noise propagation in gene networks. *Science* **307**, 1965–1969 (2005).
11. Arkin, A., Ross, J. & McAdams, H. H. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. *Genetics* **149**, 1633–1648 (1998).
12. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic Gene Expression in a Single Cell. *Science* **297**, 1183–1186 (2002).
13. Raser, J. M. & O'Shea, E. K. Noise in gene expression: origins, consequences, and control. *Science (New York, NY)* **309**, 2010–2013 (2005).
14. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA Synthesis in Mammalian Cells. *PLoS Biology* **4**, e309 (2006).
15. Murphy, K. F., Bal�zsi, G. & Collins, J. J. Combinatorial promoter design for engineering noisy gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 12726–12731 (2007).
16. Cox, R. S., Surette, M. G. & Elowitz, M. B. Programming gene expression with combinatorial promoters. *Molecular systems biology* **3**, 145 (2007).
17. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).
18. Barrera, L. O. & Ren, B. The transcriptional regulatory code of eukaryotic cells--insights from genome-wide analysis of chromatin organization and transcription factor binding. *Current opinion in cell biology* **18**, 291–298 (2006).
19. Fields, B. N., Knipe, D. M. & Howley, P. M. *Fields' virology*. 2 v. (xix, 3091, 86 p.) (Wolters Kluwer Health/Lippincott Williams & Wilkins: Philadelphia, 2007).
20. Ji, J. P. & Loeb, L. A. Fidelity of HIV-1 reverse transcriptase copying RNA in vitro. *Biochemistry* **31**, 954–958 (1992).
21. Wu, W., Blumberg, B. M., Fay, P. J. & Bambara, R. A. Strand transfer mediated by human immunodeficiency virus reverse transcriptase in vitro is promoted by pausing and results in misincorporation. *The Journal of biological chemistry* **270**, 325–332 (1995).
22. Schroder, A. R. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521–529 (2002).
23. Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C. & Bushman, F. D. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res* **17**, 1186–1194 (2007).
24. Bushman, F. *et al.* Genome-wide analysis of retroviral DNA integration. *Nat Rev Microbiol* **3**, 848–858 (2005).
25. Lassen, K. G., Bailey, J. R. & Siliciano, R. F. Analysis of human immunodeficiency virus type 1 transcriptional elongation in resting CD4+ T cells in vivo. *J Virol* **78**, 9105–9114 (2004).
26. Cullen, B. R. The HIV-1 Tat protein: an RNA sequence-specific processivity factor? *Cell* **63**, 655–657 (1990).

27. Williams, S. A. *et al.* NF-kappaB p50 promotes HIV latency through HDAC recruitment and repression of transcriptional initiation. *The EMBO journal* **25**, 139–149 (2006).
28. He, G. & Margolis, D. M. Counterregulation of chromatin deacetylation and histone deacetylase occupancy at the integrated promoter of human immunodeficiency virus type 1 (HIV-1) by the HIV-1 repressor YY1 and HIV-1 activator Tat. *Molecular and cellular biology* **22**, 2965–2973 (2002).
29. Fujinaga, K. *et al.* Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element. *Molecular and cellular biology* **24**, 787–795 (2004).
30. Gatignol, A. Transcription of HIV: Tat and cellular chromatin. *Advances in pharmacology (San Diego, Calif)* **55**, 137–159 (2007).
31. Barboric, M. & Peterlin, B. A new paradigm in eukaryotic biology: HIV Tat and the control of transcriptional elongation. *PLoS Biology* (2005).
32. Weinberger, L. S., Burnett, J. C., Toettcher, J. E., Arkin, A. P. & Schaffer, D. V Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell* **122**, 169–182 (2005).
33. Abram, M. E., Ferris, A. L., Shao, W., Alvord, W. G. & Hughes, S. H. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *Journal of virology* **84**, 9864–78 (2010).
34. Marcello, A. *et al.* Nuclear organization and the control of HIV-1 transcription. *Gene* **326**, 1–11 (2004).
35. Pereira, L. A., Bentley, K., Peeters, A., Churchill, M. J. & Deacon, N. J. A compilation of cellular transcription factor interactions with the HIV-1 LTR promoter. *Nucleic Acids Res* **28**, 663–668 (2000).
36. Feinerman, O., Veiga, J., Dorfman, J. R., Germain, R. N. & Altan-Bonnet, G. Variability and robustness in T cell activation from regulated heterogeneity in protein levels. *Science* **321**, 1081–1084 (2008).
37. 2012 UNAIDS Report on the Global AIDS Epidemic. at <http://www.unaids.org/en/resources/publications/2012/name,76121,en.asp>
38. Leibowitz, A. A. & Sood, N. Market power and state costs of HIV/AIDS drugs. *International journal of health care finance and economics* **7**, 59–71 (2007).
39. Dunne, M. Antiretroviral drug development: the challenge of cost and access. *AIDS (London, England)* **21 Suppl 4**, S73–9 (2007).
40. Lassen, K., Han, Y., Zhou, Y., Siliciano, J. & Siliciano, R. F. The multifactorial nature of HIV-1 latency. *Trends in molecular medicine* **10**, 525–531 (2004).
41. Burnett, J. C. *et al. Combinatorial latency reactivation for HIV-1 subtypes and variants.*

Chapter 2: High-throughput analysis reveals orthogonal variation of expression mean and noise across genomic locations with nucleosome occupancy underlying promoter transitions and noise

2.1 Introduction

Isogenic populations grown under identical conditions have long been known to exhibit non-genetic heterogeneity[1]. This non-genetic heterogeneity arises from the inherently random nature of biochemical processes and infrequent reactions that become significant when few molecules are present in the system. Such random fluctuations or noise can be further amplified by the flow of information through the underlying pathway to produce dramatic phenotypic variations within isogenic populations ranging from single celled bacteria and yeast to insect and mammalian cells[2–7].In particular, gene expression noise arising from stochastic fluctuations in transcription has been shown to be an important source of non-genetic heterogeneity in mammalian cells. Over the last decade, significant work has been conducted to experimentally validate and model the process of stochastic gene expression[8–13]. These advances in our understanding have been driven by the development of powerful single cell analysis techniques such as flow cytometry, high-throughput microscopy, and single molecule RNA fluorescent *in situ* hybridization (smFISH). However, the contribution of different molecular factors in regulating gene expression noise remains largely unknown.

Both the genomic location with its associated chromatin environment and the promoter architecture can potentially regulate gene expression noise. Elegant large-scale studies in simple eukaryotic organisms like *S. cerevisiae* identified the scaling of noise with mean protein expression for several endogenous genes[14]. However, the contribution of the promoter architecture and genomic location to gene expression noise was not explored in this study. Similarly, another study in the context of the well-characterized yeast Pho5 promoter[15] identified the contribution to expression noise of different transcription factor mutants on nucleosome remodeling during transcriptional activation However, it was not clear if the results were dependent on the particular endogenous locus or reflective of dynamics across the whole genome. Further, while transcriptional regulation is similar between yeast and mammals, mammalian genomes exhibit additional complexity, with differences in both large scale chromatin dynamics and promoter proximal chromatin marks, that may limit the generalization of the findings in yeast to mammals[16,17].A recent study in mammalian cells looked at the impact of *cis*-regulatory sequences on gene expression noise[18].However, the limited dataset and the use of small-molecule perturbations that globally alter chromatin were unable to provide a clear understanding of the role of the genomic location in regulating gene expression noise. Therefore, we designed a lentiviral-based system that effectively deconvolves the influence of the promoter architecture from the genomic location, thereby allowing a comprehensive and systematic study of gene expression noise regulation by the genomic environment in a mammalian system.

Due to the semi-random nature of lentiviral integration, and the molecular features of lentiviral promoters, an HIV-1 based lentiviral system provides an attractive model to study the role of genomic context in regulating gene expression noise. In particular, semi-random integration efficiently samples a myriad of genomic locations while maintaining the same promoter architecture. Further, lentiviral promoters exhibits features that are archetypal of endogenous eukaryotic promoters, such as a TATA box, extensive *cis* acting elements and well-positioned nucleosomes along the promoter[19,20].These characteristics make a HIV-1 based lentiviral system ideally suited for investigating the contribution of the genomic location to gene expression noise by decoupling the contribution of promoter sequence variation. Specifically, to develop the first large-scale data set to study the influence of genomic location on gene expression noise we isolated a large set of single-cell clones representing hundreds of integration positions.

While several studies have shown that cell-to-cell heterogeneity in genetically identical cells may produce different phenotypes with important consequences in diverse areas of biology such as development, cancer, bacterial and viral infections, little is known about the underlying molecular factors that regulate gene expression noise. Recent studies in *S. cerevisiae* have begun to unravel the role of chromatin and chromatin modifying complexes in regulating the dynamics of gene expression[15,21]. We extend these studies to a mammalian system to systematically understand the role of chromatin in regulating transcriptional bursting. Identifying the molecular players and sources of non-genetic heterogeneity may provide better tools to target diseases such as cancer and pathogenic infections[3,6,7,22] .

While others and we have previously used lentiviral based vectors to study gene expression noise, small datasets and indirect measurements of transcription have resulted in discrepancies in the results. Furthermore, these studies have been unable to infer specific molecular features underlying the regulation of expression noise. We previously found that the transcript burst size primarily modulates the mean level of gene expression whereas another study recently found that both the burst size and frequency vary with mean expression across genomic locations[23,24] .To discriminate between these conflicting results, we conducted the first large-scale analysis to accurately quantify both mRNA by smFISH and protein by flow cytometry. Furthermore, to gain deeper molecular insight we assessed chromatin accessibility across integration positions. We used this combined data to parse out the exact contribution of different model parameters to features of gene expression noise. Finally, this study quantitatively links promoter chromatin density to promoter fluctuations and the generation of noise across genomic locations.

2.2 High-throughput generation of single-integration clones to comprehensively capture observed diversity in both expression mean and expression noise

To facilitate comprehensive systematic analysis of the modulation of LTR basal transcriptional dynamics across genomic locations, a high-throughput

workflow (Fig. 1A) was designed to raise a diverse and representative set of clones. A model vector (LGM2) consisting of the HIV-1 LTR driving dual protein (GFP) and RNA (M2 smFISH) reporters was developed for the simultaneous analysis of RNA and GFP in single cells of clonal populations. Furthermore, permitting direct comparison of protein and RNA levels provides a deeper and more direct inference of the transcription dynamics than previously studied. Importantly, previous studies have used small sets of less than 40 clones[23,25] or used a short half-life GFP (d2GFP) variant that, due to reduced fluorescence, precludes analysis of a significant subset of genomic locations[24]. Furthermore, previous studies have strictly used GFP reporters of gene expression activity. Together, these features may provide a biased or limited portrait of expression dynamics across genomic locations and thereby limit the comprehensive quantitative analysis of clonal gene expression properties.

Therefore, to capture a spectrum of single-integration clones that robustly represent the diverse expression properties exhibited in the wide and highly skewed bulk distribution, single cells were sorted into 960 wells with a sorting probability of 0.5 per well from a wide gate (Fig. 1B) spanning 1.8 $\log_{10}$ RFU units and distinct from Autofluorescence. This gate was chosen to limit sampling bias toward any particular regime of mean expression level and resulted in 227 LGM2 clones, which to our knowledge represents the largest set of clones yet studied for HIV model or viral systems. Further, to validate results obtained from the 227 LGM2 clones and to ensure that the results were independent of any cell culture or sorting artifacts, we repeated the scheme described above to isolate an additional 191 single-integration clones (Appendix A, Fig.A5).

The expression of each of the 227 clonal populations was determined by flow cytometry and each clone subsequently gated (Appendix A, Fig.A1) with a data driven algorithm that samples a small region of forward scatter and side scatter space and retains at least $4*10^3$ cells per clone. We found minimal dependence between the gate chosen and the resulting GFP distribution moment scaling. (Appendix A, Fig.A2, Fig.A3). This ensemble of clones exhibited mean expression spanning 2 orders of magnitude, which is identical to the range observed in the bulk distribution. (Fig. 2.1C) Furthermore, it is apparent that for a given mean level of expression there is considerable variation in distribution width. If these distributions arose from a constant rate Poisson process in which the promoter is always in a productive state, we would expect distribution Variance to scale linearly with Mean and the coefficient of variation (CV,noise) to decrease as $CV \propto 1/\sqrt{Mean}$. In contrast, for an idealized Telegraph process in which RNA is produced in infrequent bursts arising from stochastic promoter transitions from a 'Off' state and an 'On' state, we would expect Variance to be related to Mean through a power-law relationship and the noise as measured by CV to be uncorrelated from Mean.

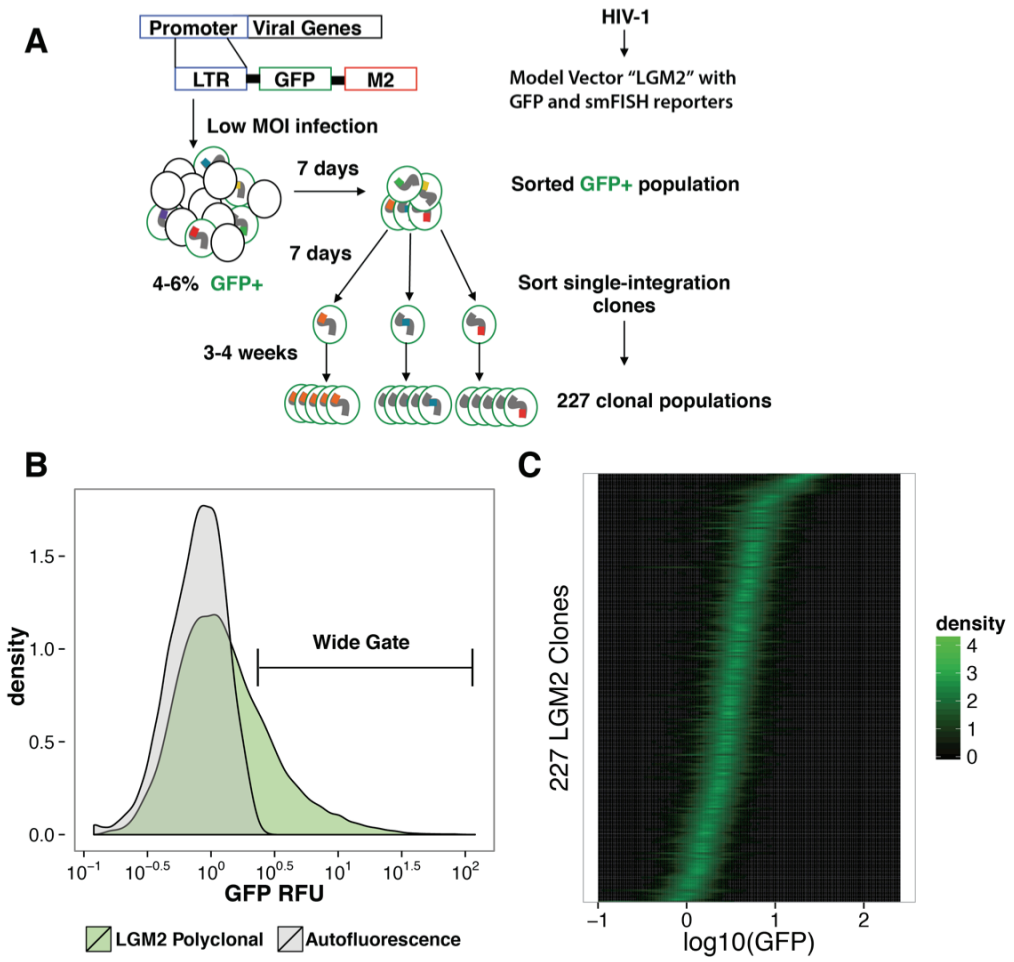


Figure 2.1|High-throughput single integration clone generation robustly captures the diversity of expression mean and noise exhibited by HIV-1 across integration sites. (A)**Clonal generation workflow:** Jurkat T cells were infected at low MOI (~ 0.04) with HIV-1 LGM2 and allowed to reach steady state expression for 7 days. To facilitate discrimination of infected cells by GFP fluorescence, the culture was stimulated with TNF-α for 16 hrs. Following stimulation, $3*10^5$ GFP positive cells were sorted and allowed to expand for 7 days. From this GFP+ population, single cells were sorted into ten 96 well plates and allowed to expand for 3-4 weeks to generate 227 clonal populations. (B) **Unbiased clone sorting from long tailed polyclonal distribution:** As indicated by the overlay of the polyclonal LGM2 GFP distribution with autofluorescence of uninfected Jurkat T cells, the HIV-1 LTR supports wide and highly skewed gene expression with cells exhibiting expression 1-2 orders of magnitude greater than the very low expression mode. For further analysis of the modulation of LTR gene expression across viral integrations, an unbiased gate distinct from autofluorescence was used to capture a diverse and representative set of single cell clones. (C) **Evidence of noise independence from mean expression across large set of clones:** Analysis of GFP expression of the 227 clones by flow cytometry reveals modulation of both the mean level of expression and expression 'noise' as indicated by significant differences in distribution width within groups of clones with similar means. Each row within the plot represents the 1D kernel smoothed density of the GFP expression of a single clone, with the clones rank ordered by mean expression.

2.3 Uncorrelated expression mean and noise suggest primarily orthogonal control across genomic locations

Therefore, in order to infer whether either of these basic models of gene expression can explain the observed distributions, we next examined the relationships between expression mean and variance and between expression mean and the coefficient of variation (expression noise). Examination of the relationship (Fig. 2A blue line) between GFP mean and GFP variance reveals that in log transformed space clonal distribution variance is highly correlated with distribution mean (R^2= .86 Spearman correlation coefficient r_s=0.92). This strongly suggests an underlying power-law relationship with $\sigma^2 \propto \mu^{2.1\pm0.1}$. This scaling is distinct from Poisson scaling (Fig. 2A, dashed black), with an expected linear relationship resulting from a constant rate of production, and stringently consistent ($p<0.01$) with distribution scaling arising through promoter transitions between Off and On states in an idealized telegraph process (theoretical slope of 2, dashed green line).

While scaling between the first two moments is a useful comparison to theoretical bounds, a normalized dimensionless assessment of expression noise with respect to mean such as the coefficient of variation (CV, σ/μ) provides a direct comparison to other experimental systems. Consistent with an underlying Telegraph process, examination of the relationship (Fig. 2B blue line) between GFP CV and GFP mean reveals an uncorrelated relationship (R^2=0.03,r_s=0.19). Interestingly, for a given mean level of expression a nearly constant range of CV) is sampled (Appendix A, Fig.A5). This is in stark contrast to recent genome wide studies of expression noise in yeast[14,26] ,which have indicated that $\sim CV \propto 1/\sqrt{\mu}$. We do not find that this lack of correlation is an artifact of the gating strategy used (Appendix A, Fig.A4). Furthermore, to ensure that these results were not influenced by artifacts introduced during infections, cell culture or cell sorting, the experimental scheme described in Figure 1A was repeated to isolate 191 single-integration clones that validated the lack of correlation between mean expression and CV. This strongly suggests that across genomic locations expression mean and expression noise are differentially controlled.

2.4 Clone subsetting for further analysis by smFISH reflects properties of the full set of clones

While the above results provide an initial, diagnostic assessment of the most likely underlying transcriptional mechanism, the observations are inherently convolved with post-transcriptional processes that may obscure important underlying information. In particular, the long half-life of the GFP reporter potentially buffers and smooths important dynamic information. Therefore, in order provide a more direct inference of the transcriptional mechanism underlying the observed differential variation of expression mean and expression noise across genomic locations, a subset of 25 clones selected for analysis by smFISH with a probe against the M2 array in the integrated LGM2 provirus.

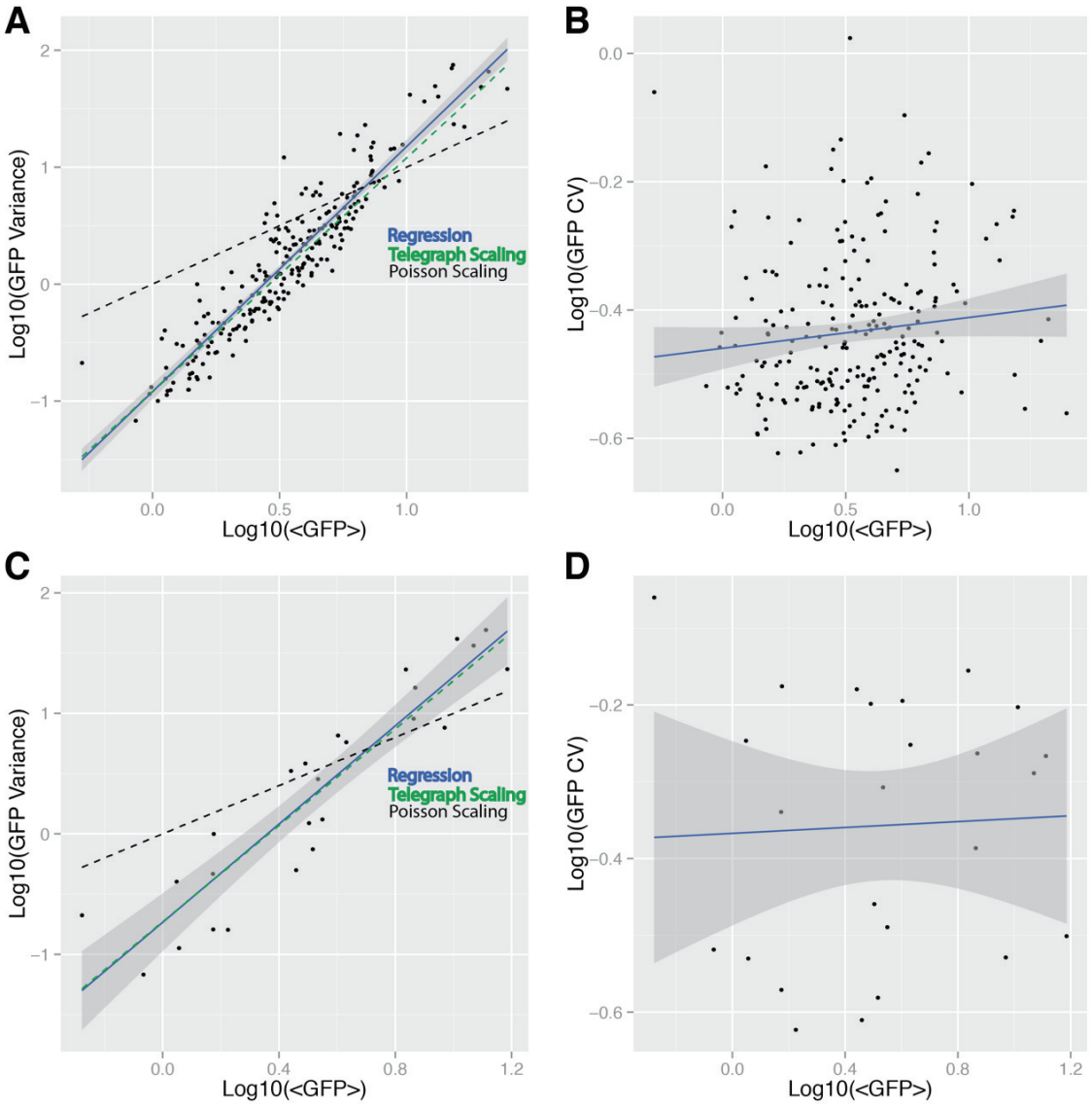


Figure 2.2|Scaling between distribution mean and variance is highly consistent with telegraph-like transcription while noise scaling suggests independent control of expression mean and noise across integration positions. (A) **Clonal distribution moment scaling:** GFP fluorescence of approximately 10^4 cells from each clone were measured via flow cytometry. The highly significant power law relationship between distribution mean and variance with a log-log linear regression slope of 2.1±0.11(R^2=0.86,r_s =0.92,p<0.001) is distinct from Poisson scaling (dashed line) and is consistent with distribution scaling arising through 'input controlled' promoter transitions between "Off" and "On" states in an idealized telegraph process (theoretical slope of 2, dashed green line) (B**) Expression Mean and noise independence:** Uncorrelated GFP expression noise, measured as log –log transformed Coefficient of Variation (CV), and GFP mean across clones (R^2=0.013, r_s=0.12) suggests independent control of mean expression and expression noise across integration positions. (C,D) Clone selection for further analysis: The 227 clones were clustered (Appendix A, Fig.A6) based on their GFP distributions to identify a subset of 25 representative clones that were chosen for further analysis to capture the span of expression means and variances observed in the full set of clones (See Appendix A for details about clustering). Specifically, pairs of

clones were selected with similar means but markedly different CV. Following this subsetting, the resulting relationships between log-log transformed variance and mean (slope=2.03±0.36, R^2=0.85, r_s =0.93, p<0.001) and between mean and CV (R^2=0.02,r_s=0.05) indicate that this subset reasonably reflects the scaling relationships observed in the full set of clones. r_s represents the Spearman correlation coefficient for the explanatory and response variables in each pairwise regression while p values represent support for the correlation.

These clones were selected by clustering 227 distributions into four clusters using a *k-means* approach (Appendix A, Fig.A6). Within each cluster, pairs of clones with similar means but markedly different CV were selected. This effectively samples both the range of expression means and noises observed in the full set of clones thereby allowing us to make detailed inferences about molecular factors regulating gene expression dynamics as a function of the genomic location. Indeed, examination of the relationship between mean and variance (Fig.2.2C) and mean and CV (Fig.2.2D) for this subset of clones reveals trends that are not statistically different (F-test, p>0.1) from those observed in the full set of clones. Using this approach we analyzed both high noise (Fig. 2.3A) and low noise (Fig. 2.3B) clones to generate exact distributions of RNA copy number per cell.

.2.5 Hybrid unsupervised image-processing enables the high-throughput analysis of over fifteen thousand single cells across many clones

Previously, owing to the intensive task of image processing, smFISH has been performed on limited numbers of cells across few conditions, which may significantly limit inference of distribution shape and model fitting. We performed smFISH at a large scale by imaging all 25 clones in 3 channels across 84 fields (capturing ~5-10 cells per field) with 91 axial slices at 0.2 μm spacing. This spacing was determined to provide optimal resolution of smFISH signals and record the entire cell volume. Importantly, performing RNA FISH at this scale presented two challenges: (1) accurate segmentation of cells in fields with touching cells, and (2) reliable identification of single molecule FISH signals across many samples. Previously reported methods involve either z-projection of image stacks[27], manually intensive threshold selection[28], or training set development[29]. Z-projection flattens the recorded image stack and exhibits a propensity to generate overlapping signals that then must be algorithmically separated. Similar to manual threshold selection, training set development requires significant manual intervention. Furthermore, the potential to over-fit limited training data makes scaling to many samples uncertain.
To overcome these issues and enable smFISH at a greater scale we implemented custom software that segments complex fields in a highly automated fashion and reliably segments smFISH signals with the minor requirement of manual estimation of a 'ballpark' threshold that weakly effects the resulting counts. Efficient cellular segmentation was achieved through a multi-step process involving normalization of heterogenous fluorescent intensities and morphological reconstruction of the raw GFP channel, followed by intensity thresholding and

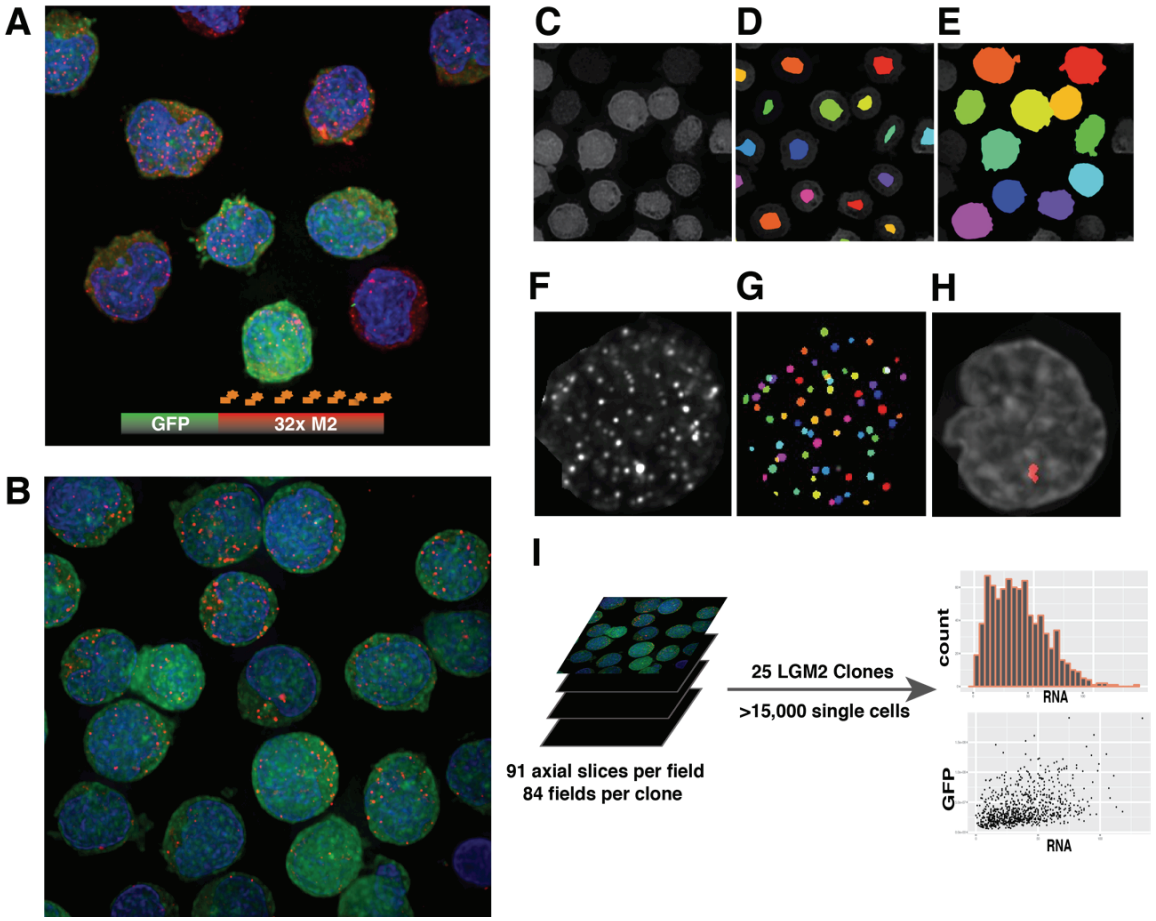


Figure 2.3| Hybrid unsupervised segmentation of both cells and smFISH signals enables high-throughput analysis of expression output of thousands of single cells (A,B) Transcriptional bursting and expression heterogeneity in high and low clones: (A) Significant heterogeneity in RNA copy number and GFP within a high noise clonal population (HIV-1 LGM2 clone GD1) and direct evidence of transcriptional bursting is revealed by a false-colored maximum intensity projection (MIP) of a deconvolved wide-field optically sectioned field of cells imaged at 100X in three channels (GFP, DAPI, TAMRA). **(C-H) High-throughput image analysis of microscopy images:** Unsupervised pseudo-3D segmentation of cells proceeds through a multistep process depicted through maximum intensity projections (MIP) of operations performed on a stack of image per field. First, the raw stack of GFP fluorescent images (C) undergo homomorphic filtering and morphological reconstruction (C) to normalize variation in intensity and fill in structural holes. Grey level thresholding followed by Euclidean distance thresholding results in unique seeds for each cell (D), which are depicted as color mapped objects overlaid on the reconstructed MIP. The resulting seeds are used together with the reconstructed stack as inputs for a seeded watershed to arrive at the final segmentation (E). **(F-H)** Unsupervised segmentation of smFISH labeled RNA in single cells occurs through a multi-layered filtering scheme to remove background, amplify particulate signals, and discriminate between single particle signals and nascent transcriptional bursts. (F) Raw MIP of smFISH channel depicts punctate and approximately spherical signals amidst slowly varying background. (G) Following Top-Hat and Laplacian of Gaussian filtering, connected components analysis followed by a heuristic multi-parameter based classification scheme robustly discriminates smFISH signals. (I) Nascent transcriptional bursts are discriminated from single molecule objects through shape and intensity based parameters. **(I) High-throughput image analysis of microscopy images:** Custom morphological segmentation of cells in fields and FISH signals enabled analysis of over 15,000 cells across 25 clones. For each clone a distribution of RNA number per cell is determined.

application of a thresholded Euclidean distance transform to generate unique seeds for each cell, which were then grown using connected components on the reconstructed image for final segmentation (Fig. 2.3 C-E, Appendix A, Fig. A9). Resulting segmented cell objects were checked for consistency against a model of cell shape and suspect fields were held for manual review. In practice, such manual review represents less than 10% of total fields. This enables the segmentation of hundreds of fields representing thousands of cells within hours.

To reliably identify smFISH signals, each single cell object was subjected to a multi-step filtering, segmentation and heuristic classification. Briefly, raw probe channel stacks were filtered by the morphological top-hat filter with an elliptical structuring element that seperates the approximately spherical FISH signals from the majority of background. A Laplacian of Gaussian filter[13] was applied to further amplify spherical signals. This transformed image is thresholded using a fixed threshold per clone (Fig. 2.4 F-H,Appendix A, Fig.A11) and segmented through connected components analysis. Rather than relying strictly on a threshold, we classified objects on the basis of size, circularity and intra-object grey level intensity standard deviation. We empirically determined thresholds on these features that yielded classification in good agreement with manual classification. The combination of these three features efficiently distinguishes aberrant blobs from FISH signals. Specifically, to segment FISH signals connected component analysis is performed on the thresholded image. Very small (<30 total pixels) and very large (>300) objects are excluded in this step, which provides an initial classification filter. Subsequently, objects are only accepted if their perimeter to area (p2a) values are highly consistent with a sphere (p2a between 0.96 and 1.02). Secondly, background blobs have very uniform pixel intensity across the object while FISH signals have radially decreasing intensity from the center of the diffraction limited spot. This is captured most succinctly in the pixel grey level standard deviation within each object. Therefore, objects with a grey level standard deviation below an empirical threshold are also rejected.. Such multi-parameter classification proved more time-efficient and robust (relative to signals distinguished by eye) to inclusion of background artifacts across many images and clones than thresholding alone.

2.6 RNA distribution shape is highly related to protein distribution shape

High-throughput computational analysis resolved counts of LGM2 RNA per cell for an average of 630 cells per clone. Previous studies[13,30] have relied on sparsely sampled distributions that may inaccurately estimate distribution properties. Furthermore, it is frequently assumed that translation is a constant rate first order process that does not vary between clonal populations. Under such an assumption we would expect RNA Mean and Variance to strongly predict protein Mean and Variance. To examine this assumption directly and to discriminate between stereotypical Poisson and Telegraph transcriptional processes, we determined the relationships for RNA variance as a function of mean (Fig. 4A, blue line) and RNA CV (Fig. 4B, blue line) as a function of mean.

We find that RNA mean and variance are highly correlated (R^2=0.89, r_s=0.93) with $\sigma_{RNA}^2 \propto \mu_{RNA}^{1.6\pm0.25}$.Furthermore, we find that RNA CV is weakly dependent RNA

mean (R^2=0.31, r_s=-0.38,p=0.03. Together, these results suggest that RNA distributions may reflect a mechanism in between idealized Poisson and Telegraph processes. Furthermore, post-transcriptional steps may augment noise and thereby further decorrelate mean from CV.

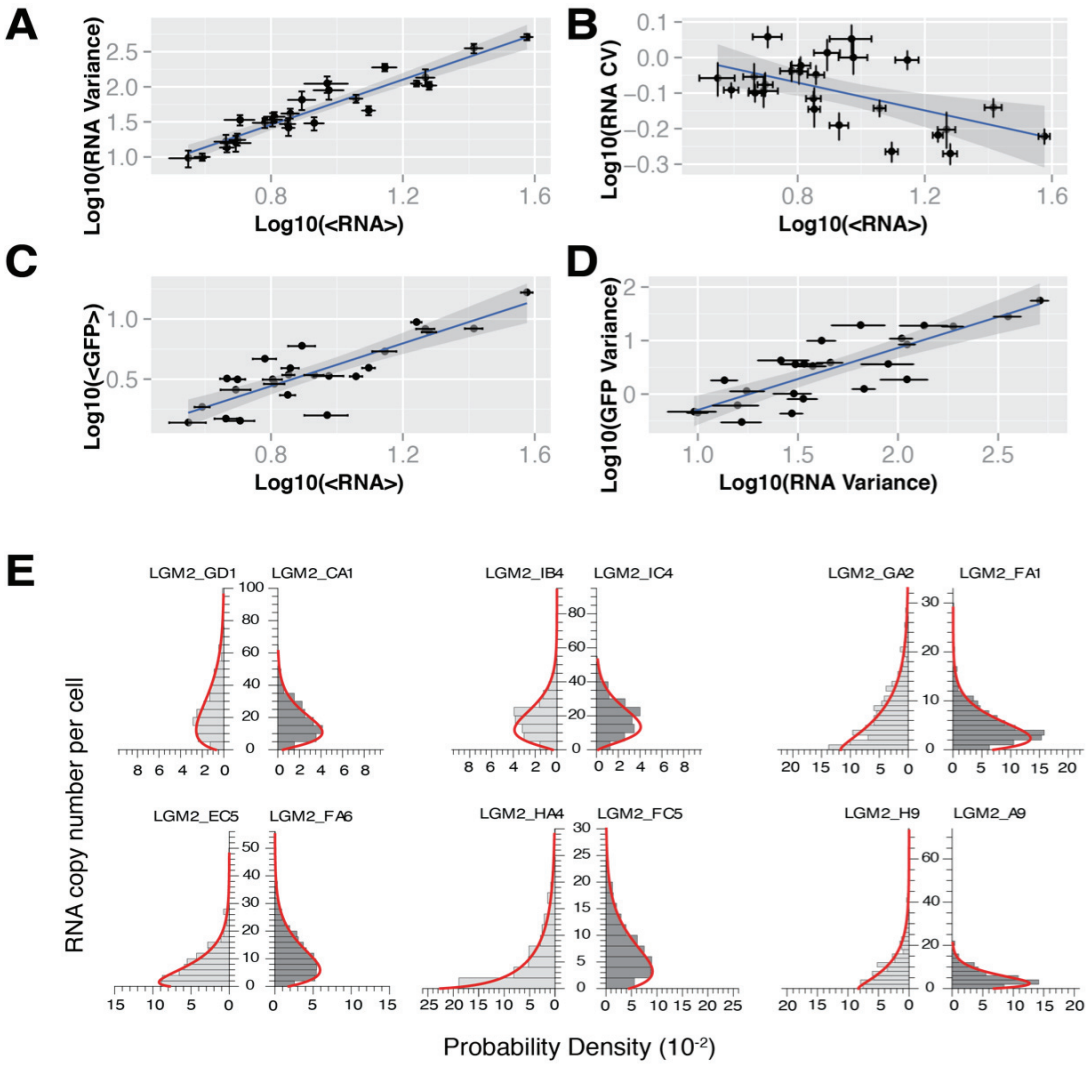


Figure 2.4|RNA distribution scaling and systematic fitting of RNA distributions reveal that HIV-1 LTR expression output is primarily controlled by intrinsic fluctuations in promoter activity (A)RNA variance correlated with mean RNA copy number:Log-Log linear regression of RNA variance as a function of RNA Mean finds a high-confidence scaling with a slope of 1.68±0.25 (R^2=0.83, r_s =0.92, p<0.001), which is intermediate to Poisson and 'input controlled' scaling. **(B)RNA mean and noise primarily independent**: Log-Log linear regression of RNA CV ('noise') as a function of mean finds an approximately uncorrelated of noise on mean expression (slope=-0.2±0.13,R^2=0.31, r_s =-0.43, p=0.03. This suggests that RNA mean and expression noise are controlled by primarily independent mechanisms. **(C,D) RNA distribution shape predominantly explains GFP distribution shape**: RNA and GFP means are highly correlated (R^2=0.79, r_s =0.86, p<0.001) (D) Similarly, RNA and GFP variances are strongly correlated (R^2=0.73, r_s =0.84, p<0.001), with RNA variance explaining the majority of variation observed in GFP variance. Together these suggest that contributions to the mean and width of the protein distributions from processes downstream of transcription are minimal. Furthermore, this strongly implies that intrinsic transcriptional dynamics are the primary determinant of protein distribution shape. Error bars on RNA moments represent

95% confidence interval estimates derived from bootstrapped smFISH distributions for each clone. r_s represents the Spearman correlation coefficient for the explanatory and response variables in each pairwise regression and p values represent support for correlation. **(E) Maximum Likelihood fitting of a two-state model to RNA distributions:** copy number distributions were determined by smFISH and automated analysis of the set of 25 selected LGM2 clones. Resulting distributions were fit to the analytical probability density function (*pdf*) of a stochastic two-state transcriptional model through maximum likelihood parameter estimation. RNA copy number distributions are depicted as paired density histograms for six representative pairs of high noise (pair left, light gray) and low noise (pair right,dark gray). Histograms are shown overlaid (red curve) with the *pdf* evaluated using best-fit model parameters for each clone

To determine how significantly RNA distribution shape controls GFP distribution shape, the RNA mean and variance of distributions for each clone were compared to their corresponding GFP moments. We find that variation in RNA mean explains variations in GFP mean well (R^2=0.79,p<0.001) while RNA variance predominantly explains GFP variance (R^2=0.73,p<0.001). This suggests that GFP distribution shape is predominantly determined by the underlying RNA distribution with only minor contributions from post-transcriptional processes.

2.7 Systematic fitting of RNA distributions finds a two-state model can describe both low and high noise clones

While scaling of the RNA distributions is highly suggestive of an underlying telegraph process, with distribution shape predominantly determined by stochastic promoter transitions between 'On' and 'Off' states, full consideration of the distribution shape is necessary to describe the underlying process. Therefore, we performed maximum-likelihood estimation (MLE) of kinetic parameters in the 'standard' two-state transcription model (Appendix A, Fig.A13), which has received considerable attention by ourselves and other recent studies[13,23,30]. While this model is an idealization of complex molecular phenomena, it has a convenient analytical solution[13,31] and has been found to parsimoniously explain observed protein and RNA distributions for synthetic and endogenous promoters in yeast and mammalian systems. MLE best-fit parameters for all 25 clones quantitatively account for the measured smFISH histograms (Fig. 2.4E and Appendix A, Fig.A14). Specifically, we find that each clone can be described by the average rate of promoter transitions to the 'On' state and the average number of transcripts produced in the 'On' state (burst-size). In particular, we find this model is sufficient to explain both the low (Fig. 4E right/dark gray in each pair) and high noise (Fig. 4E left, light gray) pairs with similar mean expression levels previously selected for further analysis. Furthermore, we find that all fitted clones are non-Poissonian with inferred burst-sizes larger than 1 in all cases. Together with our analysis of distribution shape scaling at the level of RNA and protein, this reveals that expression output is primarily controlled by intrinsic fluctuations in promoter activity with transcription occurring by a burst-like process.

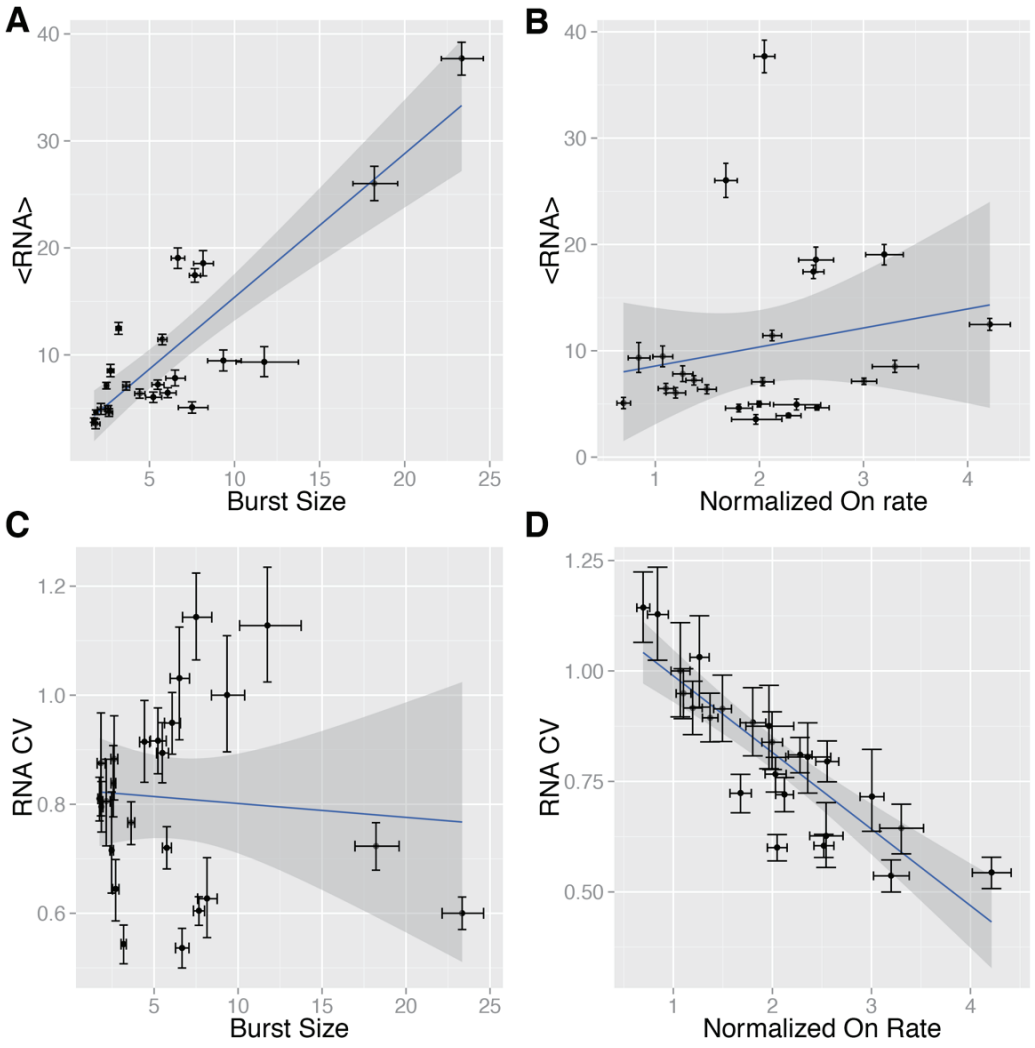


Figure 2.5|Maximum likelihood fitting of a stochastic model of transcription to smFISH distributions reveals burst size primarily accounts for expression mean while promoter transitions from off to on primarily account for expression noise. (A)**Burst size accounts for variation in mean expression:** Maximum likelihood estimation of burst size for each clone significantly accounts for a majority of the variation observed in mean RNA copy number across integration positions (slope=1.35±0.35,R^2=0.75, r_s =0.8, p<0.001). (C)**Burst size is uncorrelated from expression noise:** In contrast, burst size cannot account for the variation in expression noise (CV) observed (R^2=0.01, r_s =0.01). (B) **Promoter ON transitions are uncorrelated from expression mean:** Maximum likelihood estimation of promoter on rate normalized by the rate RNA degradation cannot account for variation in mean RNA copy number (R^2=0.05, r_s =0.2, p=0.37).(D) **Promoter On transitions accounts for expression noise variation:** Normalized on rate (by rate of RNA degradation) significantly accounts for variation in expression noise (CV) across integration positions (slope=-0.52±0.14,R^2=0.75, r_s =-0.80, p<0.001). Interestingly, this independence in cross-correlations suggests that while both burst size and promoter on rate vary across integration positions, expression mean and noise are controlled by primarily orthogonal mechanisms. Error bars on RNA moments represent 95% confidence interval estimates derived from bootstrapped smFISH distributions for each clone. Error bars on model parameters represent 95% confidence intervals estimated using 1.92 log-likelihood ratio units. r_s represents the Spearman correlation coefficient for the explanatory and response variables in each pairwise regression and p values represent support for correlation.

2.8 Differential control of expression mean and noise by burst size and rate of promoter ON transitions

We previously found through systematic fitting of GFP distributions that while both burst sizes and the rate of promoter ON transitions vary with integration position; the transcript burst size primarily modulates the mean level of GFP expression[23].Recently, parameters derived from polyclonal time-courses and assumed analytical expressions of noise scaling found that burst-size and burst-frequency (inverse of 'On' rate) vary equally with respect to mean across genomic locations[24].Given this apparent discrepancy, we sought to determine which model features predominantly relate to distribution mean and noise across genomic locations. In agreement with our previous work[23],we found that both burst-size and frequency vary across genomic locations. Systematic determination of RNA distributions with smFISH and fitting to a stochastic model of transcription revealed that burst-size primarily explains RNA mean (Fig. 2.5A, R^2=0.75, r_s=0.8) while the normalized promoter 'On' rate explains expression noise (Fig. 5D, R^2=0.75, r_s=-0.89). Consistent with our previous findings at the protein level, increased burst-sizes drive higher mean RNA expression[23]. Surprisingly, increasing the rate of promoter transitions to the 'On' state reduces expression noise in a highly monotonic relationship. While such a relationship has been theorized this is to our knowledge the first experimental demonstration.

To confirm that burst-size and burst frequency provide orthogonal explanations of expression mean and noise, respectively, we also examined the corresponding cross-correlations between burst-size and RNA noise (Fig. 5C) and burst frequency and RNA mean (Fig. 5B) and found that neither provide significant explanations (p>0.3). Furthermore, we did not find a signficiant correlation between burst-size and burst frequency (Appendix A, Fig. A15). This strongly suggests that while burst-size and burst-frequency vary across genomic locations they do not vary proportionately with mean as recently suggested[24]. Rather, burst-size explains variation in the mean while burst-frequency explains variation in expression noise. The apparent discrepancy with previous findings may be a consequence of the previous use of clones biased toward high expression (burst sizes ~100-200) or the lack of systematic parameter estimation. Together, these results significantly enhance our mechanistic understanding. Specifically, by linking burst-size to Mean and burst-frequency to CV and demonstrating the independence of these relationships we provide a concise explanation of the apparent orthogonality between mean and expression noise initially observed in both GFP and RNA moment scaling (Figs. 2 and 3).

2.9 Nucleosome Occupancy at the Transcription Start Site Regulates Gene Expression Noise and Burst Frequency

To identify the underlying molecular mechanisms regulating gene expression noise, we hypothesized that the nucleosome organization at the LTR and the chromatin density at the site of integration may play a critical role in influencing transitions between the OFF and ON promoter states, thereby regulating the width of the RNA/protein distributions. Since HIV-1 integrations sample a large diversity

of genomic locations, such a mechanistic study would also provide the first genome wide analysis in a mammalian system detailing the importance of chromosome location of endogenous genes in regulating gene expression noise.

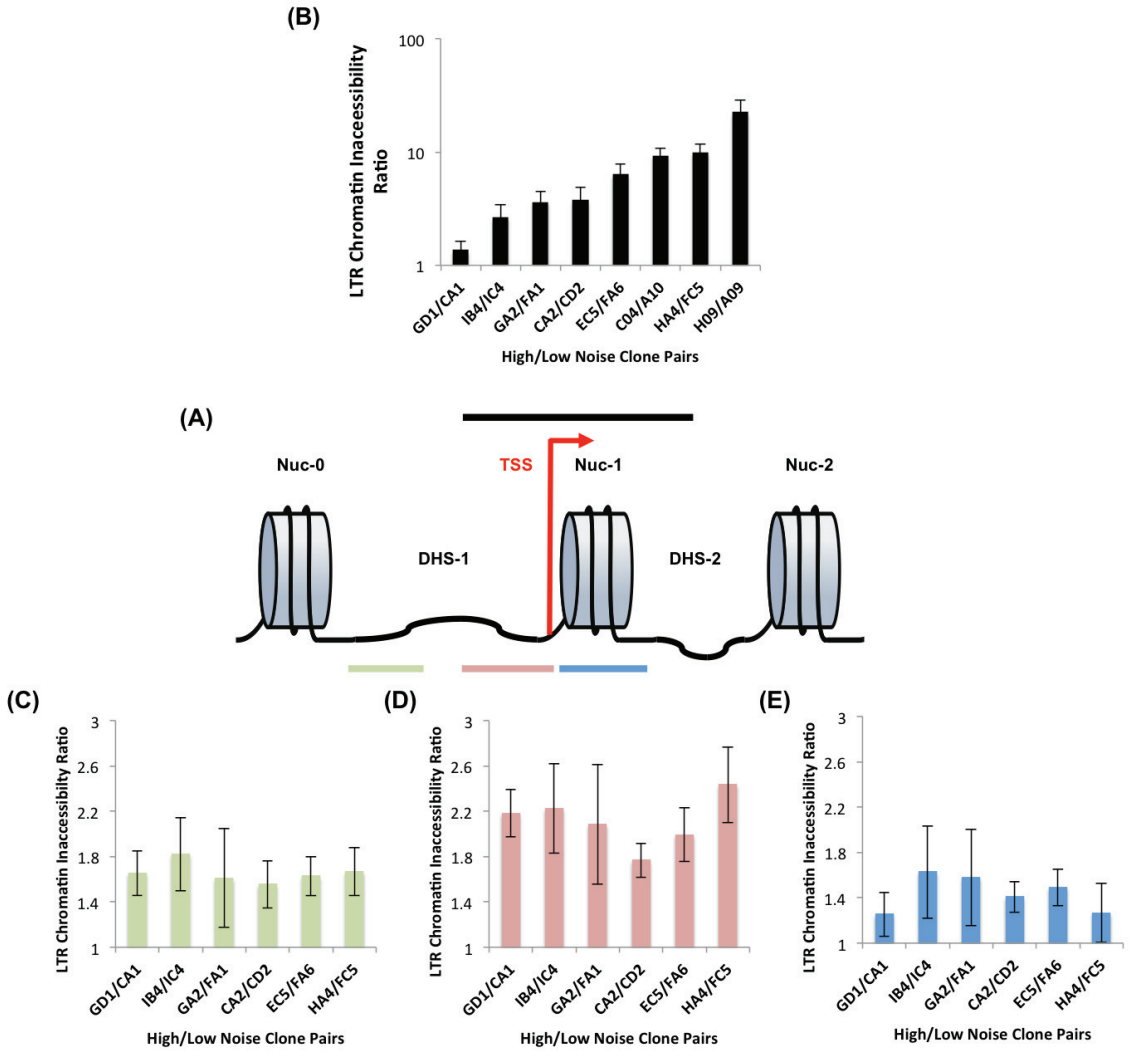


Figure 2.6|Clones with wider GFP distributions exhibit more inaccessible promoters. (A) HIV-1 LTR exhibits positioned nucleosomes: Schematic figure showing the well characterized placement of nucleosomes along the HIV-1 promoter. Nuc-1 is positioned at the transcription start site (TSS) and Nuc-0 is placed further upstream along the promoter. (B**) Significant differences in DNA accessibility around TSS between high and low noise clones**: The region highlighted in black was probed after digestion with DNase I for 16 clones. The figure plots the level of chromatin inaccessibility for pairs of clones that exhibit similar mean levels of expression. Clones that exhibit noisier gene expression also have more closed chromatin. (C), (D) and (E) **Differential DNA accessibility between high and low noise clones across three distinct sites in the LTR**: The HIV-1 promoter were probed in greater detail with the color scheme matching that shown in (A). As in (B), it was found that for similar mean levels of gene expression, noisier clones exhibited more closed chromatin along the entire length of the promoter (Ratios >1 in all cases). Interestingly, compared to the two other regions of the promoter, the nucleosome free or hypersensitive site (HSS) showed

maximum differences in chromatin inaccessibility between high and low noise clones. All qPCR was performed in triplicate and the error bars reflect the standard deviation from the mean.

HIV-1 has well-positioned nucleosomes along the entire length of the viral genome[20,32].Specifically, the LTR has two nucleosomes within the promoter, one (called Nuc-1) that is immediately adjacent to the transcription start site (TSS) and another further upstream along the promoter (called Nuc-0) (Fig. 2.6A). Such well-positioned nucleosomes are stereotypical of TATA containing promoters in yeast and mammalian cells and are poised to regulate gene expression. Therefore, to quantitatively measure the chromatin accessibility of the LTR across different genomic locations, we used DNase I sensitivity assays as described previously[33]. Initially, to assess if the chromatin accessibility at the promoter regulates gene expression noise in general, we decided to probe a large region of the promoter centered on the TSS as indicated by the black bar (Fig 2.6A,B). We quantified the chromatin accessibility over this site for 6 pairs of clones (Fig 2.4E), with each pair expressing similar mean levels of RNA/protein but different expression noise. In agreement with our hypothesis, we found that the ratio of chromatin inaccessibility between high and low noise clone pairs were greater than 1 in all cases, implying that clones that are integrated into more closed chromatin display noisier gene expression. Thus, it appeared that chromatin features within the promoter potentially regulate gene expression noise, independent of the mean level of expression.

To gain a more detailed molecular picture of the factors regulating gene expression noise, we performed DNase I sensitivity assays and probed 3 shorter regions along the length of the viral promoter[32] . First, we quantified the chromatin density around Nuc-1, a nucleosome that has previously been shown to be important for LTR mediated gene expression. Next, we probed the chromatin density around a hypersensitive site (HSS) between Nuc-1 and Nuc-0, a region that contains binding sites for critical transcription factors NF-κB and Sp1. Recently, it was shown that the presence of a nucleosome at Nuc-1 or HSS could potentially significantly influence the gene expression state of the viral promoter[32]. Finally, we probed the chromatin state at a site close to the upstream nucleosome Nuc-0 within the promoter. Analysis of the chromatin state at these three sites showed that the ratio of chromatin inaccessibility between high and low noise clone pairs is always greater than 1, further validating our hypothesis that high noise clones have more repressed chromatin (Fig. 2.6 C,D and E) along the entire length of the promoter. We found that while the ratios are greater than 1 at all three sites, the difference in the chromatin state between high and low noise clones is maximized at the HSS (Fig. 2.6D). The raw chromatin inaccessibility scores showed that while the high noise clones appear to have consistently higher values at all three sites, the HSS site for the low noise clones has significantly lower levels of chromatin than the other two sites in the promoter (Appendix A, Fig. A16). Thus low noise clones appear to have particularly open chromatin at HSS that might arise from the binding of transcription factors such as NF-κB and Sp1 to sites within this region[34] . Thus, these results show that for a given mean level of gene expression, noise in expression (or

RNA/protein distribution) is regulated by integration site-specific chromatin accessibility at the promoter (Fig. 2.6).

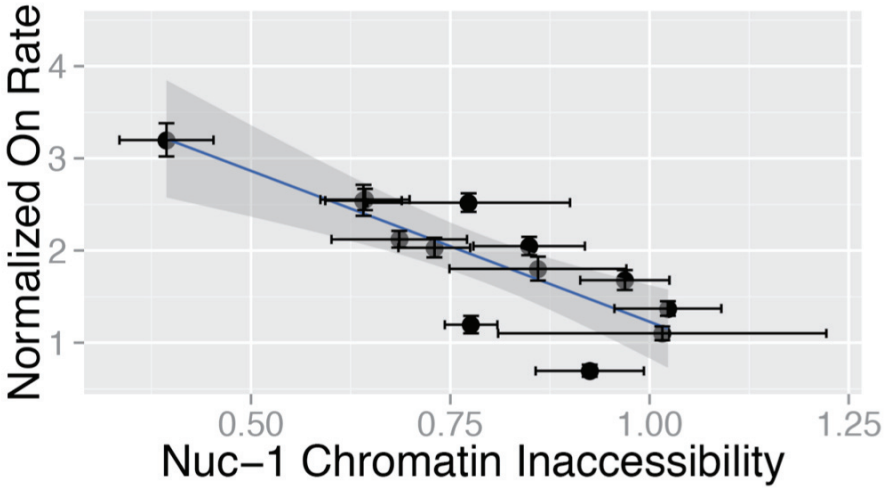


Figure 2.7|Nuc-1 occupancy can explain variation in promoter activation rates across viral integration positions. Chromatin state around the transcription start site is strongly correlated to the promoter activation rate. Unlike other regions of the HIV-1 promoter (See Supp Fig – ka vs HSS and Nuc0), chromatin inaccessibility around Nuc-1 was strongly correlated to the frequency of promoter transition from the Off to On state (slope=-1.1±0.5,R^2=0.68, r_s =-0.85, p<0.001),. Clones with more closed chromatin produced more infrequent transitions from the Off to On promoter state. All qPCR was performed in triplicate and error bars reflect the standard deviation from the mean. . Error bars on ka represent 95% confidence intervals estimated using 1.92 log-likelihood ratio units. r_s represents the Spearman correlation coefficient for the explanatory and response variables in each pairwise regression and p values represent support for correlation.

We next sought to quantify whether chromatin accessibility at the promoter can explain the primary features describing 'bursty' transcription – the transcript burst size and the rate of promoter transitions from the OFF to ON state. Specifically, we wanted to estimate the contribution of the chromatin environment at different sites within the promoter to regulating these parameters. Interestingly, we found that the rate of promoter transitions correlates very strongly to the chromatin density around the transcription start site at Nuc-1 (R^2 = 0.69, r_s=-0.85,p<0.001) (Fig. 2.7). This relationship was much weaker in the case of the HSS (R^2 = 0.49) or Nuc-0 (R^2 = 0.41) site (Appendix A, Fig.A17). Finally, to understand if some linear combination of the chromatin density around Nuc-1, HSS and Nuc-0 is a better predictor of the rate of promoter transitions, we performed principal components analysis (PCA) to establish that the rate of promoter transitions is almost inversely correlated to the chromatin density at Nuc-1 and essentially orthogonal to the HSS and Nuc-0 axis (Appendix A, Fig. A18A). Thus, the chromatin state at Nuc-1 is the best predictor of the burst frequency with increased chromatin density at the transcription start site resulting in more infrequent transitions to the ON promoter state. In contrast, the burst size appears to be uncorrelated to the chromatin state of the promoter (Appendix A, Fig. A18B).

These studies demonstrate that the chromatin environment around the LTR at different genomic locations, especially the nucleosome Nuc-1 proximal to the transcription start site, regulates the rate of promoter transitions to the ON state and the CV of the RNA/protein distributions, which to our knowledge provides the first detailed description of the molecular features regulating gene expression noise across different genomic locations in a mammalian system.

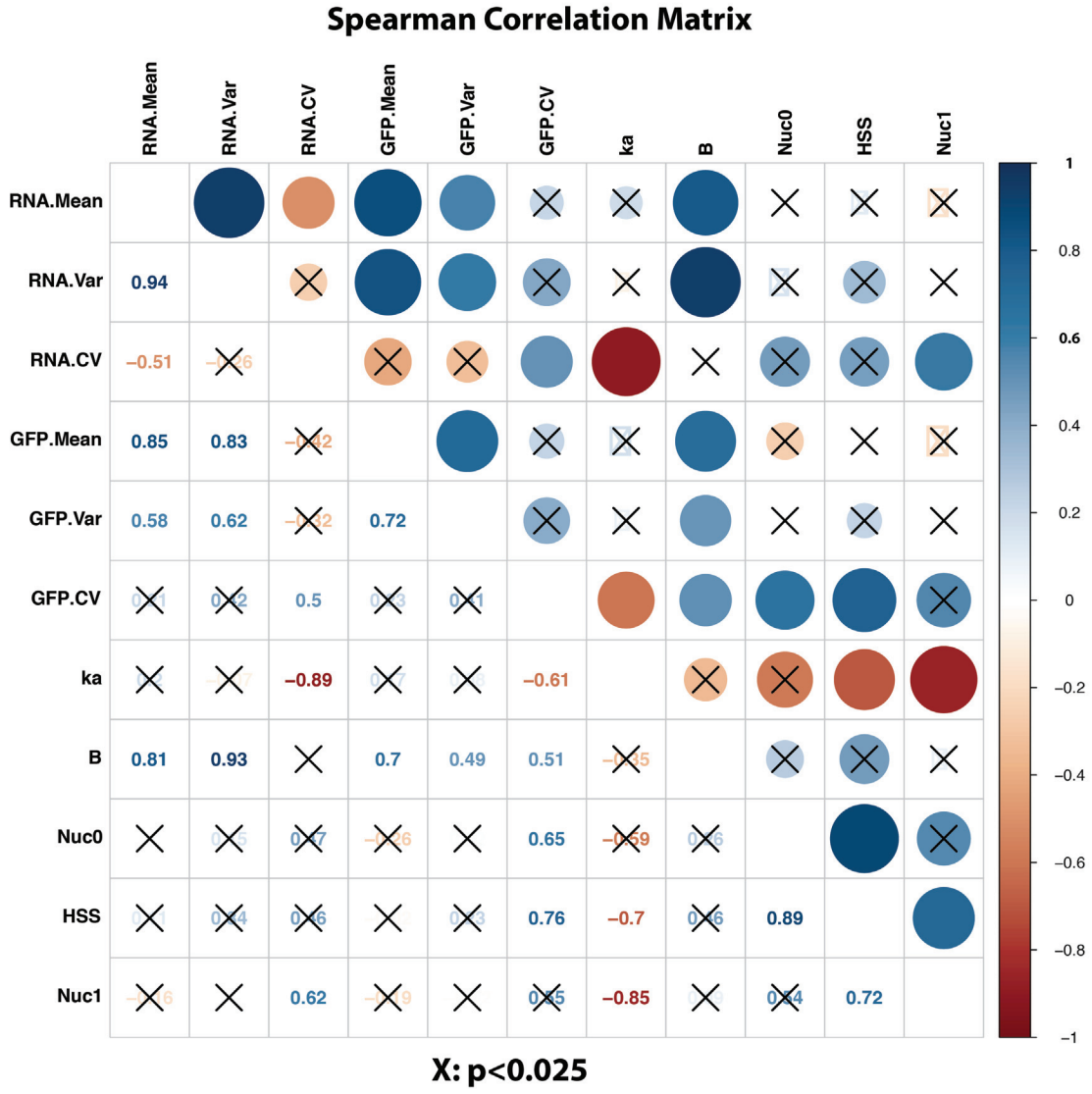


Figure 2.8|Spearman Correlation Matrix bolsters all major moment and parameter relationships: The top-half of the matrix visualizes Spearman correlation values for each row and column pair as circles where both the size and the color (legend on right) indicate the degree of correlation. The bottom-half of the matrix visualizes the Spearman correlation for each row and column pair as the numerical value and colored accordingly. In both halves of the matrix, correlations for which the p-value is greater than 0.025 are indicated by an 'X'.

2.10 Discussion

Quantitative investigation of the flow of genetic information from dynamic chromatin regulation to transcription and translation is fundamental to our understanding of genome function and cellular behavior. While the nature of noisy stochastic gene expression has received intense interest in both synthetic and natural systems[2,13,35,36] the molecular features underlying noisy expression phenotypes have remained elusive. Here, using a dual-reporter lentiviral model system, we analyzed over 200 clonal populations by flow cytometry and over 15,000 single cells from 25 clonal populations by smFISH. This unparalleled scale of analysis, combined with measurements of promoter chromatin occupancy, enabled three key insights. First, that expression mean and CV are uncorrelated across integration positions. Second, that this observed large-scale independence between expression mean and CV can be explained by inferred orthogonality between transcriptional burst size and promoter on rate. Lastly, that chromatin density can explain the promoter activation rate but does not provide an explanation for burst-size. In particular, in this study we systematically demonstrate that promoter chromatin density can explain promoter activation rate, which in turn can explain the CV or noise of the RNA distribution. This suggests that local chromatin density may modulate the frequency of transcriptional bursting and thereby tune expression noise.

These findings are in contrast to recent studies suggesting equal modulation of burst size and frequency with respect to expression mean across genomic sites[24], and a limited role for chromatin on bursting dynamics[18] .Equal modulation of burst size and frequency by Dar *et al*[24] were inferred from assumed expressions relating the CV of clustered single-cell trajectories, which are arbitrarily clustered on time-course endpoints, in polyclonal populations to model parameters. The CV of these clusters is only demonstrated to be *qualitatively* similar to a small sample of isogenic clones. Therefore, it is unclear how the clustering procedure may obscure the estimation of kinetic parameters across genomic locations. Furthermore, reported burst sizes (>100 for most clones), are not reflective of the full range of LTR driven expression (Skupsky *et al*[23], and this study). Together, these features significantly question claims about the role of genomic location on noise characteristics and limit the generalization of their findings. Sutter *et al* strongly rejects a role for the influence of chromatin on bursting dynamics. However, this claim rests on the use of TSA on very limited clonal data, a highly cytotoxic drug that *globally* remodels chromatin and has unknown molecular effects on the specific genomic locations studied. Further, clone-to-clone variability with a few clones showing that TSA influences its noise characteristics make the claims of Sutter *et al* inconclusive. Another recent study using small molecule perturbations such as TSA and 5-AzaC suffers similar drawbacks. Their conclusions about how the chromatin environment regulates gene expression noise are biased by large-scale chromatin remodeling and toxic effects that are induced by these drugs. It is likely that both promoter architecture and the local chromatin environment couple to give rise to observed expression distributions and inferred transcription dynamics. Further studies are needed to systematically study the nature of this coupling, which may

influence the evolution of promoters and the non-random positioning of genes within the genomes[37].

Our assertion that the chromatin density modulates promoter activation rate can be intuitively understood from genome-wide studies in Drosophila quantifying nucleosome dynamics showing that different regions of the genome could have variable nucleosome turnover rates[38] thereby directly affecting promoter switching rates. Further, we provide quantitative evidence that the position of nucleosomes relate to the TSS influence the noise characteristics from the promoter to different extents. Chromatin density around the TSS plays the most important role in regulating the promoter ON rate. Additionally, our finding that burst size couldn't be explained by chromatin density may suggest that other molecular features such as transcription factor or RNA polymerase occupancy underlie burst size regulation. Further investigation of the molecular underpinnings of burst size modulation will further deepen our molecular interpretation of bursting dynamics

.

Nucleosomes have been known to regulate transcription by setting a threshold for initiating transcription[33,39]. This work shows that in mammalian systems nucleosomes might also be important to fine-tune transcription. Just as miRNA's have been implicated in imparting robustness[40], nucleosomes may perform a similar function. Therefore, this study quantitatively establishes another function of nucleosomes in regulating noise, in addition to its known functions in regulating expression levels and imparting cellular memory.

2.11 Materials and Methods

2.11.1 Viral cloning

To facilitate single molecule detection of transcripts, the M2 array from pGEM-M2-32x[13] was cloned as a directional SalI-XhoI fragment into the single XhoI site of HIV CLG[22] to generate HIV CLGM2 used in this study.

2.11.2 Cell Culture

Jurkat cells, used for creating clonal LGM2 cell-lines and HEK293T cells, used for packaging virus were cultured in RPMI 1640 and Isocove's DMEM, respectively. Cells were maintained at 37^0C and 5% CO_2 with the cell-media supplemented with 10% fetal bovine serum (FBS) and 100 U/mL of penicillin-streptomycin.

2.11.3 Viral Harvesting, Titering and Infections

To package the LGM2 construct, HEK293T cells were transfected with 10μg of the plasmid along with the helper plasmids pMDLg/pRRE, pcDNA3 IVS VSV-G and pRSV-Rev as described previously[41] and harvested. Harvested lentivirus was concentrated by ultracentrifugation to yield between 10^7 and 10^8 infectious units/ml. To titer, 10^5 Jurkat cells per well, were infected with a range of concentrated virus doses and six days post infection, gene expression was stimulated with 20 ng/ml tumor necrosis factor-α (TNF-α,

Sigma-Aldrich). After stimulation for 18 hours, GFP expression was measured by flow cytometry, and titering curves were constructed by determining the percentages of cells that exhibited GFP fluorescence greater than background levels.

2.11.4 Clone Generation

Assuming a Poisson distribution, the well with 5% GFP infected cells was selected for expansion. This corresponds with a low MOI of ~0.05 and as previously demonstrated[22] ensures at most one integration event per infected cell within the population. The selected population was expanded for 6 days and stimulated with TNF-α as described above. Approximately 0.1*10^6 cells were sorted from the GFP^+ population on a DAKO-Cytomation MoFlo Sorter. The resulting population, which represents a polyclonal ensemble of single-integration clones, was expanded for 7 days. Single cell clones were sorted from a wide (~1.5 log10(RFU)) gate distinct from background fluorescence into multi-well plates and cultured for 3-4 weeks to facilitate expansion.

2.11.5 Flow Cytometry

GFP expression of expanded clonal cultures was assessed by flow cytometry using a Beckman Coulter FC500 analytical cytometer. Multiple reads over the course of a week were obtained to ensure that measured fluorescence represented stationary distributions of gene expression. Flow cytometry data was processed using Bioconductor (Gentleman 2004) packages in custom R scripts as described in the Supplementary Information.

2.11.6 Cell fixation

For each sample, 1-2 million cells washed once in PBS and allowed to adhere to 0.01mg/ml poly-L-Lysine (Sigma-Aldrich P5899) coated chambered #1 cover-glass (Nunc Lab-Tek #155380). Cells were fixed with 4% Formaldehyde (Sigma-Aldrich 252549) solution in PBS. Slides with fixed cells were washed once with PBS and 70% cold Ethanol added to permeabalize cellular membranes. Slides were stored at 4C until hybridization.

2.11.7 In situ hybridization

Single molecule labeling of LGM2 transcripts was performed as previously described (Raj 2006) with the following exceptions: a) 35% Formamide was used in the hybridization buffer b) A shorter, TAMRA dual end labeled probe (BioSearch Inc. , 5'-GTCGATCAGCTGGCTGGTGCTCTTCGTCCACAAAC-3') was used and c) the hybridization reaction was carried out for 16hrs. at 30^0C.Prior to imaging, slides were washed twice with 2X SSC 35% Formamide. Cell nuclei were counterstained with 0.5μg/ml DAPI. Just prior to imaging, samples were mounted under a coverlsip with aqueous oxygen scavenging buffer system[28] (GLOX).

2.11.8 Imaging and Deconvolution

84 randomly selected fields were imaged per sample. To facilitate imaging of the entire cell volume, each field was imaged in three channels (GFP, DAPI, TAMRA) 90 z-axial slices at 0.2 μm spacing with a 100X oil immersion objective on a Deltavision Core (API) widefield deconvolution microscope equipped with standard fluorescent filters. Image stacks were iteratively deconvolved using an experimentally determined point-spread function (PSF) with Huygens Core 3.3 (SVI) running on a small Linux cluster. Optimal deconvolution parameters were empirically determined to yield the best subjective image quality. Custom Tcl scripts running under Silicon Grid Engine (RefXXXX) were used to manage job creation and scheduling.

2.11.9 Image Processing

Deconvolved image stacks were processed using custom software developed in MATLAB (Mathworks Inc.) using the DIPImage Toolbox (Quantitative Imaging Group, TU Delft). Under minimal user intervention, the software automatically segments cells and smFISH objects within single cells. A detailed protocol for the processing pipeline is contained in Appendix A. Source code is available on request.

2.11.10 Model Fitting

RNA distributions were fit using Maximum Likelihood Estimation (MLE) of model parameters using the full analytical solution to the two-state stochastic gene expression model[31] MLE was implemented using custom code in Mathematica 8 (Wolfram Inc.) as numerical minimization over the negative log-likelihood function defined over the *pdf* given the observed RNA counts the rate of RNA degradation set to our experimentally determined rate and transcription rate assumed to be constant across integration sites as previously discussed[23]. In this manner, the effective fit parameters are the burst frequency and burst size. All model parameters and summary statistics for each clone can be found in Table A2.

2.11.11 Statistical Analysis

95% confidence intervals on descriptive statistics of RNA distributions were estimated from the 2.5% and 97.5% quantiles of bootstrapped copy number counts per cell. 95% confidence intervals on fit parameters were estimated from the log-likelihood function assuming asymptotic normality of the estimates. These analyses were performed in Mathematica 8. All regression and correlation analysis was performed in R using using the *lm* and *rcorr* functions . The regression p-values of all primary inferences reported in this study fall below an α of 0.05. Distribution clustering and principal component analysis were performed in MATLAB (Mathworks Inc). A full summary of regression best-fit slopes with associated p-values and Spearman correlation coefficients are contained in Appendix A.

2.11.12 DNase I sensitivity assay

The assay was performed using the EpiQ Chromatin Analysis Kit (Bio-Rad) as previously described[33,42]. Briefly, $2.5x10^5$ cells were either treated with DNase I or left

untreated for 1 hour at 37^0C. After quenching the reaction, DNA was extracted and quantified by qPCR using the EpiQ Chromatin SYBR Supermix (Bio-Rad). The following primers were used to quantify the chromatin density at the HIV promoter:
LTRfor (5'-GGACTTTCCGCTGGGGACTTTCCAGGG-3')
LTRrev (5'-GCGCGCTTCAGCAAGCCGAGTCCTGCGTCGAG-3')
Nuc-1for (5'- AGCTCTCTGGCTAACTAGGG-3')
Nuc-1rev (5'- AAAGGGTCTGAGGGATCTCTAG-3')
HSSfor (5'- GGGACTTTCCGCTGGGGAC-3')
HSSrev (5'- CCCAGTACAGGCAAAAAGCAGC-3')
Nuc-0for (5'- GAGCCTGCATGGGATGG-3')
Nuc-0rev (5'- CTCCGGATGCAGCTCTC-3')
The qPCR results were normalized by the chromatin density at the hemoglobin promoter using the primers:
hHBBfor (5'-AAGCCAGTGCCAGAAGAGCCAAGGA-3')
hHBBrev (5'-CCCACAGGGCAGTAACGGCAGACTT-3')
All qPCR was performed in triplicate with melt curves to ensure product specificity.

2.11.13 mRNA Extraction and RT- qPCR

To determine the half-life of transcripts, a polyclonal LGM2 population was stimulated by α-Amanitin and total RNA was extracted from cells at different time points using Trizol (Invitrogen). RNA was also extracted from unstimulated cells at these time points. LGM2 and β-Actin mRNA was quantified by RT-qPCR using the single step Quantitect SYBR Green RT-PCR kit (Qiagen). All qPCR was performed in triplicate. For additional details and primers, see Appendix A.

2.12 References

1. Spudich, J. L. & Koshland, D. E. Non-genetic individuality: chance in the single cell. *Nature* **262**, 467–471 (1976).
2. Arkin, A., Ross, J. & McAdams, H. H. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. *Genetics* **149**, 1633–1648 (1998).
3. Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L. & Leibler, S. Bacterial persistence as a phenotypic switch. *Science (New York, N.Y.)* **305**, 1622–5 (2004).
4. Weinberger, L., Burnett, J., Toettcher, J., Arkin, A. & Schaffer, D. Stochastic Gene Expression in a Lentiviral Positive-Feedback Loop: HIV-1 Tat Fluctuations Drive Phenotypic Diversity. *Cell* **122**, 169–182 (2005).
5. Wernet, M. F. *et al.* Stochastic spineless expression creates the retinal mosaic for colour vision. *Nature* **440**, 174–180 (2006).
6. Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E. & Huang, S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature* **453**, 544–7 (2008).
7. Spencer, S. L., Gaudet, S., Albeck, J. G., Burke, J. M. & Sorger, P. K. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature* **459**, 428–32 (2009).

8. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic Gene Expression in a Single Cell. *Science* **297**, 1183–1186 (2002).
9. Blake, W. J., K**?**rn, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. *Nature* **422**, 633–637 (2003).
10. Raser, J. M. Control of Stochasticity in Eukaryotic Gene Expression. *Science (New York, NY)* **304**, 1811–1814 (2004).
11. Paulsson, J. Summing up the noise in gene networks. *Nature* **427**, 415–8 (2004).
12. Kaern, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics* **6**, 451–464 (2005).
13. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA Synthesis in Mammalian Cells. *PLoS Biology* **4**, e309 (2006).
14. Bar-Even, A. *et al.* Noise in protein expression scales with natural protein abundance. *Nature genetics* **38**, 636–643 (2006).
15. Mao, C. *et al.* Quantitative analysis of the transcription control mechanism. *Molecular systems biology* **6**, 431 (2010).
16. Rando, O. J. & Chang, H. Y. Genome-wide views of chromatin structure. *Annual review of biochemistry* **78**, 245–71 (2009).
17. Court, F. *et al.* Modulated contact frequencies at gene-rich loci support a statistical helix model for mammalian chromatin organization. *Genome biology* **12**, R42 (2011).
18. Suter, D. M. *et al.* Mammalian genes are transcribed with widely different bursting kinetics. *Science (New York, N.Y.)* **332**, 472–4 (2011).
19. Pereira, L. A., Bentley, K., Peeters, A., Churchill, M. J. & Deacon, N. J. A compilation of cellular transcription factor interactions with the HIV-1 LTR promoter. *Nucleic Acids Res* **28**, 663–668 (2000).
20. Verdin, E., Paras, P. & Van Lint, C. Chromatin disruption in the promoter of human immunodeficiency virus type 1 during transcriptional activation. *The EMBO journal* **12**, 3249–59 (1993).
21. Weinberger, L. *et al.* Expression noise and acetylation profiles distinguish HDAC functions. *Molecular cell* **47**, 193–202 (2012).
22. Weinberger, L. S., Burnett, J. C., Toettcher, J. E., Arkin, A. P. & Schaffer, D. V Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell* **122**, 169–182 (2005).
23. Skupsky, R., Burnett, J. C., Foley, J. E., Schaffer, D. V & Arkin, A. P. HIV promoter integration site primarily modulates transcriptional burst size rather than frequency. *PLoS computational biology* **6**, 14 (2010).
24. Dar, R. D. *et al.* Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 17454–9 (2012).
25. Singh, A., Razooky, B., Cox, C. D., Simpson, M. L. & Weinberger, L. S. Transcriptional bursting from the HIV-1 promoter is a significant source of stochastic noise in HIV-1 gene expression. *Biophysical journal* **98**, L32–4 (2010).

26. Newman, J. R. S. *et al.* Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441**, 840–846 (2006).
27. Trcek, T. *et al.* Single-mRNA counting using fluorescent in situ hybridization in budding yeast. *Nature protocols* **7**, 408–19 (2012).
28. Raj, A., Van den Bogaard, P., Rifkin, S. A., Van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* **5**, 877–879 (2008).
29. Rifkin, S. A. Identifying fluorescently labeled single molecules in image stacks using machine learning. *Methods in molecular biology (Clifton, N.J.)* **772**, 329–48 (2011).
30. Zenklusen, D., Larson, D. R. & Singer, R. H. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural & Molecular Biology* **15**, 1263–1271 (2008).
31. Peccoud, J. & Ycart, B. Markovian Modeling of Gene-Product Synthesis. *Theoretical Population Biology* **48**, 222–234 (1995).
32. Rafati, H. *et al.* Repressive LTR nucleosome positioning by the BAF complex is required for HIV latency. *PLoS biology* **9**, e1001206 (2011).
33. Miller-Jensen, K. *et al.* Chromatin accessibility at the HIV LTR promoter sets a threshold for NF-κB mediated viral gene expression. *Integrative biology : quantitative biosciences from nano to macro* **4**, 661–71 (2012).
34. Burnett, J. C., Miller-Jensen, K., Shah, P. S., Arkin, A. P. & Schaffer, D. V Control of stochastic gene expression by host factors at the HIV promoter. *PLoS Pathog* **5**, e1000260 (2009).
35. Thattai, M. & Van Oudenaarden, A. Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci U S A* **98**, 8614–8619 (2001).
36. Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-time kinetics of gene activity in individual bacteria. *Cell* **123**, 1025–36 (2005).
37. Batada, N. N. & Hurst, L. D. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat Genet* **39**, 945–949 (2007).
38. Deal, R. B., Henikoff, J. G. & Henikoff, S. Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones. *Science (New York, N.Y.)* **328**, 1161–4 (2010).
39. Lam, F. H., Steger, D. J. & O'Shea, E. K. Chromatin decouples promoter threshold from dynamic range. *Nature* **453**, 246–50 (2008).
40. Ebert, M. S. & Sharp, P. A. Roles for microRNAs in conferring robustness to biological processes. *Cell* **149**, 515–24 (2012).
41. Dull, T. *et al.* A third-generation lentivirus vector with a conditional packaging system. *Journal of virology* **72**, 8463–71 (1998).
42. Dey, S. S. *et al.* Mutual information analysis reveals coevolving residues in Tat that compensate for two distinct functions in HIV-1 gene expression. *The Journal of biological chemistry* **287**, 7945–55 (2012).

Chapter 3: An informatics driven approach to the generation of highly diverse combinatoric mammalian enhancers

3.1 Introduction

Eukaryotic gene transcription is regulated by a number of complex properties, including the identity, orientation, and multiplicity of transcription factor binding sites (TFBS) proximal and distal to a RNA polymerase II (RNAPII) core promoter. The complexity of the genetic programs[1,2] observed in eukaryotes derive in part from the substantial logic encoded in the *cis* acting architecture of promoters and related enhancer elements[3,4]. In particular, the need to regulate genes in space and time[5] within multicellular organisms has driven the evolution[6,7] of large enhancer elements[8] composed of ensembles of DNA motifs that direct the binding of transcription factors. These transcription factors orchestrate gene expression, and the precise sequence of the motifs[9,10]to which they bind can determine their binding energy[11], conformation, and co-binding partners[12,13]. Furthermore, the organization of TFBS into *cis* regulatory modules (CRM)[14–16], their organization into chromatin, and the genetic and epigenetic environment of the surrounding chromosomal region all contribute to gene expression output[17–20].

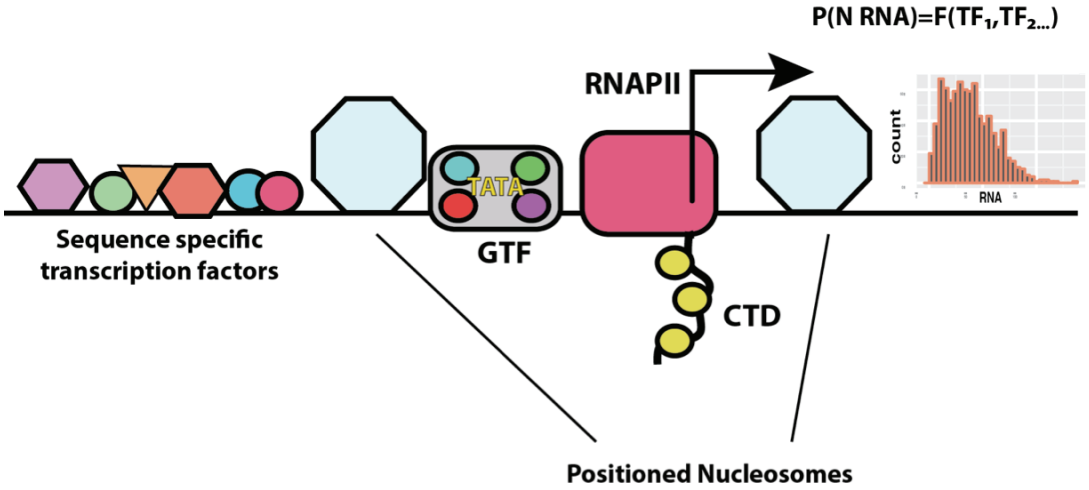


Figure 3.1|The general structure of eukaryotic promoters: Complex genetic programs are orchestrated through the logic encoded in large promoter proximal enhancer elements. These enhancers are composed of DNA sequence motifs that direct the binding of sequence specific transcription factors. Sequence specific transcription factors interact with general transcription factors (GTF) such as TFIID and TATA binding protein. These transcription factors catalyze post-translational modification of positioned nucleosomes and the RNAPII complex. For example, such modifications may include histone tail lysine acetylation or RNAPII carboxyl terminal domain phosphorylation. The precise ensemble and binding dynamics of these transcription factors determines the resulting transcriptional output; a probability function expressing the expected number of RNA per cell over time.

Transcription is initiated and elongation regulated through the dynamic interaction between core components of the RNAPII complex and enhancer bound

transcription factors[21–24]. These promoter bound factors can post-transcriptionally modify RNAPII and histone proteins[25], and remodel the local chromatin environment to increase accessibility by other factors or to allow RNAPII to transit the body of the gene. These modifications of the local chromatin environment and gene bound factors serve to regulate the timing and kinetics of transcription[26,27]. By integrating multiple regulatory inputs from the both the physiological state of the cell and the local genomic context, promoters thereby serve as integrative devices that determine the temporal probability of mRNA production and ultimately protein expression. Recent efforts to decipher the logic of such promoter transfer functions have begun to decipher the fundamental metrics of transcriptional control[16,28–31]. However, *cis* acting TFBS comprise a poorly understood grammar that syntactically specifies complex gene expression programs. Knowledge of this grammar would enable understanding the natural function of eukaryotic promoters and enhancer elements and would by extension aid in the synthetic design of promoters for synthetic biology and gene therapy applications. In contrast to the current limitations in the genetic manipulation available for endogenous promoters, retroviral promoters are highly genetically tractable. Furthermore, due to core promoter and enahcer elements that are highly similar to endogenous mammalian promoters, retroviruses represent an attractive model system to the study of transcriptional control by promoter sequence elements

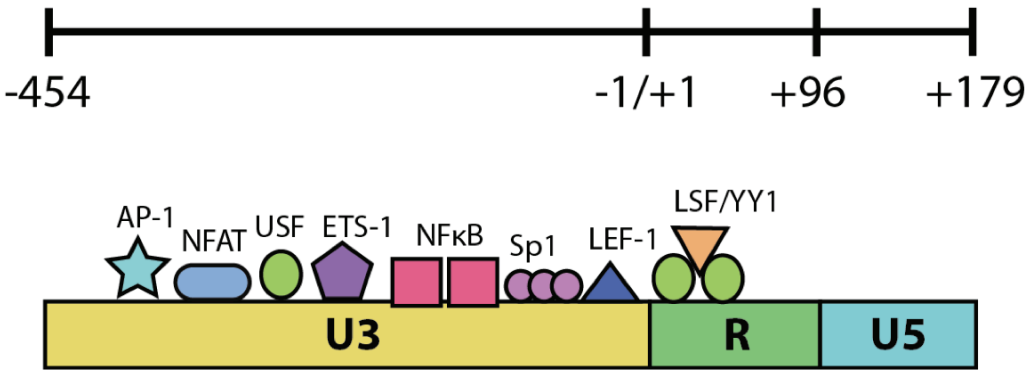


Figure 3.2| Enhancer structure of the HIV LTR: The HIV LTR is subdivided into three regions: U3, R and U5. The majority of the *cis* acting logic of the LTR is contained in U3, which extends approximately 450 nucleotides 5' of the transcriptional start site (coordinates relative to HIV HXB2). U3 is composed of numerous and frequently repeated transcriptional factor binding motifs such as NFkB and Sp1. How this specific architecture of this *cis* acting logic

Fundamental aspects of the retroviral lifecycle include the reverse transcription of two genomic RNA into a full length cDNA copy and the integration of the resulting proviral cDNA at semi-random[32] sites within the host chromosomes. During reverse transcription, recombination due to reverse transcriptase strand transfer, as well as short insertions and deletions due to strand slippage alter the *cis* acting architecture of the LTR, can occur. Once integrated, the provirus must engage a genetic program to generate the viral gene products necessary for viral production and further replication. HIV encodes a terminal promoter element termed the Long-

Terminal-Repeat (LTR), which integrates multiple inputs from the host cell environment to produce viral genes via an unknown transfer function. Consisting of a core TATA based promoter (U3/R boundary) that facilitates assembly of the RNA Polymerase II transcriptional apparatus, and a large 5' proximal enhancer (U3 Region) that directs binding of multiple host cell transcription factors, the overall structure of the LTR (Fig. 3.2) is highly similar to endogenous mammalian promoters. HIV circulates globally as 11 major subtypes (A1, A2, B, C, D, F1, F2, G, H, J and K) and several major recombinant forms (e.g. A/G and B/F). Interestingly, the U3 regions of these subtype LTRs exhibit considerable differences in the composition, ordering, and multiplicity of TFBS[33] (Fig. 3.3). Furthermore, within individual patients and across geographic distributions of a single subtype, numerous polymorphisms in LTR TFBS are observed[34,35]. Recent work suggests that these subtype architectural differences and TFBS polymorphisms quantitatively alter LTR mediated expression[33,36]. However, the precise relationship between LTR *cis* acting architecture and gene expression output is unknown.

Due to the lack of efficient tools for mammalian genome engineering, even with the advance of sequence-specific nucleases endogenous mammalian promoters are challenging to manipulate *in situ*. Therefore, the HIV LTR presents an ideal model system to systematically study promoter structure-function relationships by introducing variability into the promoter *ex situ* and then introducing the genome into a cell via viral infection. Here, we develop a synthetic approach based on the random non-homologous recombination of short dsDNA elements to form long synthetic enhancers that can be subjected to experimental selection in order to broadly infer the metrics of promoter structure-function relationships.

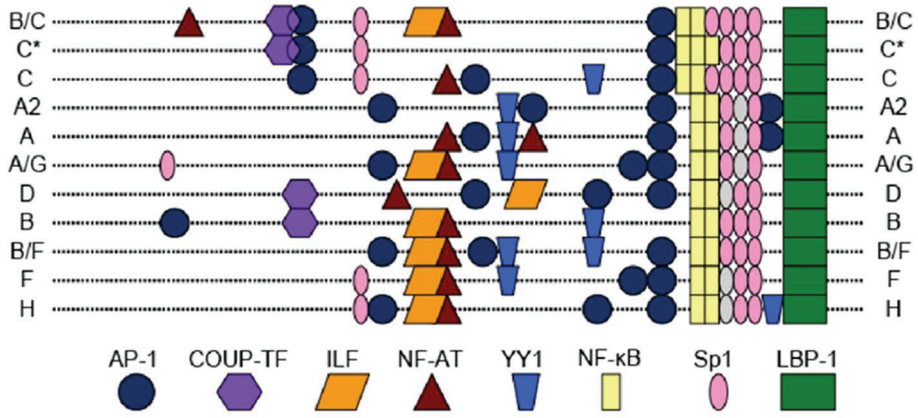


Figure 3.3| LTR architectural differences amongst major HIV subtypes: Globally, HIV circulates as 11 major subtypes and a number of stable recombinant types. These subtypes roughly correspond to geographic areas and are likely reflective of regional transmission patterns, host haplotypes or other host genotype dependent factors. The LTR of these subtypes differ markedly in the precise composition, ordering and multiplicity of transcription factor bind sites. However, the precise relationship between the LTR *cis* acting architecture and gene expression output is poorly understood. Figure adapted from Burnett *et al* 2010[33].

3.2 Traditional DNA shuffling of HIV subtype LTR

As an initial attempt to generate LTR architectural diversity, we employed traditional DNA shuffling, which is based on PCR based recombination of DNAse I treated parental DNA species[37]. DNA shuffling has been widely employed to evolve novel functions in biomolecules through directed evolution[38,39]. Eleven parental subtype LTR obtained as molecular clones from the NIH AIDS reagent program were amplified using PCR. Amplicons were subsequently digested with DNAse I for amounts of time ranging from 5 to 30 minutes, yielding fragments of decreasing average size with respect to digestion time. Unfortunately, multiple attempts to assemble these fragments and ultimately amplify full length recombinant LTR species according to established DNA shuffling protocols[39] failed to yield viable products. The failure of such a well-established technique is likely due to the short length of the LTR (~700bp) and the resulting small size of the majority of DNAse fragments (typically <50bp). Another potential reason for the observed failure of fragment assembly may be attributed to the significant sequence diversity (nucleotide homology <70%) between the parental subtype species, which may have limited overlap between parental fragments and ultimately precluded PCR based assembly.

3.3 Nucleotide exchange and excision technology (NExT) DNA shuffling of subtype LTR

DNAse I is a highly non-specific DNA nuclease that generates broad fragment size distributions and affords little specific control over the ultimate points of recombination in traditional DNA shuffling. To investigate whether uncertain or short fragment sizes led to the observed failure of traditional DNAse based DNA shuffling, we subsequently attempted to perform DNA shuffling via nucleotide exchange and excision technology[40]. Briefly, this method relies on the doping of PCR amplicons of desired parental species with Uridine triphosphate (dUTP). The degree of doping can be controlled through the molar ratio of Uridine to Thymidine triphosphates in an otherwise standard PCR reaction. Uridine incorporations serve as specific points of fragmentation through enzymatic uracil base excision with uracil-DNA-glycosylase and subsequent phosphate backbone cleavage with piperidine. Through modulation of the dUTP:dTTP ratio, this method achieves deterministic and repeatable fragment size distributions. These fragments can be used directly in traditional homology based assembly and re-amplification.

As in 3.2, 11 parental LTR representing major HIV subtypes and recombinant forms were subjected to levels of dUTP doping ranging from 5% to 50%. This yielded the expected differences in fragment size distributions. Unfortunately, no level of dUTP doping and resulting fragmentation subsequently resulted in successful assembly and re-amplification of full-length recombinant species (Fig. 3.4). Therefore, it is likely that small fragment size is not the primary reason for the observed failure of DNA shuffling. We conjecture that the lack of significant

homology between the parental subtype LTR was simply beyond the limits of homology based recombination techniques.

3.4 A synthetic approach to systematically explore the LTR structure function relationship

Synthetic biology is based on the premise of predictable forward engineering of biological systems. Within this paradigm, recent efforts have focused on building predictable biological systems from well-characterized 'parts' that represent control elements such as core promoter elements, initiators, terminators, and feedback elements. Complementary to this forward-engineering paradigm is the generation of large libraries of biomolecule diversity through mutagenic and combinatoric techniques that for instance generate highly diverse libraries of RNA apatamers or enzymes. These diverse libraries can then be subjected to genetic selection or directed evolution, which involves application of numerous rounds of functional selection to identify novel biomolecule properties of interest. Recent efforts have combined part based forward-engineering with techniques from directed evolution by subjecting large combinatoric libraries to functional selections. In particular, elegant experiments in both yeast and mammalian cells have performed selections on combinatoric promoter libraries to achieve promoters of different noise characteristics or levels of expression. Importantly, these studies involved relatively limited rearrangements of larger promoter fragments that reflected *a priori* hypotheses of promoter structure-function relationships. Furthermore, these studies did not specifically address the relationship between TFBS sequence, multiplicity, or ordering and gene expression output. Given the high degree of complexity of *cis* acting logic observed in eukaryotic promoters, we sought to develop a synthetic approach capable of generating very large combinatoric promoter libraries. In particular, we sought to treat individual transcription factor binding motifs as individual parts. In this manner, we hypothesized that the roles in determining patterns of gene expression of motif sequence and their architectural arrangement into *cis* regulatory modules could be broadly inferred.

3.5 Identification of regulatory sites within the HIV LTR and empirical determination of their position weight matrices

In adopting a synthetic part based approach, our first goal was to establish a set of parts that corresponded to both well studied and poorly studied regulatory sites within the LTR. Furthermore, we sought to include both positive and negative regulatory elements that are observed within naturally circulating LTR. Due to their central role in contributing to LTR mediated gene expression, sites most proximal to the core TATA promoter, such as NFkB and Sp1, have been extensively studied. However, large regions such as the Negative Regulatory Element (NRE) are weakly characterized and have uncertain regulatory roles. Furthermore, during the course of an infection, in addition to its primary target, the CD4+ T cell, HIV is known to infect other immune cell lineages such as macrophages and at very late stages after

the establishment of AIDS, cells of the central nervous system[41]. Therefore, it is likely that the LTR may contain motifs or CRM that support gene expression in non-T cell types. Each cell type presents a unique cell biology context with different levels of expression of transcription factors, which has a global impact on gene expression. Our experimental system only involved T cell lines (Jurkat and SupT1), therefore we sought to ascertain motifs that were primarily involved within the physiological state of T cells.

Motif Name	IUPAC sequence
NFkB-1	GACTTTCCR
NFkB-2	GGMCKTTCCRG
Sp1-1	GGGACGTGGYBA
Sp1-2	TGGGCGGGAC
Sp1-3	RGGYGHGGYTT
Ap1	TKMMWSA
LSF/YY1	NNDCCATNN
USF	DCCATNRWCACVKRGC
LEF-1	WYTACAARRACTGCA
RBF-2	ACTGMTRAC
C/EBP	GCAYKDMRWCAYVK
E2F-1	TTTCCRCTWR
p53	GGYGTGGYYT
core-NRE	CAYKDMRWCACVTRGC
COUP	GGRCCAGGGRYYAGATWYCCACTGWSHTTTGGRT
NFAT	RRDGARGTGARRARGMCGGGMAATCRAAGGAGAR
CBF	NNTGGGAANN
Spacer6	ATGCGT
Spacer10	ATGCGTAGGC
NMF-1	SMSGAMAGAGAAGTGTT
NMF-2	RRWRTGGAAGTTGACACC
NMF-3	YMGCTAGCAYKHMRWCACAVTGGCCC
EBS	MWRCATCCGGAGT
Ebox	KYKRGCYABGT
RCS	RTYTGAGCCYGGGAGCTCTCTGG

Table 3.1| Degenerate parts representing diversity of LTR motifs identified through bioinformatic analysis

Using an extensive compendium of known LTR TFBS[42] as well as a search of more recent literature, we developed a set of parts (Table 3.1) representing both annotated motifs (e.g. p53, NFAT, COUP-TF) as well as unannotated regions from the NRE that had evidence of regulatory roles in T cells. Previous work found differential roles in recruiting transcription factors for the two NFkB sites and three

Sp1 sites within the HIV Subtype B LTR[36]. Interestingly, while these multiple sites are thought to bind the same transcription factor, the sequences observed over each motif are markedly different. Therefore, in light of these observations each site was treated as a unique motif. In order to limit potential steric hindrance from factor binding to densely concatenated motifs and to allow for spacing between CRM in synthetic enhancers, two spacer parts were constructed. 'Neutral' 6 bp and 10 bp spacers were developed by iteratively swapping nucleotides within a candidate sequence and searching for homology with known eukaryotic TFBS using the JASPAR[43] database. The lengths were chosen to be on the order of 1 or 2 binding motifs and roughly correspond to a half and full turn of the B-form DNA helix. We hypothesized that such geometric effects in addition to the simple linear arrangement of motifs could impact the assembly of larger multi-factor complexes.

While TFBS are frequently represented by their consensus sequence, such representations are highly idealized and rarely exist as such within genomes. Rather, transcription factors bind to families of related motifs, and the precise sequence of the motif may determine the contribution of the binding event to the temporal pattern of transcription[9,10]. Therefore, in order to explicitly and systematically study the relationship between the precise sequence of TFBS and gene expression output, we developed a bioinformatics based approach to develop parts. The 2010 reference HIV alignment curated by the Los Alamos National Laboratory (LANL, http://www.hiv.lanl.gov/) was used to estimate position weight matrices (PWM) over each included motif. This alignment consists of 1492 full-length molecular clones representing all major HIV subtypes and major recombinant forms. Using the LANL database web interface, the alignment was cropped for each individual motif.

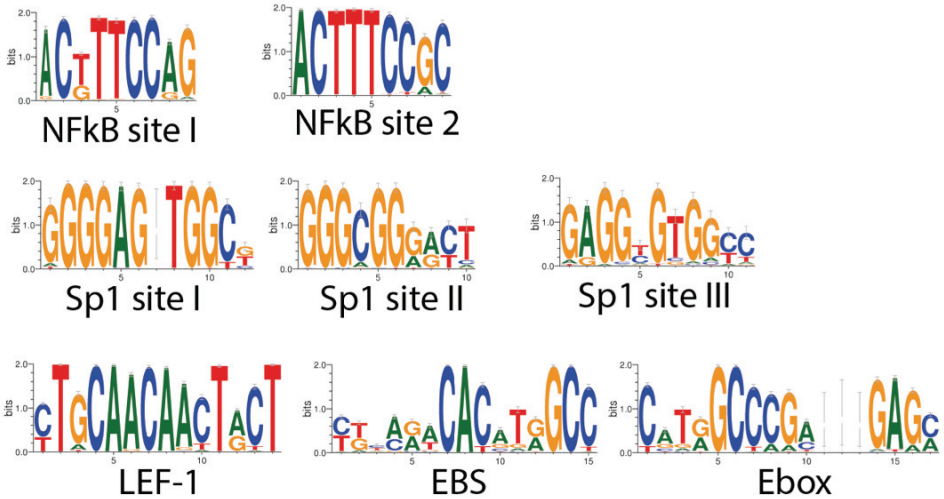


Figure 3.4| Representative Logo visualizations of LTR transcription factor binding motifs: The 2010 Los Alamos National Laboratory (LANL) reference HIV alignment of 1492 sequences representing all major subtypes and major recombinant forms was used to generate estimates of LTR position weight matrices (PWM) for motifs that were identified through a literature search. The alignment was cropped to coordinates specifying each individual motif and a logo visualization generated using WebLogo3[44]. Interestingly, motifs observed for the two NFkB sites and three Sp1 sites display different nucleotide preferences over the motif. Such intramotif differences for the same transcription factor may for instance differentially regulate binding kinetics or the assembly of larger multi-factor complexes and therefore underlie differences in expression output.

PWM are a frequentist representation of the probability of seeing a given nucleotide (A,C,G,T) at each position within a motif. This probability is expressed using the bit entropy metric, which ranges from 0 to 2 bits and can be expressed graphically as a logo (Fig. 3.4). To limit the overall degeneracy allowed at each position and limit inclusions of extreme sequence outliers, an arbitrary cutoff of 0.1 bits was used to determine what nucleotides were allowed in each part. Following this inclusion cutoff, degenerate sequence representations of the motif were output using the IUPAC alphabet for nucleotides. For example, if G or T is allowed at a given position within a motif then this is represented as K. In this manner, each part represents a family of related motifs that captures much of the diversity observed over within globally circulating LTR. Synthetic DNA oligonucleotides corresponding to these IUPAC representations can be directly obtained from commercial oligonucleotide manufacturers. An important caveat of this approach is that during solid phase oligo synthesis all the allowed nucleotides at a given position are injected during a given incorporation step. This results in an uncertain distribution of inclusion of the allowed nucleotides. In the best case, the incorporation would be uniform with respect to the number of nucleotides present during the incorporation step. Therefore, this approach likely does not perfectly represent the observed nucleotide frequencies over each position within the motif. However, despite this caveat, our approach represents a novel approach to capturing such diversity in a comprehensive manner.

3.6 Generation of highly degenerate dsDNA representations of motif PWM

Once a strategy for generating parts that capture TFBS diversity observed in key LTR motifs was established, our second goal was to devise a method to experimentally generate physical dsDNA parts that represent each motif family. Initially, we attempted to anneal degenerate sense and antisense oligos that corresponded to each motif. However, this failed to yield clean dsDNA products. It is likely that was due to the high degree of sequence diversity and lack of guarantee that each sense sequence has a corresponding antisense partner. Subsequently, to ensure that each degenerate oligo within each part would be reliably duplexed we devised a strategy based on primer extension that takes advantage of the unique off-target digestion properties of Type IIs restriction enzymes. Specifically, each part is specified as a longer oligonucleotide consisting of the desired degenerate motif, the Type IIs restriction site for MlyI, and a constant region that facilitates primer extension from a common primer. MlyI generates a blunt end six nucleotides 5' of its GACTC recognition site. This property was highly instrumental in the generation of blunt dsDNA representations of the motif PWM that are free of any constant sequence. To facilitate capture and removal of the constant sequence used for primer annealing and extension, the common primer used was 5' biotinylated. Due to the necessity of performing enzymatic reactions at 37°C, some motifs were padded on either side with random nucleotides (N) to increase their melting

temperature above 37°C. While unavoidable given currently available enzymes, this has the benefit of generating additional diversity that potentially samples additional TFBS not explicitly included in the design.

The common constant primer region facilitates processing of the parts as a pooled ensemble (Fig. 3.5). Specifically, the part oligonucleotides are first combined in equimolar quantity and extended using Vent polymerase under PCR conditions with a saturating (1.25X equimolar) amount of 5' biotinylated extension primer. The resulting dsDNA parts are then digested with MlyI, and the constant portion of the digestion product removed from the reaction mixture using Streptavidin coupled paramagnetic beads. This process results in short (10-25 bp) dsDNA parts that capture the observed sequence diversity over included LTR motifs. For further information, full protocol details can be referenced in Appendix B.

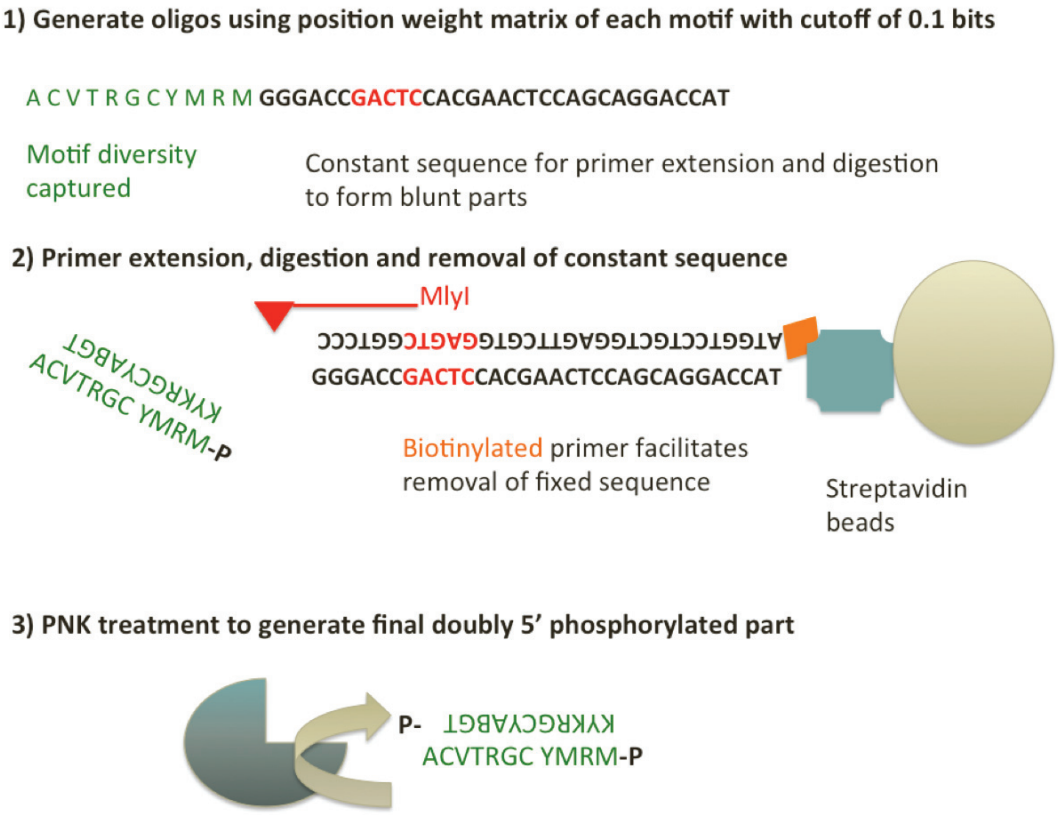


Figure 3.5| Experimental workflow to generate short dsDNA parts that capture sequence diversity observed in transcription factor binding sites: Each motif is represented as a degenerate oligonucleotide using the IUPAC nucleic acid alphabet. To facilitate generation of duplex parts from such degenerate oligos, the base motif oligo is joined with a constant region that allows annealing of a 5' biotinylated common primer to all parts and subsequent primer extension. The biotin allows for removal of the constant portion using Streptavidin paramagnetic beads following digestion of the duplex oligonucleotide with the Type IIs restriction enzyme Mly I. Mly I cuts adjacent to its recognition site and generates a blunt end, allowing recovery of the duplexed generate oligo representing each motif. To facilitate efficient ligation, the hemiphosphorylated dsDNA parts are treated with T4 Polynucleotide Kinase (PNK).

3.7 Combinatoric assembly of long DNA polymers through random recombination (NRR) of short double stranded oligonucleotide parts

Following the succesful generation of diverse part families representing LTR motif sequence diversity, our next goal was to assemble these parts into longer combinatoric DNA polymers that resemble the TFBS ensembles observed in natural enhancers. Previous approaches[45,46] to generate combinatoric promoters have relied on fixed overhangs that introduce a large amount of repetitive background sequence that is dissimilar to the architecture of natural promoters. Additionally, previous studies have used larger blocks[47] consisting of either multiple TFBS, which limits the potential to infer the contribution of single sites to gene expression output, or large repeats[48] (10X) of single TFBS, which lack similarity to patterns observed in eukaryotic promoters. Therefore, to address the limitations of previous approaches, we sought a method that could efficiently ligate small dsDNA elements into larger combinatoric polymers.

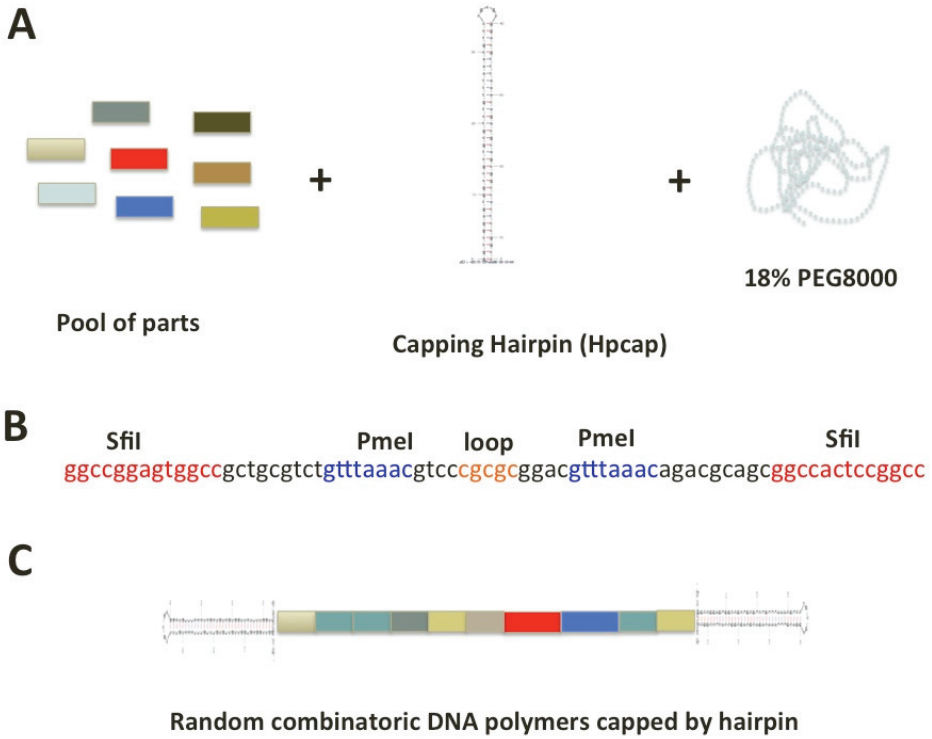


Figure 3.6| Non-homologous random recombination (NRR) of blunt dsDNA parts to generate synthetic enhancers: (A) We adapted a previously devised method of recombining blunt DNAse I generated fragments to assemble random combinatoric DNA polymers of various sizes. Efficient blunt-end ligation of short dsDNA elements is facilitated by the presence of a high volume percentage of PEG8000, which serves as a volume occupying agent and highly increases the effective concentration of ligatable part ends through molecular crowding. To achieve products of a desired size range, a hairpin is included in the assembly reaction to terminate polymer growth upon random incorporation of the hairpin into a growing polymer chain. The product size distribution is determined by the molar ration of hairpin to parts. **(B)** Hairpin sequence used in NRR contains a

PmeI site for product linearization by removal of the terminal loop sequence prior to PCR , and a SfiI site to facilitate cloning of products **(C)** The NRR assembly process generates random combinatoric polymers composed of input parts and capped on each end by a hairpin.

Toward this aim, we identified a previously developed[49] method for the non-homologous random recombination (NRR) of blunted DNAse I generated fragments from parental DNA aptamer sequences. This method was conceived of to address the need for a high degree of homology between parental species in traditional DNA shuffling. The homology requirement inherently limits the size of the sequence space that can be explored and prevents generation and analysis of sequences that have significantly different architectures than the parental species. The universe of naturally occurring promoter sequences reflects the selective pressures encountered over the course of sequence evolution and does not necessarily reflect all possible configurations and architectures that would be functional in an organism or cell type of interest.

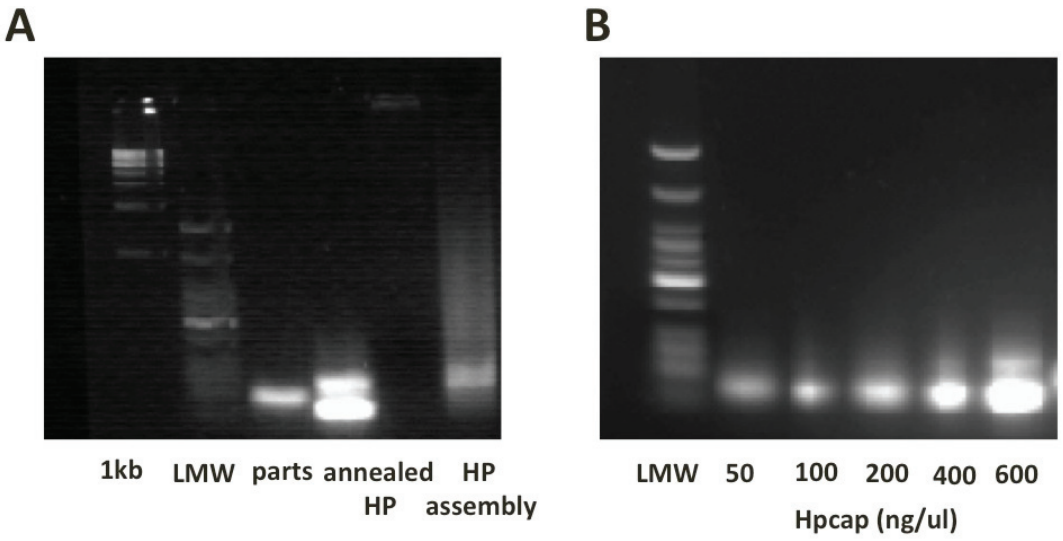


Figure 3.7| Kinetic-trapping of hairpin yields a single product and prevents self-assembly of hairpin element: (A) Use of a capping hairpin in NRR requires that the hairpin be present in its lowest energy intramolecular stem-loop configuration and not a duplex. However, slow annealing of the HPcap hairpin generates both the stem-loop structure and the higher molecular weight duplex (lane 4: 'annealed HP'). This duplex specie is problematic because it is capable of self-assembling into long repetitive polymer under NRR conditions. (B) To overcome this, we employed kinetic-trapping of the stem-loop structure by boiling in the HP oligo in the presence of 100mM NaCl, and immediately quenching the reaction vessel on ice. We find that this efficiently generates the desired single stem-loop product at concentrations up to 400 ng/ul and strongly limits self-assembly. Products were visualized on a 2% Agarose gel and '1kb' and 'LMW' represent 1kb and low molecular weight molecular weight ladders.

By using NRR with our set of TFBS parts we hypothesized that we could generate a very large sequence space (theoretically >10^12 unique sequences) that

widely samples the universe of functional enhancer architectures. NRR is itself based on a much older[50] report of efficient ligation of short and blunt dsDNA oligos under conditions of molecular crowding with high volume percentages (>15% v/v) of Poly-Ethylene-Glycol (PEG). NRR achieves product size selection through the introduction of a capping hairpin element into the ligation reaction. When a hairpin is incorporated at the end of a growing multimer, further ligation is terminated. The probability of incorporation of a given part or hairpin is dependent on its molarity relative to other species in the reaction. Therefore, by adjusting the molar ratio of capping hairpin to motif parts the average probability of hairpin incorporation is modulated, which in turn determines the average length of DNA polymer that is assembled.

During our adaptation of NRR, we encountered a technical problem with the folding and use of a hairpin element not reported by the original authors. For the hairpin element to efficiently terminate polymer chain growth, it must be present as a stable stem-loop structure in the ligation reaction. A common method of folding hairpin sequences into their lowest-energy stem-loop structure is to slowly cool the oligo in an annealing protocol. However, we found that slow annealing generated both the desired stem-loop structure as well as a bimolecular duplex structure (Fig. 3.7A). This duplex structure is problematic because it is capable of self-assembly into long polymers, which limits its utility in functioning as a chain terminator. Furthermore, HP self-assembly would pollute the resulting sequence library with repetitive HP polymers. To address this problem, we employed kinetic trapping of the stem-loop structure (Fig, 3.7B). Kinetic trapping avoids the two species observed under equilibrium thermodynamics through a rapid temperature transition from well above the critical folding temperature (the temperature at which base pairing begins to occur along the folding trajectory) to well below it. Under these circumstances, the intramolecular folding reaction is kinetically favored over the bimolecular duplex formation. In the presence of 100 mM salt, we found that kinetic trapping efficiently generated the desired stem-loop structure at HPCap oligo concentrations up to 400 ng/ul. At higher concentrations, it is likely that equilibrium exchange between the stemp-loop and duplex structures is driven through increased mass action kinetics.

An attractive feature of using NRR to generate synthetic enhancers is the potential to explore sequence length as a parameter influencing gene expression output. Recent studies suggest that the length of enhancers and the inherent relationship between length regulatory complexity may influence promoter robustness to genetic and environmental perturbations[51] or more generally the evolvability of gene expression[8,52,53]. Therefore, we sought to explore the relationship between enhancer length and gene expression output by selecting a range of combinatoric enhancer elements from 200 bp to 2 kbp, representing on average 20 to 200 assembled parts. To yield a high density of assembled products within this desired size range, the ratio of hairpin to parts was empirically determined through titration of the hairpin while keeping the input parts constant (Fig. 3.8). Due to the high degree of diversity of the parts both in terms nucleotide composition and length, the exact molarity in the reaction is uncertain. Therefore, we empirically determined the range of mass percentages that yielded products

over the desired size range. Specifically, we found that with 1.5 µg of input parts and relative mass percentages of hairpin to parts ranging from 13% to 53% (Fig. 3.8) yielded sufficient density of sequences from 200 bp to 2 kbp to facilitate recovery by gel extraction.

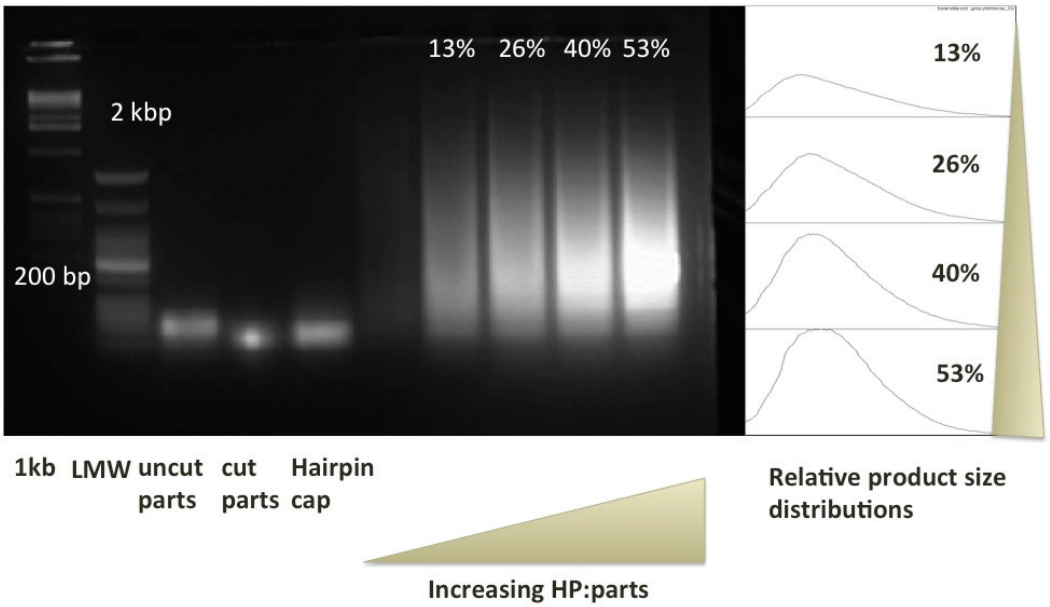


Figure 3.8| Efficient assembly of long DNA polymers from short blunt DNA parts using NRR: NRR was performed with pooled, processed parts representing LTR diversity. With 18% PEG, the weight percentage of capping hairpin (lane 5) to parts (lane 4) was empirically determined to yield product sizes ranging from 200 bp to 2 kbp. Increasing weight percentages of hairpin shift the resulting assembled product size distribution toward shorter average lengths. We found that mass percentages of hairpin as a fraction of parts ranging from 13% to 53% yielded products of the desired lengths. Products were visualized on a 2% agarose gel with '1kb' and 'LMW' representing 1 kbp and low molecular weight ladders.

3.8 Strategies for multi-template amplification of combinatoric products

Owing to the small quantities of assembled products recovered (<100 ng) from gel extraction and subsequent enzymatic processing, we found an amplification step to be required. The stem of the hairpin element that caps each end of an assembled product functions as a fixed sequence adapter and facilitates amplification with a single primer. While the original authors of NRR indicated use of traditional PCR in amplifying products, they do not specifically comment on the preservation of the size distribution or the generation of undesirable chimeric PCR products. Unfortunately, we found that traditional PCR amplification of our assembled products resulted primarily in the generation of chimeric products that are significantly longer than the 200 bp to 2 kbp size range we selected for. This is likely due to the inherent homology between assembly products arising from incorporation of degenerate sequences representing the same input part. Due to this

homology, individual species within the reaction can likely prime each other and thereby generate chimeric products[54]. Even in the absence of such chimeric product formation, traditional PCR kinetically favors amplification of shorter templates and can exhibit sequence bias[55].

To address the limitations of traditional PCR in amplifying our combinatoric sequence library, we tested the application of emulsion PCR[56,57].Emulsion PCR (ePCR) is an elegant technique that partitions single to few templates and standard PCR reagents as nanoliter aqueous vesicles within an oil and surfactant phase. In this manner, each template molecule is amplified in isolation. This approach has been shown to both preserve multi-template size distributions and limit the formation of chimeric products[56]. To facilitate stochastic partitioning of template molecules within the $\sim 10^9$ reaction vessels per 50 µl reaction, 1.66 fmol of template DNA is used. The considerable heterogeneity in template sizes and molecular weights of our combinatoric assembly products made estimates of template molarity approximate at best. However, assuming an average assembled product size of 500 bp with an average molecular weight of 0.33 ng/fmol, we empirically tested input template amounts ranging from 0.1 ng to 1 ng in 0.1 ng increments (~0.33 fmol-3 fmol). Unfortunately, performing emulsion PCR across this range of template quantities failed to generate appreciable amplification. The reason for the observed failure is uncertain. However, slight deviations in the preparation of the emulsion may yield dramatically different vessel size distributions and impact reagent partitioning. Such deviations could arise during the drop wise introduction of the aqueous phase into the oil phase. Specifically, the speed of mixing or specific vessel and micro sized stir bar used during drop-wise addition may impact the resulting emulsion quality. Alternatively, recent work has reported that not all combinations of BSA, which is used as a stabilization reagent, and polymerase yield amplification[57]. These issues highlight an overall challenge in the adaptation of emulsion PCR for applications beyond those currently demonstrated.

3.9 Modification of viral vector and library cloning

To enable systematic study of the LTR structure-function relationship we sought to introduce our combinatoric sequence library into a model replication competent HIV vector (pRCHIVsLTR[58], Fig. 3.9A). Toward this aim, we modified the LTR of the vector to allow for introduction of a synthetic enhancer sequence while maintaining a constant minimal TATA promoter (-58 to -1 relative to transcriptional start site). Furthermore, we sought to generate two restriction sites that would involve the introduction of minimal mutations to the LTR. We identified a site near the 5' end the U3 region (centered around -459) and another centered on SpI site I (-58). The placement of these sites permits removal of the majority of U3 *cis* acting logic, thereby allowing its replacement with our combinatoric enhancer library (Fig. 3.9B). In the absence of a viable amplification strategy, we have made multiple attempts to directly clone combinatoric products into the *Sfi* I modified vector. These attempts have resulted in estimated library diversities of 10^4 to 5*10^4. While this indicates the overall feasibility of our approach, we do not consider this to be sufficient to systematically study the structure-function relationship following

the initial selections imposed by viral packaging. We hypothesize that initial selection due to the constraints of packaging will yield at least a 10-fold reduction in library size. Therefore, we continue to seek a strategy that can capture >10^6 unique sequences.

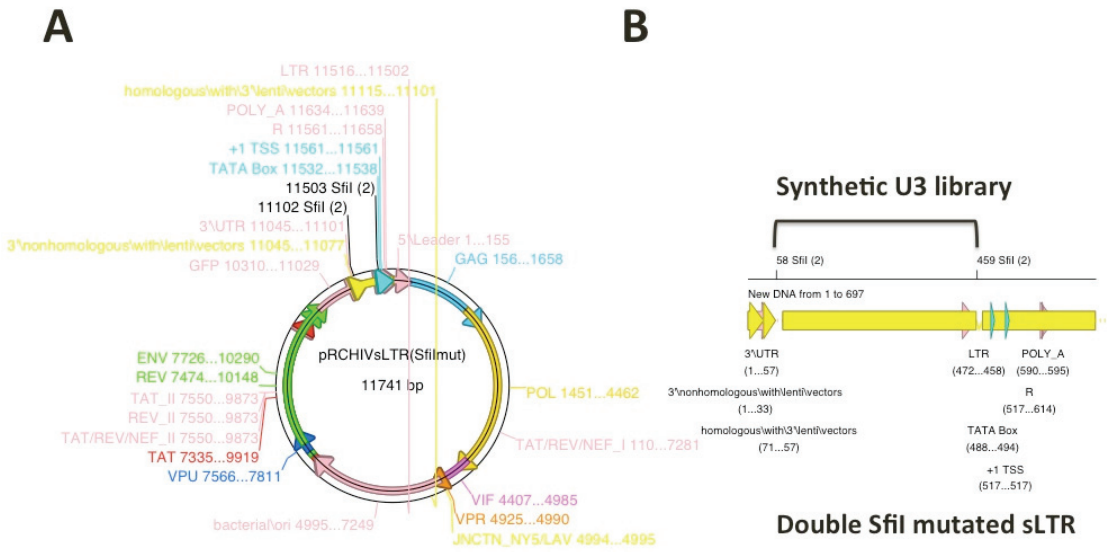


Figure 3.9| Modification of replication competent HIV vector for cloning of combinatoric enhancer library: (A) To facilitate cloning of combinatoric U3, SfiI sites were introduced at two sites within the single LTR of pRCHIVsLTR, a replication competent model HIV vector. (B) Schematic of locations of the two SfiI sites relative to the TATA box and the transcriptional start site (TSS). This cloning strategy preserves a core TATA promoter in U3 and leaves R and U5 regions unchanged.

3.10 Computational annotation of combinatoric sequences

We envisioned that downstream analysis of combinatoric sequences following experimental selections would require software to efficiently identify and annotate both rationally designed and novel motifs within recovered sequences. Toward this aim, we developed a modular sequence annotation framework in Python that takes as input a list of IUPAC motifs corresponding to the degenerate input parts, automatically annotates combinatoric sequences, and displays annotations in a graphical user interface (GUI). Expected position weight matrices of the motifs in input parts are computed assuming uniform incorporation of nucleotides at each position allowed by the specified IUPAC character. To allow for identification and annotation of additional motifs not specifically included as parts, additional motifs culled from curated databases such as JASPAR can be included. Such motifs could for instance arise from oligo synthesis errors, novel motifs arising from the concatenation of input parts, or from the random nucleotide padding of some input parts. Due to its best in class motif identification accuracy and precision, we chose to use FIMO[59] from the MEME[60] suite as the core motif finding algorithm.

However, owing to the modular nature of our framework, recent algorithms such as BAMBI[61] can easily be used. Within the annotation GUI, each input part is mapped to a unique color, which permits quick assessment of the composition and qualitative architecture of annotated sequences (Fig. 3.10). In addition to graphical output, the identity and position of identified motifs is output as human and machine-readable text (Table 3.2).

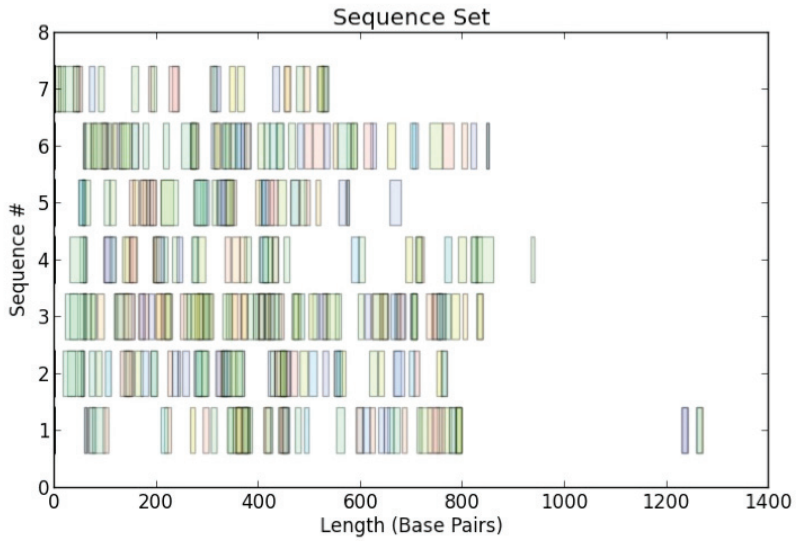


Figure 3.10| Computational annotation of combinatoric ehancers reveals sequences of different lengths and motif compositions: A prototype graphical user interface was developed that maps motifs identified with FIMO to colors in a user defined colorspace. Computational annotatation of the Sanger sequencing of seven experimentally assembled enhancers indicates that our novel assembly method yields enhancers of differing length and *cis* acting architecture

Table 3.2: Representative annotation of combinatoric U3

* prime (') indicates that the motif is flipped relative to the orientation of the sequence. A ~1.2 kbp combinatoric enhancer was sequenced using Sanger sequencing and computationally annotated.

>Synthetic_LTR_U23.seq---1273

NFkB1-2'	1262	1272	GGCCCTTCCCG
EBS'	1232	1244	AAACATTCGGAGA
NMF3'	342	367	CCGCTATCAGGACATAGCGTTGGCTA
RBF2'	357	371	CGGACCGCTATCAGG
Ebox2	775	788	AGTACAGCTGCGCA
Ap1'	473	485	GCATGCCCGACGG
Spacer	636	647	CAAGCGTATGCA
LSF-YY1'	447	459	GTGACCCATGGCG
CoreNRE	67	82	CCTTTCGCCAGCTGGC
C/EBP	653	668	CGCATTGCATCAGCCA
Spacer'	647	658	AATGCGACGGCT
NMF3	358	383	CTGATAGCGGTCCGCCACACCCAGCC
USF'	350	365	GCTATCAGGACATAGC
RBF2'	216	230	AGTTCTTCTGACTGT
RBF2	354	368	TGTCCTGATAGCGGT
p53'	368	379	GGGTGTGGCGGA
NMF-2'	762	779	GTACTCGACGTTGTCACT
CBF	785	796	CGCAAGGAACGC

Ebox2'	375	388	TGGCCGGCTGGGTG
NMF3	717	742	GCCCCGGCACTTCGCCCCAATAGCAG
CBF	1259	1270	GAGCGGGAAGGG
Sp1-2'	711	722	CGGGGCAGGATC
Sp1-3'	787	799	GGGGCGTTCCTTG
CBF'	372	383	GGCTGGGTGTGG
p53'	440	451	TGGCGATGCCTG
NFkB1-2'	788	798	GGGCGTTCCTT
CoreNRE'	605	620	CATTCGACCACCAAGC
Sp1-2'	370	381	CTGGGTGTGGCG
LEF1	554	568	TCGACAAGACCGGCT
Ebox'	68	79	AGCTGGCGAAAG
Sp1-1	293	304	AGGAAGCGGTCA
CBF'	749	760	AAGCGGGAAGGG
LSF-YY1	444	456	CATCGCCATGGGT
RCS'	357	379	GGGTGTGGCGGACCGCTATCAGG
CoreNRE'	412	427	CATGGTGGAAAATGGC
NFkB1-2'	267	277	CGCCGCTCCCG
Ap1'	666	678	GTATCCATCATGG
RCS'	743	765	CACTGAAGCGGGAAGGGGACTGG
NMF3'	100	75	CAGCTGGCGCTAATAGCGAAGAGGCC
Spacer'	1230	1241	CATTCGGAGACC
Ebox'	490	501	GCCAGGCTCAAG
Sp1-2	107	96	CGGTGCGGGCCT
Sp1-2	618	629	ATGGGCAGGTAG
Sp1-2	682	693	CTCGGCAGGAGC
CoreNRE'	448	463	CGTCGTGACCCATGGC
p53'	787	798	GGGCGTTCCTTG
LSF-YY1'	594	606	GCGAAACATCGCA
E2F-1	415	428	ATTTTCCACCATGA
Sp1-3'	368	380	TGGGTGTGGCGGA
NFkB1-1	411	425	GGCCATTTTCCACCA
NFkB1-1'	210	224	TCTGACTGTCAGACC
Ebox2'	775	788	TGCGCAGCTGTACT
USF	442	457	GGCATCGCCATGGGTC
CBF'	307	318	GGCGGCGAATGG
Ebox	449	460	CCATGGGTCACG
Sp1-3	592	604	GATGCGATGTTTC
RCS'	732	754	GAAGGGGACTGGCTGCTATTGGG
Ebox	339	350	GGGTAGCCAACG

3.11 Discussion

We have demonstrated a novel approach to the generation of highly diverse combinatoric enhancer like sequences. Specifically, we developed a novel experimental approach of representing transcription factor binding site (TFBS) polymorphisms as highly degenerate and short dsDNA parts. By using TFBS as parts, we conjecture that this will enable the explicit and systematic study of the relationship between the organization of TFBS into larger modules and gene expression output. Previous combinatoric approaches[31,46,47] have been largely driven by *a priori* expectations of functional architectures and may have fallen prey to confirmation bias. By limiting their efforts to modest shuffling of larger promoter fragments, they ultimately explored small sequence spaces that do not explicitly sample outside the universe of known functional architectures.

Our approach here uses limited *a priori* assumption of functional enhancer architectures. The primary design constraint is inclusion of motifs as parts for

assembly. A fundamental advantage of this is that the ensemble of input motifs can be tailored to study alternative promoters or to broadly select for promoter architectures that support gene expression in cell or tissue types of interest. Specifically, beyond such fundamental studies of promoter structure-function relationships and associated transfer functions, we envision the application of our approach in the development of predictable expression output for fully engineered systems, and as an enabling technology for the development of cell or tissue type specific transcriptional control.

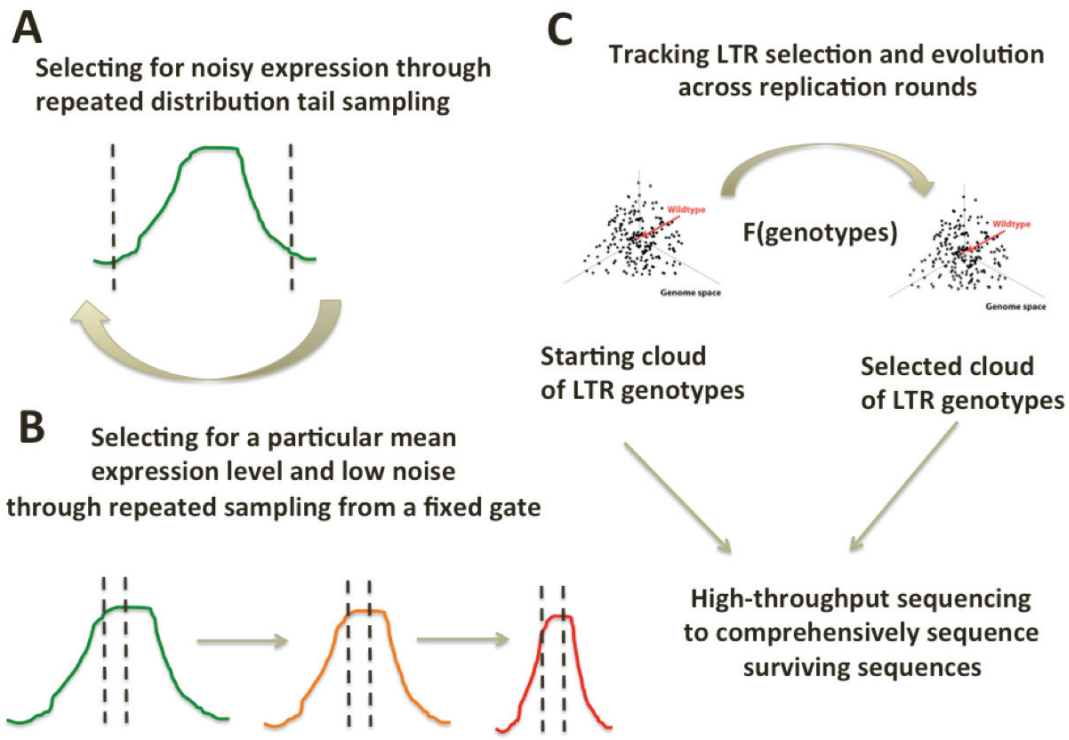


Figure 3.11| Functional selection strategies: (A) To systematically infer the relationship between promoter architecture and expression noise, architectures that enhance noise can be enriched by repeatedly sampling from alternate tails of a polygeneic population representing a diverse library of LTR genotypes. (B) Similarly, to select for maintenance of constant mean and reduced expression noise , a narrow gate centered around a fixed mean level of expression may be used. (C) In addition to such quantitative inferences of the LTR structure-function, we envision tracking the selection of a large input pool of LTR genotypes through rounds of HIV replication in a tissue culture model. High-throughput sequencing of the genotypes that survive each round would enable comprehensive reconstruction of selection trajectories.

We hypothesize that through synthetic generation of a large LTR architectural space, our approach will enable systematic inference of the LTR structure-function relationship. While additional technical challenges in the cloning of sequence and generation of a highly diverse library remain, we have demonstrated the potential of our approach to generate highly diverse combinatoric

sequences that can ultimately be subjected to functional selections. While recent studies have suggested that different endogenous promoters[62,63], which differ in their *cis* acting architectures, display significantly different transcriptional dynamics, we lack a broad quantitative understanding between promoter architecture and expression noise. By using a non-replicating GFP reporter vector, we envision performing functional selections to infer the relationship between LTR architecture and fundamental gene expression metrics such as the mean and CV or noise of expression. In particular, LTR architectures that support high expression noise may be inferred through repeatedly selecting from alternate tails of a polygenic GFP reporter population (Fig. 3.11A). Alternatively, architectures supporting lower heterogeneity may be selected for by sampling from a narrow gate around a fixed mean level of expression. Over repeated rounds of selection, we anticipate this would lead to reduction in the expression noise within the population and a convergence toward the mean level of expression selected for (Fig. 3.11B). In complement to such studies, we envision tracking LTR architectures across rounds of viral replication. By providing a replicating virus with a large pool of LTR architectures, we anticipate being able to reconstruct the trajectories of genotypes that survive through the selective pressures imposed by the functional needs of a replicating virus. Specifically, we anticipate that numerous architectures alternative to those similar to naturally circulating architectures may survive selection.

3.12 Materials and Methods

3.12.1 DNA Shuffling:

Eleven subtype LTR were amplified from full-length molecule clones (Table 3.2) using PCR under standard conditions (primers below) with the addition of 0.5M Betaine using primers listed below. PmeI and BsaI restriction sites in forward and reverse primers reflect eventual cloning of shuffled LTR as a PmeI-BsaI, representing the U3 and R regions into pRCHIVsLTR. This represents U3 and R LTR regions. For DNA shuffling, LTR amplicons were digested with DNAseI (Roche) according the manufacturer's protocol on ice for 5 to 30 minutes. Attempts to assemble and reamplify shuffled sequences were performed as previously described[39].

Table 3.2: Subtype sequences obtained from NIH AIDS Reagent Program

Subtype	Reagent Name	Accession No.	Country of isolation
B/C	P98CN009.8	AF286230	China
C	P93IN904	AF067157	India
C	P93IN999	AF067154	India
A2	P94CY017.41	AF286237	Cyprus
A	P92UG037.1	U51190	Uganda

A/G	P92NG003.1	U88825	Nigeria
D	P94UG114.1	U88824	Uganda
B	pCLGIT	K03455	USA
B/F	P93BR029.4	AF005495	Brazil
F	P93BR020.1	AF005494	Brazil
H	P90CF056.1	AF005496	Central African Republic

PCR primers used in DNA Shuffling and NeXT:

5' RCHIV B_BC_C_BF_F_H_**Pme**
CCTC**gtttaaac**gaccggtggcaagtggtcaaaaagtagtgtgattggatgatttaagaccaatgacttacaaggcagctgtag

5p RCHIV A2_**Pme**
CCTC**gtttaaac**gaccggtggcaagtggtcaaaaagtagtgtgattggatgaTTTAAGGCCAATGAC

5p RCHIV A_**Pme**
CCTC**gtttaaac**gaccggtggcaagtggtcaaaaagtagtgtgattggatgaTCTAAGGCCAATGAC

5p RCHIV AG_**Pme**
CCTC**gtttaaac**gaccggtggcaagtggtcaaaaagtagtgtgattggatgaTTTGAGACCAATGAC

5p RCHIV D_**Pme**
CCTC**gtttaaac**gaccggtggcaagtggtcaaaaagtagtgtgattggatgaATTAAGACCAATGAC

3p RCHIV_**Bsa**_BC ggctcagatctggtctaaccaga**gagacc**cagtacaggcgaaaagcagctgc
3p RCHIV_**Bsa**_C1 ggctcagatctggtctaaccaga**gagacc**cagtacaggcgaaaagcagctgc
3p RCHIV_**Bsa**_C2 ggctcagatctggtctaaccaga**gagacc**cagtacaggcgaaaagcagctgcttatatgc
3p RCHIV_**Bsa**_A2 ggctcagatctggtctaaccaga**gagacc**cagtacatgcgagaagcagctgc
3p RCHIV_**Bsa**_A ggctcagatctggtctaaccaga**gagacc**cagtacaagcgagaagcagctgc
3p RCHIV_**Bsa**_AG ggctcagatctggtctaaccaga**gagacc**cagtacaggcgagaagcggctgc
3p RCHIV_**Bsa**_D ggctcagatctggtctaaccaga**gagacc**cagtacaggcagaaagcagctgc
3p RCHIV_**Bsa**_B ggctcagatctggtctaaccaga**gagacc**cagtacaggcaaaaagcagctgc
3p RCHIV_**Bsa**_BF ggctcagatctggtctaaccaga**gagacc**cagtacaggcagaaagcagctgc
3p RCHIV_**Bsa**_F ggctcagatctggtctaaccaga**gagacc**cagtacaggcgaaaagcggctgc
3p RCHIV_**Bsa**_H ggctcagatctggtctaaccaga**gagacc**cagtacaagcgaaaagcagctgc

5p_U3_**Pme**_reamp CCTC**gtttaaac**gaccggtggcaagtggtcaaaaagtagtgtgattggatga
3p_U3_**Bsa**_reamp ggctcagatctggtctaaccaga**gagacc**

3.12.2 NeXT Shuffling:

Initial amplification of initial parental LTR was performed as in DNA shuffling except that 100mM dUTP (Invitrogen) was doped in place of dTTP at molar percentages ranging from 5 to 50%. Uracil was excised from UTP doped amplicons using Uracil DNA-deglyclosolase (UDG, New England Biolabs) under the manufacturer's recommended conditions. The DNA backbones of excised sequences

were subsequently cleaved with Piperidine (Sigma-Aldrich) and attempts to assemble resulting fragment distributions was performed as previously described[40].

3.12.3 Generation of position weight matrices and IUPAC oligonucleotides for LTR motifs:

Custom alignments over LTR motifs identified from literature sources as described were generated using the 2010 Los Alamos National Laboratory reference HIV alignment, which consists of 1492 curated sequences. The full alignment was cropped to the coordinates for each motif. Custom software in Python using Biopython (http://www.biopython.org) packages was used to estimate position weight matrices from cropped alignments. Degenerate IUPAC oligonucleotides were then output using a 0.1 bit cutoff over the resulting motif PWM. The melting temperature range of degenerate motifs was determined using OligoCalc[64]. Motifs with average T_m with padded with N's where necessary to increase the average melting temperature above 40°C. Source code is available on request.

3.12.4 Kinetic trapping of capping hairpin :

HPCapSfiI:

5'-ggccggagtggccgctgcgtctgtttaaacgtcccgcgcggacgtttaaacagacgcagcggccactccggcc-3'

A stable-stem loop structure was generated by resuspending the above 5' phsophosylated 'HPCapSfiI ' oligonucleotide in 1X New England Biolabs Buffer #2 at a final concentration of 200 ng/ul. The solution was heated to 90°C for 5 minutes and immediately quenched on ice. The desired stem-loop structure was verified on a 3% agarose gel.

3.12.5 Generation of blunt dsDNA parts and combinatoric assembly of parts:

Synthetic oligonucleotides consisting of degenerate IUPAC motifs, a Mly I restriction site, and a constant primer region were designed *in silico* (Table 3.3) and ordered from Invitrogen Inc. as standard custom desalted oligonucleotides in 96-well plate format. To facilitate primer extension, a 5' biotinylated primer (Bt-extend) was also designed. Part generation and combinatoric assembly was adapted from the previously described[65] method of nonhomologous random recombination and is fully described in Appendix B. Assembly products were resolved on a 2% agarose gel and products ranging from 200bp to 2 kbp gel extracted using standard methods.

Table 3.3 Synthetic oligonucleotides used in LTR motif part generation

Oligo sequence (5' to 3')	Oligo name
NNNGACTTTCCRNNNGGACCGACTCCACGAACTCCAGCAGGACCAT	NFkB1-1
GGGACGTGGYBAGGACCGACTCCACGAACTCCAGCAGGACCAT	Sp1-1
GNNTKMMWSANNGGGACCGACTCCACGAACTCCAGCAGGACCAT	Ap1
DCCATNRWCACVKRGC GGACCGACTCCACGAACTCCAGCAGGACCAT	USF

NNGNDCCATNGNNGGACCGACTCCACGAACTCCAGCAGGACCAT	LSF-YY1
WYTACAARRACTGCA GGACCGACTCCACGAACTCCAGCAGGACCAT	LEF1
NGN ACT GMT RAC NGNGGACCGACTCCACGAACTCCAGCAGGACCAT	RBF2
NGC AYK DMR WCA YVK NGGACCGACTCCACGAACTCCAGCAGGACCAT	C/EBP
NGT TTC CRC TWR GN GGACCGACTCCACGAACTCCAGCAGGACCAT	E2F-1
NGG YGT GGY YTNGGACCGACTCCACGAACTCCAGCAGGACCAT	p53
CAYKDMRWCACVTRGCGGACCGACTCCACGAACTCCAGCAGGACCAT	CoreNRE
GGRCCAGGGRYYAGATWYCCACTGWSHTTTGGRTGGACCGACTCCACGAACTCCAGCAGGACCAT	COUP
RRDGARGTGARRARGMCGGGMAATCRAAGGAGAR A GGACCGACTCCACGAACTCCAGCAGGACCAT	NFAT
NGN TGG GAA NGNGGACCGACTCCACGAACTCCAGCAGGACCAT	CBF
NAT GCG TAG GCNGGACCGACTCCACGAACTCCAGCAGGACCAT	Spacer
S M S G A M A G A G A A G T G T T GGACCGACTCCACGAACTCCAGCAGGACCAT	NMF-1
R R W R T G G A A G T T G A C A C C GGACCGACTCCACGAACTCCAGCAGGACCAT	NMF-2
Y M G C T A G C A Y K H M R W C A C A V T G G C C CGGACCGACTCCACGAACTCCAGCAGGACCAT	NMF3
M W R C A T C C G G A G T GGACCGACTCCACGAACTCCAGCAGGACCAT	EBS
A C V T R G C Y M R M GGGACCGACTCCACGAACTCCAGCAGGACCAT	Ebox
RTYTGAGCCYGGGAGCTCTCTGGGGACCGACTCCACGAACTCCAGCAGGACCAT	RCS
G G M C K T T C C R GGGACCGACTCCACGAACTCCAGCAGGACCAT	NFkB1-2
NTGGGCGGGACNGGACCGACTCCACGAACTCCAGCAGGACCAT	Sp1-2
NRG GYG HGG YTT NGGACCGACTCCACGAACTCCAGCAGGACCAT	Sp1-3
NGNNCANNTGNGNNGGACCGACTCCACGAACTCCAGCAGGACCAT	Ebox2
Biotin-ATGGTCCTGCTGGAGTTCGTGGAGTCGGTCC	Bt-extend

3.12.6 Vectors

The LTR from pRCHIV-sLTR[58] was cloned as a PmeI-KasI fragment into pUC19 to generate pUC19-sLTR. SfiI sites were introduced at two locations via QuickChange site-specific mutagenic PCR (Stratagene) using the primers below to generate pUC19-sLTR-Sfimut. The mutant sLTR was subcloned into pRCHIV-sLTR as a PmeI-KasI fragment to generate pRCHIV- sLTR-Sfimut.

U3Lib_SfiMut3p_F: 5'-ggaggcgtggcctgggcgggacggccggagtggccgagccctcagatgctg-3'
U3Lib_SfiMut3p_R: 5'- cagcatctgagggctcggccactccggccgtcccgcccaggccacgcctcc-3'
U3Lib_SfiMut5p_Fb:
5'-ccacttttaaaagaaaaggggccactccggcctaattcactcccaaagaagacaag-3'
U3Lib_SfiMut5p_Rb:
5'- cttgtcttctttgggagtgaattaggccactccggccccctttctttaaaaagtgg-3'

3.12.8 Library cloning:

Gel extracted combinatoric polymers and pRCHIV- sLTR-Sfimut were digested with SfiI (New England Biolabs), phenol-chloroform extracted, and ethanol precipitated. 150 ng of digested backbone was ligated overnight at 16°C with 50-450 ng of combinatoric inserts using T4 DNA ligase (NEB). Ligation products ethanol precipitated and electroporated into electrocompetent DH10B (Invitrogen). Library diversity was estimated through serial dilution of electroporated cells and colony counting following overnight growth on LB-Agar.

3.12.7 Computational annotation of combinatoric sequences:

Custom GUI Python software to output text based and graphical annotations of Sanger sequenced combinatoric LTR was implemented using FIMO[59] as well as Biopython and Wx-Python packages. Source code is available on request.

3.13 References

1. Yuh, C. H., Bolouri, H. & Davidson, E. H. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* **279**, 1896–1902 (1998).
2. Davidson, E. H. *et al.* A genomic regulatory network for development. *Science (New York, N.Y.)* **295**, 1669–78 (2002).
3. Riethoven, J.-J. M. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. *Methods in molecular biology (Clifton, N.J.)* **674**, 33–42 (2010).
4. Setty, Y., Mayo, A. E., Surette, M. G. & Alon, U. Detailed map of a cis-regulatory input function. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 7702–7707 (2003).
5. Wilczynski, B., Liu, Y.-H., Yeo, Z. X. & Furlong, E. E. M. Predicting spatial and temporal gene expression using an integrative model of transcription factor occupancy and chromatin state. *PLoS computational biology* **8**, e1002798 (2012).
6. Tautz, D. Evolution of transcriptional regulation. *Current opinion in genetics & development* **10**, 575–9 (2000).
7. Gordon, K. L. & Ruvinsky, I. Tempo and mode in evolution of transcriptional regulation. *PLoS genetics* **8**, e1002432 (2012).
8. Landry, C. R., Lemos, B., Rifkin, S. A., Dickinson, W. J. & Hartl, D. L. Genetic properties influencing the evolvability of gene expression. *Science* **317**, 118–121 (2007).
9. Maerkl, S. J. & Quake, S. R. Experimental determination of the evolvability of a transcription factor. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 18650–5 (2009).
10. Maerkl, S. J. & Quake, S. R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science (New York, N.Y.)* **315**, 233–7 (2007).
11. Fordyce, P. M. *et al.* De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nature biotechnology* **28**, 970–5 (2010).
12. Barrera, L. O. & Ren, B. The transcriptional regulatory code of eukaryotic cells--insights from genome-wide analysis of chromatin organization and transcription factor binding. *Current opinion in cell biology* **18**, 291–298 (2006).
13. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
14. Ben-Tabou de-Leon, S. & Davidson, E. H. Gene regulation: gene control network in development. *Annual review of biophysics and biomolecular structure* **36**, 191 (2007).

15. Davidson, E. H. Network design principles from the sea urchin embryo. *Current opinion in genetics & development* **19**, 535–40 (2009).
16. Ben-Tabou de-Leon, S. & Davidson, E. H. Modeling the dynamics of transcriptional gene regulatory networks for animal development. *Developmental biology* **325**, 317–28 (2009).
17. Tharakaraman, K., Bodenreider, O., Landsman, D., Spouge, J. L. & Marino-Ramirez, L. The biological function of some human transcription factor binding motifs varies with position relative to the transcription start site. *Nucleic Acids Res* **36**, 2777–2786 (2008).
18. Komurov, K. & White, M. Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Mol Syst Biol* **3**, 110 (2007).
19. Wilson, C., Bellen, H. J. & Gehring, W. J. Position effects on eukaryotic gene expression. *Annual review of cell biology* **6**, 679–714 (1990).
20. Kumaran, R. I., Thakar, R. & Spector, D. L. Chromatin dynamics and gene positioning. *Cell* **132**, 929–934 (2008).
21. Montanuy, I., Torremocha, R., Hern**&#xfffd;**ndez-Munain, C. & Su**&#xfffd;&#xfffd;**, C. Promoter influences transcription elongation: TATA-box element mediates the assembly of processive transcription complexes responsive to cyclin-dependent kinase 9. *The Journal of biological chemistry* **283**, 7368–7378 (2008).
22. Gorski, S. A., Snyder, S. K., John, S., Grummt, I. & Misteli, T. Modulation of RNA polymerase assembly dynamics in transcriptional regulation. *Mol Cell* **30**, 486–497 (2008).
23. Orphanides, G. & Reinberg, D. RNA polymerase II elongation through chromatin. *Nature* **407**, 471–5 (2000).
24. Croston, G. E. & Kadonaga, J. T. Role of chromatin structure in the regulation of transcription by RNA polymerase II. *Current opinion in cell biology* **5**, 417–23 (1993).
25. KADONAGA, J. Regulation of RNA Polymerase II Transcription by Sequence-Specific DNA Binding Factors. *Cell* **116**, 247–257 (2004).
26. Lam, F. H., Steger, D. J., O'Shea, E. K. & O**&#xfffd;**shea, E. K. Chromatin decouples promoter threshold from dynamic range. *Nature* **453**, 246–250 (2008).
27. Mao, C. *et al.* Quantitative analysis of the transcription control mechanism. *Molecular systems biology* **6**, 431 (2010).
28. Beer, M. A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**, 185–198 (2004).
29. Mantzaris, N. V From single-cell genetic architecture to cell population dynamics: quantitatively decomposing the effects of different population heterogeneity sources for a genetic network with positive feedback architecture. *Biophys J* **92**, 4271–4288 (2007).
30. Isaacs, F. J., Hasty, J., Cantor, C. R. & Collins, J. J. Prediction and measurement of an autoregulatory genetic module. *Proc Natl Acad Sci U S A* **100**, 7714–7719 (2003).
31. Murphy, K. F., Bal**&#xfffd;**zsi, G. & Collins, J. J. Combinatorial promoter design for engineering noisy gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 12726–12731 (2007).

32. Bushman, F. *et al.* Genome-wide analysis of retroviral DNA integration. *Nat Rev Microbiol* **3**, 848–858 (2005).
33. Burnett, J. C. *et al. Combinatorial latency reactivation for HIV-1 subtypes and variants.*
34. Wain-Hobson, S. The fastest genome evolution ever described: HIV variation in situ. *Current opinion in genetics & development* **3**, 878–83 (1993).
35. Charpentier, C., Nora, T., Tenaillon, O., Clavel, F. & Hance, A. J. Extensive recombination among human immunodeficiency virus type 1 quasispecies makes an important contribution to viral diversity in individual patients. *Journal of virology* **80**, 2472–2482 (2006).
36. Burnett, J. C., Miller-Jensen, K., Shah, P. S., Arkin, A. P. & Schaffer, D. V Control of stochastic gene expression by host factors at the HIV promoter. *PLoS Pathog* **5**, e1000260 (2009).
37. Stemmer, W. P. Rapid evolution of a protein in vitro by DNA shuffling. *Nature* **370**, 389–91 (1994).
38. Bloom, J. D. *et al.* Evolving strategies for enzyme engineering. *Current opinion in structural biology* **15**, 447–52 (2005).
39. Maheshri, N., Koerber, J. T., Kaspar, B. K. & Schaffer, D. V Directed evolution of adeno-associated virus yields enhanced gene delivery vectors. *Nature biotechnology* **24**, 198–204 (2006).
40. Müller, K. M. *et al.* Nucleotide exchange and excision technology (NExT) DNA shuffling: a robust method for DNA fragmentation and directed evolution. *Nucleic acids research* **33**, e117 (2005).
41. Krebs, F. C., Hogan, T. H., Quiterio, S., Gartner, S. & Wigdahl, B. Lentiviral LTR-directed Expression , Sequence Variation , and Disease Pathogenesis. *Microbiology* **10**, 29–70 (2001).
42. Pereira, L. A., Bentley, K., Peeters, A., Churchill, M. J. & Deacon, N. J. A compilation of cellular transcription factor interactions with the HIV-1 LTR promoter. *Nucleic Acids Res* **28**, 663–668 (2000).
43. Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research* **32**, D91–4 (2004).
44. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome research* **14**, 1188–90 (2004).
45. Gertz, J., Siggia, E. D. & Cohen, B. A. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* **457**, 215–8 (2009).
46. Kinkhabwala, A. & Guet, C. C. Uncovering cis regulatory codes using synthetic promoter shuffling. *PloS one* **3**, e2030 (2008).
47. Cox, R. S., Surette, M. G. & Elowitz, M. B. Programming gene expression with combinatorial promoters. *Molecular systems biology* **3**, 145 (2007).
48. Schlabach, M. R., Hu, J. K., Li, M. & Elledge, S. J. Synthetic design of strong promoters. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 2538–43 (2010).
49. Bittker, J. A., Le, B. V & Liu, D. R. Nucleic acid evolution and minimization by nonhomologous random recombination. *Nature biotechnology* **20**, 1024–9 (2002).

50. Upcroft, P. & Healey, A. Rapid and efficient method for cloning of blunt-ended DNA fragments. *Gene* **51**, 69–75 (1987).
51. Ludwig, M. Z., Manu, Kittler, R., White, K. P. & Kreitman, M. Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. *PLoS genetics* **7**, e1002364 (2011).
52. Aldana, M., Balleza, E., Kauffman, S. & Resendiz, O. Robustness and evolvability in genetic regulatory networks. *J Theor Biol* **245**, 433–448 (2007).
53. Ciliberti, S., Martin, O. C. & Wagner, A. Robustness can evolve gradually in complex regulatory gene networks with varying topology. *PLoS Comput Biol* **3**, e15 (2007).
54. Meyerhans, A., Vartanian, J. P. & Wain-Hobson, S. DNA recombination during PCR. *Nucleic acids research* **18**, 1687–91 (1990).
55. Kanagawa, T. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of bioscience and bioengineering* **96**, 317–23 (2003).
56. Williams, R. *et al.* Amplification of complex gene libraries by emulsion PCR. *Nature methods* **3**, 545–50 (2006).
57. Schütze, T. *et al.* A streamlined protocol for emulsion polymerase chain reaction and subsequent purification. *Analytical biochemistry* **410**, 155–7 (2011).
58. Shah, P. S., Pham, N. P. & Schaffer, D. V HIV develops indirect cross-resistance to combinatorial RNAi targeting two distinct and spatially distant sites. *Molecular therapy : the journal of the American Society of Gene Therapy* **20**, 840–8 (2012).
59. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)* **27**, 1017–8 (2011).
60. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* **37**, W202–8 (2009).
61. Jajamovich, G. H., Wang, X., Arkin, A. P. & Samoilov, M. S. Bayesian multiple-instance motif discovery with BAMBI: inference of recombinase and transcription factor binding sites. *Nucleic acids research* **39**, e146 (2011).
62. Zenklusen, D., Larson, D. R. & Singer, R. H. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural & Molecular Biology* **15**, 1263–1271 (2008).
63. Suter, D. M. *et al.* Mammalian genes are transcribed with widely different bursting kinetics. *Science (New York, N.Y.)* **332**, 472–4 (2011).
64. Kibbe, W. A. OligoCalc: an online oligonucleotide properties calculator. *Nucleic acids research* **35**, W43–6 (2007).
65. Bittker, J. A., Le, B. V & Liu, D. R. Nucleic acid evolution and minimization by nonhomologous random recombination. *Nat Biotechnol* **20**, 1024–1029 (2002).

Chapter 4: Shape based clustering and moment analysis of highly heterogeneous Tat feedback distributions reveals characteristic scaling relationships

4.1 Introduction

The virally encoded factor Tat is a multi-function protein that coordinates the subversion of eukaryotic transcriptional control to strongly trans-activate HIV gene expression through a transcriptional feedback circuit. Tat mediates positive feedback by orchestrating multiple transcriptional regulatory steps to overcome initially repressed and weak basal LTR expression and amplify viral gene expression 100-fold. Specifically, Tat has been shown to recruit host-cells factors involved in chromatin remodeling[1,2], transcriptional initiation[3], transcriptional elongation[4–6], and RNA processing. However, the kinetics of these processes and their roles in determining Tat feedback dynamics are largely unknown. Previous work has demonstrated that Tat feedback is highly susceptible to stochastic effects and displays integration position dependent dynamics[7,8]. These stochastic and position dependent effects have confounded efforts to systematically model Tat feedback dynamics across viral integration positions. Specifically, previous models[7,8] have focused on a bimodal Tat feedback, which represents a subset of observed outputs of the Tat feedback circuit. Therefore, a model that systematically explains feedback dynamics across integration would significantly our understanding of the coupling between basic gene circuits and the local genomic in which they operate.

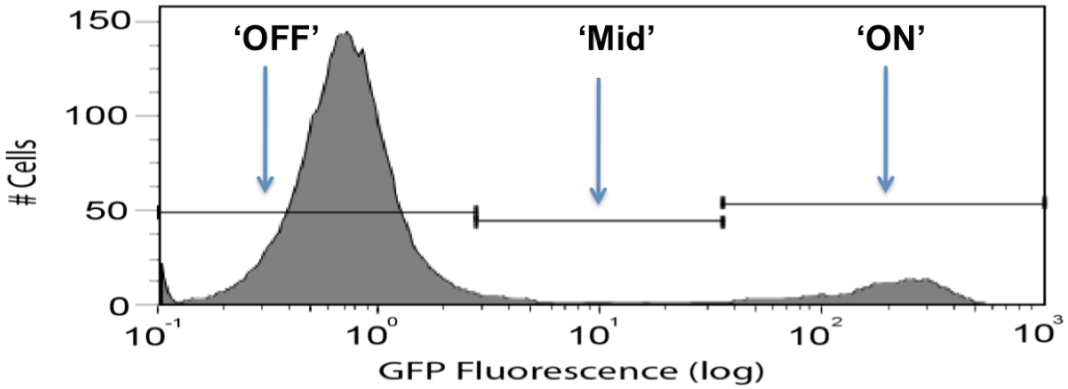


Fig.4.1|HIV LGIT minimal feedback circuit gene expression exhibits bimodality: As demonstrated previously[7], analysis of the steady-state expression profile of a polyclonal population following sorting of TNFα stimulated GFP positive cells reveals a characteristic bimodal distribution with two well separated modes. This distribution is divided into three regions ' OFF', 'Mid' and 'ON', with 'OFF' and 'ON' corresponding to low and high levels of Tat respectively. The 'OFF' population suggests that the Tat feedback circuit exhibits an intrinsic capability to entire the latent state. Interestingly, a subset of single cell clones isolated from the 'Mid' region exhibit dynamic gene expression with isogenic cells transiting between the 'OFF' and 'ON' states.

Polyclonal populations representing single integrations of a minimal Tat feedback vector (HIV LGIT) exhibit bimodal expression[7,9] (Fig. 1). The Tat feedback circuit is though to be deterministically mono-stable, with a single stable state corresponding to the 'Off' mode of the bimodal population[10]. Therefore, in the absence of deterministic bistability, the observed bimodality may arise from feedback amplified excitatory transients[8], or noise driven steady-state bimodality[11,12]. Interestingly, recent theoretical[13,14] and experimental[15] work has suggested the potential role of rate-limiting stochastic nucleosome remodeling in generating steep, switch-like activation and bimodality through the regulated occlusion of essential activating transcription factor (TF) binding sites. Stochastic nucleosome remodeling and binding of transcription factors to these occluded sites that recruit chromatin-remodeling activity are thought to form a kind of epigenetic positive feedback loop with TF binding reinforcing the remodeled state. The frequency of the initial stochastic remodeling is likely influenced by the local histone code[16], which biases the modification-directed recruitment of factors that reinforce the local chromatin state.

Our findings in Chapter 2 together with our previous work[17] demonstrate that viral integration position has a dramatic effect on the mean and noise of basal expression. Furthermore, in Chapter 2 we demonstrate a link between Nuc-1 occupancy on the LTR and expression noise. However, the impact of these observations on Tat feedback dynamics is unknown. A fundamental molecular function of Tat is the recruitment of SWI/SNF complexes to remodel the repressive Nuc-1, which serves as rate-limiting barrier to productive elongation. Therefore, differential occupancy of Nuc-1 across viral integration positions may significantly impact the rate of Tat-mediated Nuc-1 modeling and the resulting gene expression output. Furthermore, Tat feedback may couple with other features of the local integration context. Therefore, integration position is a confounding factor that represents a significant barrier to systematically model Tat feedback.

Tat feedback generates highly heterogeneous distribution shapes across integration positions. These different distribution shapes are likely indicative of alternative parameterizations of the kinetic feedback mechanism. In order to reduce the size of the experimental space that must be explored to infer the kinetic parameter phase space that generates the observed distributions, we develop a clustering based approach that infers a minimum number of clusters to succinctly represent observed distribution classes. Using this approach we demonstrate clustering of 342 minimal Tat feedback clones into 30 representative classes. Furthermore, we develop and demonstrate the efficacy of an experimental system to directly relate basal expression and noise to Tat feedback mediated expression while controlling for integration position.

4.2 Development of an experimental system to systematically relate basal expression noise to Tat feedback dynamics

While in Chapter 2 we systematically inferred kinetic parameters and chromatin occupancy underlying differential basal expression and noise across viral integration positions, we lack a fundamental understanding of the relationship between the molecular and kinetic features of basal and Tat feedback dynamics. Recent studies[11,12] have indicated the potential for noisy expression coupled with short half-life transcription factors to generate bimodal gene expression patterns similar to those we have observed[7] for Tat feedback distributions. Owing to the 33 million[18] individuals that in the face of HIV infection unfortunately have the Tat feedback circuit underlying their current and future health, understanding the mechanisms underlying Tat feedback is of acute biomedical importance. Specifically, understanding how the local context of the integration position and its demonstrated modulation of expression noise influences Tat feedback would greatly enhance our understanding of the mechanisms of that allow the Tat feedback circuit to stochastically reactivate to regenerate active infection[19–21] after long periods of quiescence.

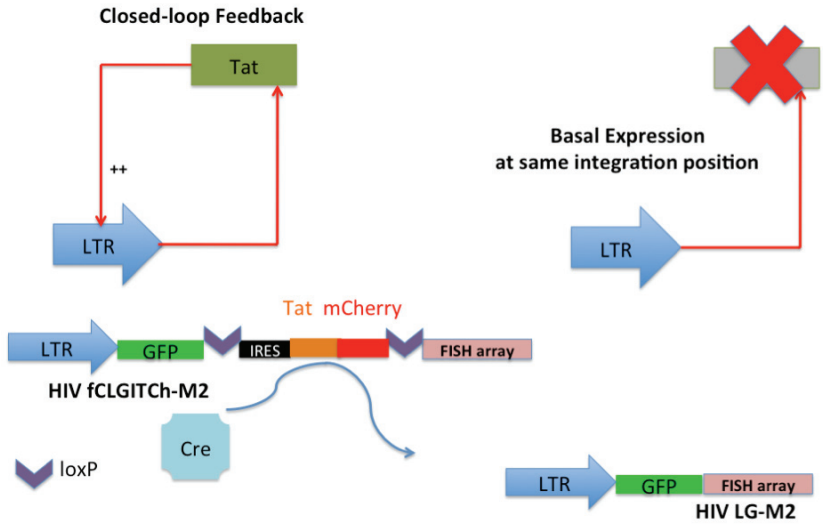


Figure 4.2| *In situ* conversion of close-loop Tat feedback circuit to basal expression cassette: To enable direct systematic analysis of the relationship between basal LTR expression noise and heterogeneous Tat mediated gene expression we developed a *floxed* variant of LGIT to enable excision of Tat while controlling for the confounding effects of integration position. We hypothesized that this system would enable direct conversion of a closed-loop feedback system to a basal expression cassette. To facilitate monitoring of excision, Tat is fused to a mCherry reporter. The vector,fCLGITCh-M2 contains a GFP reporter of protein expression and a repeated array for smFISH analysis of mRNA as performed in Chapter 2.

Toward this aim, we developed an experimental system to directly relate basal expression to Tat feedback while controlling for the confounding effects of viral integration position (Fig. 4.2). We hypothesized that by surrounding Tat by *loxP* sites, we could employ bacteriophage P1 Cre recombinase to efficiently excise Tat, thereby disabling the feedback circuit and converting the system to a basal expression cassette similar to that studied in Chapter 2. We modified our previous LGIT model system through introduction of *loxP* around the internal ribosome entry site (IRES) and Tat. The placement of *loxP* sites was chosen to generate a system highly similar to our basal, LG, expression system. Due to the likelihood of incomplete excision of Tat in clonal populations, and to facilitate tracking of excision, Tat was fused to a mCherry reporter. We reasoned that this system would permit systematic inference of the relationship between basal expression kinetics and Tat feedback across viral integration positions. Specifically, we sought to infer where in basal kinetic parameter phase space different Tat feedback distributions arise.

4.3 Large scale clonal analysis of highly heterogeneous feedback distribution shape

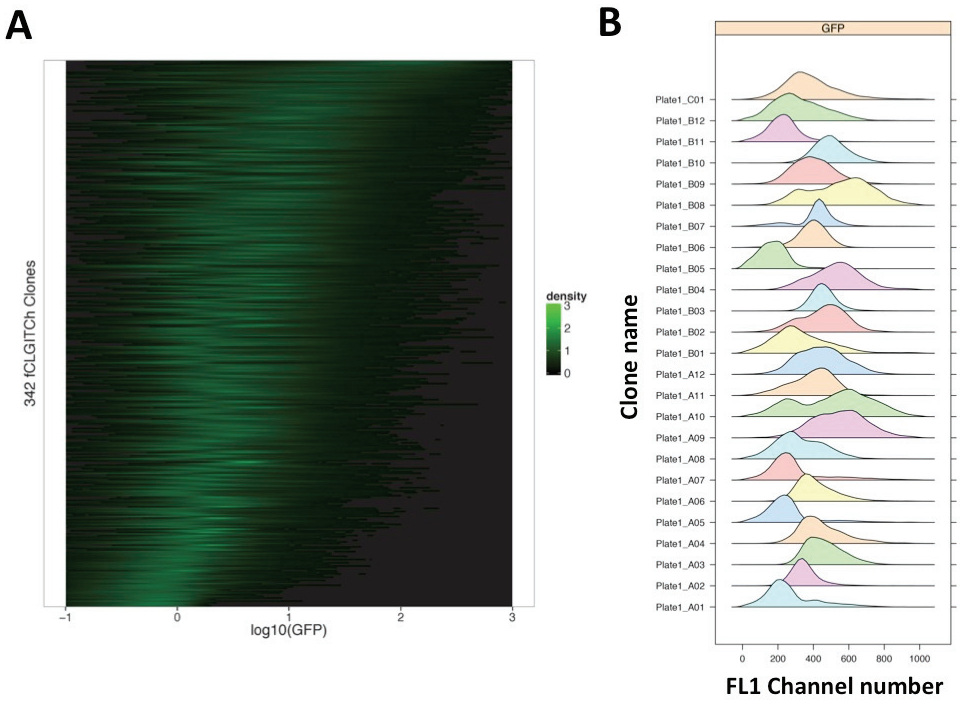


Figure 4.2| Highly heterogeneous distribution shape observed in large set of fCLGITCh clones: (A) The GFP expression 342 single-integration fCLGITCh clones was determined by flow cytometry. Clonal distributions were log10 transformed and rank ordered on mean. This large set of clones densely samples the variety of minimal Tat feedback distributions observed, capturing a wide variety of distribution shapes from monomodal dim through wide and bimodal to monomodal bright. (B) A subset of 25 clones demonstrates the considerable variation in distribution shape, which presents a challenge to systematically modeling Tat feedback kinetics.

While previous Tat feedback models have strictly focused on modeling bimodal distributions[7,8,10], we sought to systematically infer a model that can explain all observed distribution classes. We hypothesized that exploration of a large parameter state-space would be necessary to reasonably develop and infer such a model. Therefore, to support this aim we experimentally generated a large set of 342 fCLGITCh single-integration clones. The clones are highly heterogeneous with mean levels of expression spanning ~2.4 orders of magnitude and distributions that differ widely in their overall shape (Fig. 4.2A). A characteristic feature is their significant width, with a given distribution typically spanning 2 orders of magnitude. Qualitatively, observed distribution shapes include monomodal distributions with dim (10^{-1}-10^{0} RFU) or bright (10^{2}-10^{3} RFU) mean expression levels, bimodal distributions, and wide distributions spanning orders of magnitude. However, the overall distribution shape within these broad classes is widely different (Fig. 4.2B).

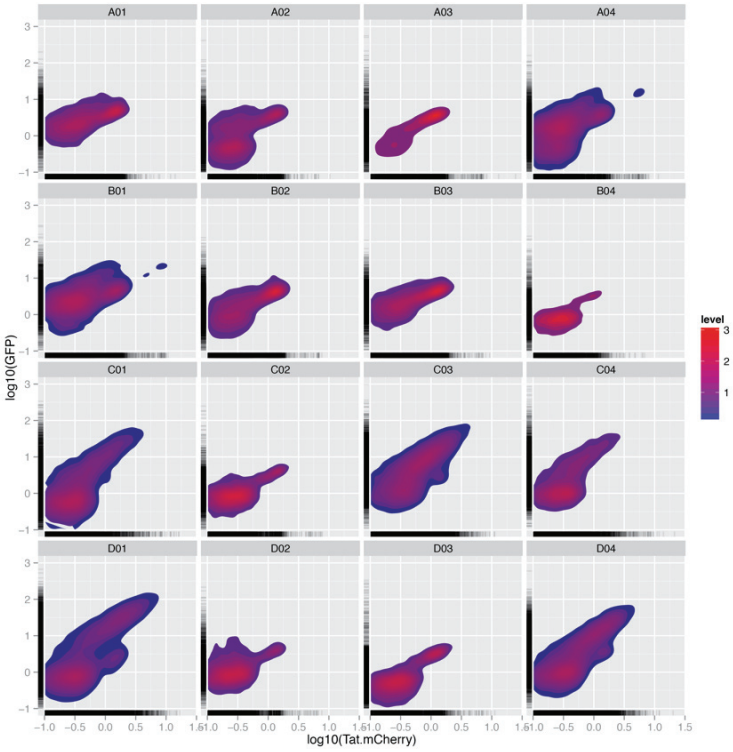


Figure 4.3| Representative contour plots of GFP versus Tat-mCherry qualitatively indicate correlation between two reporters: Kernel smoothed contour plots of GFP expression versus Tat-mCherry expression were plotted for the first four clones from the first four plates of clones. Contour level corresponds to the smoothed joint-density. This qualitatively indicates correlation between GFP and mCherry reporters. However, from the marginal rug-plots on each axis, it is apparent than the dynamic range of mCherry is compressed relative to GFP. This is expected due to the lower efficiency of IRES dependent translation and lower quantum efficiency of mCherry relative to GFP.

Previously, it was assumed that GFP expression shape is correlated with the hidden Tat expression. Due to its alternative recruitment of the ribosome in a non-cap dependent fashion, the two open reading frames in an X-IRES-Y configuration, are expected to be expressed in an approximately 10X:1 Y ratio[22]. However, the different mechanisms associated with GFP and Tat translation mechanisms and their differences in protein half-life (GFP~24 hrs, Tat~3 hrs) could serve to alter the effective dynamic of the two proteins. Therefore, as an initial *qualitative* assessment of the relationship between GFP and Tat-mCherry, we examined the joint-density of GFP and mCherry in sixteen representative clones (Fig. 4.3). We find qualitative evidence of linear correlation between the two reporters. However, it is apparent that for a given level of Tat-mCherry, a range of GFP is accessible. This may stem from differences in translational efficiency, or half-lives of the two reporters. Underlying birth-death processes of the two reporters may introduce additional noise into the system and therefore serve to partially deccorelate the two reporters.

To quantitatively investigate the correlation between GFP and Tat-mCherry we determined Pearson correlation coefficient across all 342 clones (Fig. 4.4). We found that a majority (72%) of clones exhibit a reasonable to high degree of correlation with r>0.7. However, a subset consisting of ~12% of all clones exhibit very low (r<0.15) to zero correlation. These observations do not appear to be an artifact of our use of a strictly linear correlation coefficient. We also found similar fractions of high and low correlation using Spearman r (data not shown).

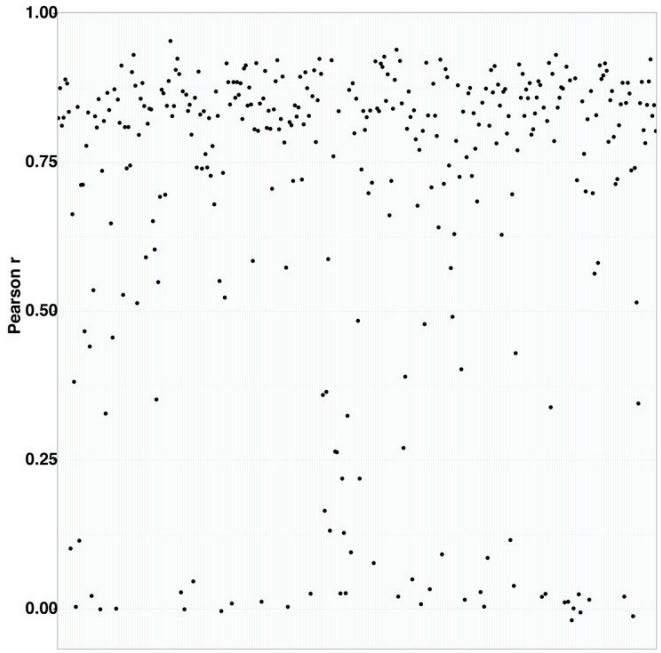


Figure 4.4| Pearson correlation coefficients for 342 fCLGITCh clones reveals fraction of aberrant clones with low to 0 correlation between GFP and Tat-mCherry: To investigate the correlation between GFP and mCherry reporters acrosss all clones, we determined the Pearson correlation coefficient of GFP versus Tat-mCherry for each clone. Each point represents a single clone. While a majority (~70%) of the clones exhibited r>0.75, a fraction exhibited unexpectedly low to zero correlation.

To further investigate the potential source of low to zero correlation for a the observed subset of clones, we examined the joint-distribution of clones with Pearson r<0.15 (Fig.4.3). We find that this subset is primarily composed of two classes. The first exhibits little dynamic range in GFP and is likely reflective of GFP or Tat mutations that limit its expression or fluorescence (for example see 'Plate2_B08'). The second exhibits dynamic range in both GFP and Tat-mCherry, yet the two are uncorrelated. This underlying cause is uncertain; however, these may also reflect mutations that effect translation or stability of either reporter or Tat. Such disabling or partially disabling mutations could arise from reverse transcription, which is a known low-fidelity process[23,24]. While we have never observed direct evidence of RT mutations, we have not previously systematically studied a two-reporter system. Therefore, without a second reporter expressed from the same provirus, such mutations likely went unnoticed. Therefore, our observation is an important consideration for future large-scale clonal studies of lentiviruses.

Figure 4.5| Evidence of RT induced mutations in low reporter correlation clones: We examined the joint GFP-mCherry distributions of clones with unexpectedly low correlation (r<0.15) between the two reporters. A fraction of these clones exhibit little dynamic range in GFP (e.g. 'Plate2_B06') with respect to Tat-mCherry and may represent defects in GFP or Tat expression or GFP fluorescence Other clones exhibit general disregulation and may represent mutations that alter translation or protein stability. Such mutations could arise during reverse transcription. Reverse transcriptase is a low-fidelity DNA polymerase and generates multiple mutations in each round[23,24].

4.4 Clustering of feedback distributions using the generalized minimum distance (GMD) of distributions:

In order to reduce the computational costs associated with systematically modeling a large set of clones, we sought to apply clustering and infer a minimal set of classes that compactly represent the observed variance in feedback distribution shape. Toward this aim, we assessed the ability of *k-means* clustering over the first 6 central moments to discriminate distributions on the basis of shape.
K-means is commonly applied method in distribution clustering that seeks to separate *n* observations into *k* clusters with partitioning occurring based on the closest mean between an observation and a cluster. Mean is broadly defined over the vector of features used for partitioning. We found that for reasonable values of *k* in the tens of clusters, *k-means* yields clusters with frequent occurrences of outliers. Due to their heterogeneity, it is likely that many of the distributions are poorly described and discriminated on the basis of central moments.

Therefore, we investigated methods that aim to establish distances between distributions on the basis of bin-to-bin comparisons of density. We reasoned that such bin-to-bin approaches would more efficiently describe and discriminate between distributions of differing shape. We initially identified the Earth Mover's Distance (EMD) as a candidate algorithm[25]. The EMD algorithm is based on the transportation problem from linear optimization and seeks to minimize the cost of the flow of material necessary to transform one distribution into another. Therefore, this computed cost effectively describes a distance metric that can then be used to efficiently cluster distributions in a non-parametric fashion on the basis of shape. EMD was originally developed for image retrieval applications[25] and has been applied to the clustering of multi-dimensional distributions arising from telecommunication networks[26]. While EMD is particularly well suited for comparing distributions over multiple dimensions, the linear optimization it entails incurs significant computational complexity and inefficient scaling to large datasets. A related metric is the Minimum Distance of Pair Assignments (MDPA). While less general than EMD and not suited for multi-dimensional distributions, MDPA is algorithmically efficient and scales to large sets of univariate distributions. MDPA effectively counts the number of "shifts" of elements required to transform one distribution into another.

Here we use an R implementation of MDPA termed the Generalized Minimum Distance of Distributions (GMD) that was recently used to systematically cluster transcriptional start site distributions at genome scale[27]. Specifically, we computed all versus all GMD metrics for all 342 fCLGITCh clones. This distance matrix was then used as input for hierarchical clustering. A number of methods for hierarchical clustering including mean, median, single linkage, complete linkage, and Ward's method were evaluated for their ability to generate clusters from the GMD distance matrix. We found that Ward's method, which aims to minimize the total within-cluster variance, generally yielded compact clusters and minimal outliers. To determine the smallest number of clusters that would provide reasonable discrimination of distribution shapes, we explored values of *k* ranging from 10 to 50. We found that 30 clusters provide the best compromise in terms of cluster number

and cluster quality on the basis of visual assessment (Fig.4.6). While some clusters contain members that differ somewhat in their peak density around their modes (e.g. Fig. 4.6, cluster 14), we observed overall excellent clustering and clear discrimination of classes such as narrow unimodal distribtions (e.g. cluster 3), unimodal long-tailed distributions (e.g. cluster 1), and bimodal distributions (e.g. cluster 28).

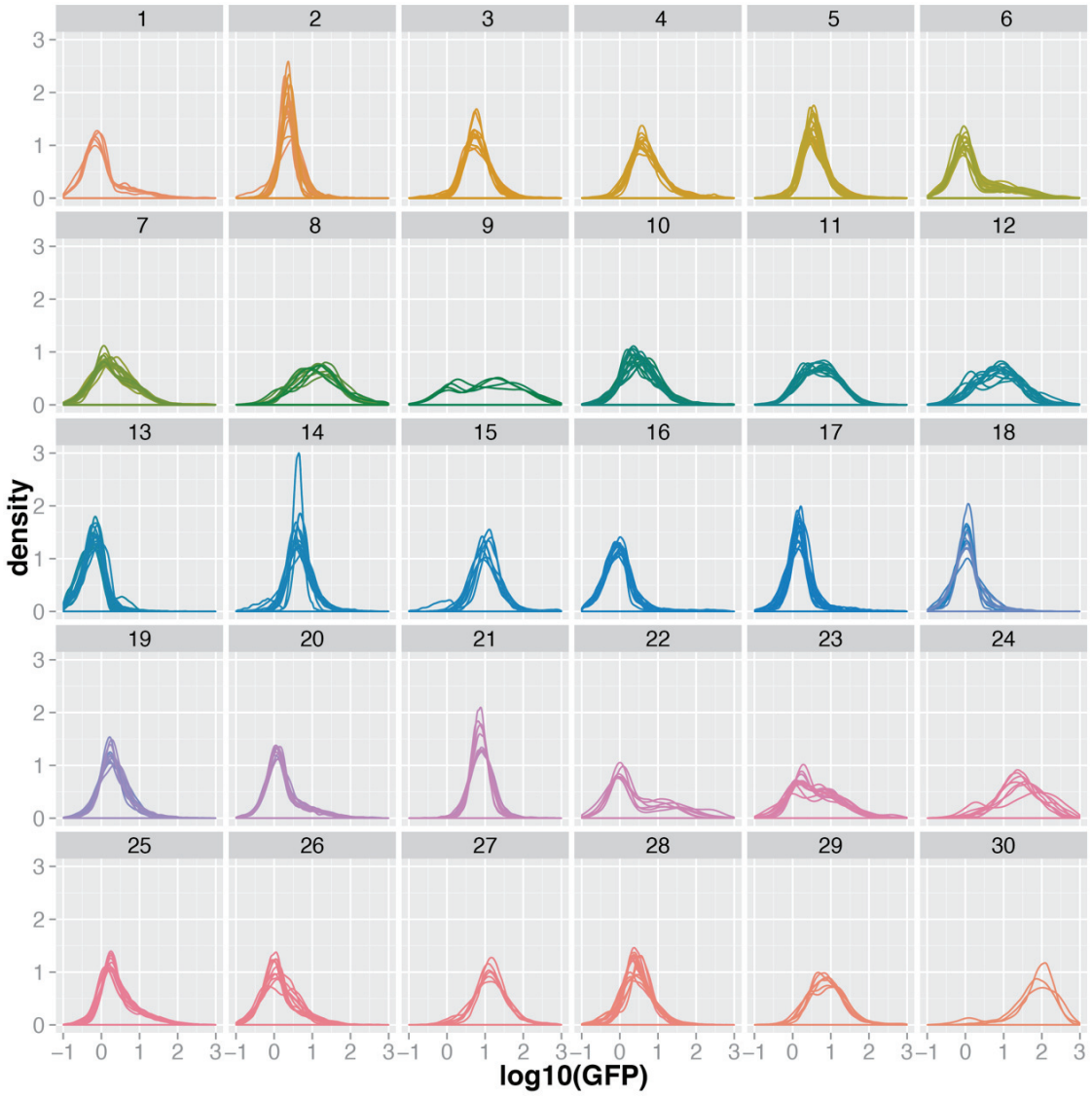


Figure 4.6| Hierarchical clustering of fCLGITCh distributions using the Generalized Minimum Distance of Distributions: In order to reduce the analytical and computational complexity associated with modeling a large set of clones, we determined the distance between distributions using the Generalized Minimum Distance of Distributions (GMD). An all versus all matrix of GMD distances was computed. Distributions were subsequently hierarchically clustered using Ward's method. We find that 30 clusters efficiently discriminate between distribution observed classes and results in limited outliers.

4.5 Distribution moment analysis

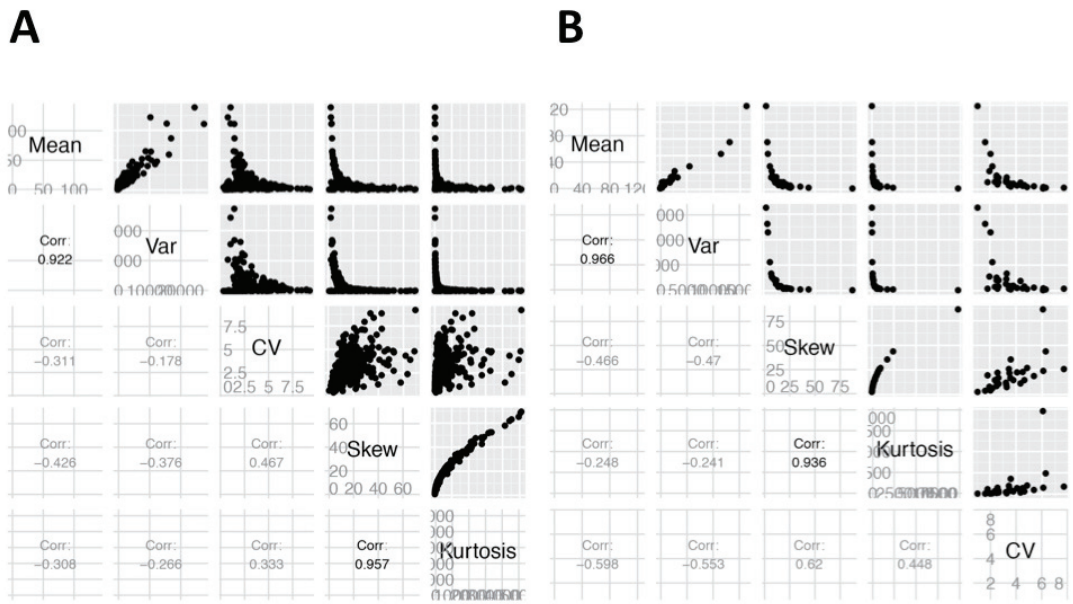


Figure 4.7| Exploratory moment based analysis of individual clones and clusters reveal highly correlated Mean~Variance and Kurtosis~Skew: (A) The first four central moments were compared against one another for all fCLGITCh clones. We find that Mean and Variance and Skew and Kurtosis are highly correlated (Pearson r=0.92 and r=0.96 respectively). This reveals fundamental scaling relationships that can be used to discriminate candidate models and approximations to the probability density function underlying the observed distributions. (B) When moments are calculated across the 30 clusters, we find statistically similar relationships ($p<0.05$), suggesting that clusters capture and describe fundamental properties of the underlying distributions.

Subsequent to establishing a clustering approach, we sought to establish whether analyzing clusters could provide a similar quantitative view of the system as analyzing individual clones. Furthermore, as we discussed in Chapter 2, the scaling between central moments can be useful in inferring a likely underlying kinetic transcriptional model. Therefore, we examined the scaling between the first 4 central moments and Coefficient of Variation (CV) for all 342 clones as well as for our set of 30 clusters (Fig. 4.7). Of the moment scaling relationships we examined across all clones, we find significant linear correlations between log-log transformed Mean and Variance (slope=1.81±0.12, R^2 =0.74, r=0.92,$p<0.001$) as well as between log-log transformed Skewness and Kurtosis (slope=1.93±0.03,R^2=0.98, r=0.96,$p<0.001$). Variances are large with respect to mean expression, with CV ranging from 0.4 to 9.2. This suggests that Tat feedback dramatically amplifies expression noise relative to basal noise (Chapter 2), with CV values typically 6-7X greater. Furthermore, distributions are strictly positively skewed and with significantly more density in their tails than Gaussian distributions, with Skewness ranging from 0.75 to 69 and Kurtosis greater than 3. Together, this suggests that

Tat feedback generates a high probability of extreme values that are far from the mode of the distribution. Interestingly, we find that Skewness is related to Kurtosis via a power law relationship with $K \sim S^{1.93 \pm 0.03}$.

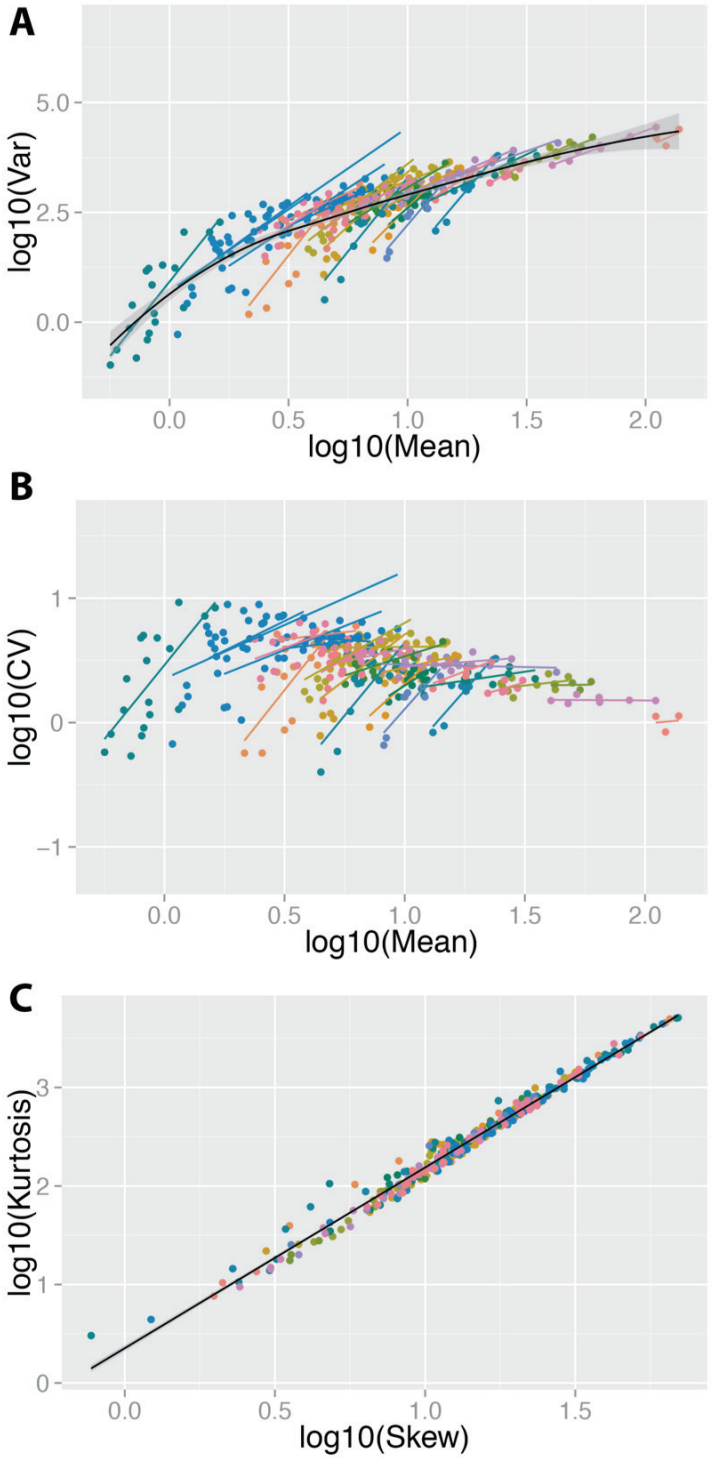


Figure 4.8| Clones partition by cluster into regions of the Mean~Variance and Mean~CV planes but not Kurtosis~Skew: (A) Variance and Mean are highly correlated, however residual curvature in log-log space indicates a strictly linear model may be insufficient. We find that clusters roughly segregate into distinct bands in the moment plane and clones within clusters (colored points) with different intracluster relationships (colored lines)(C) Similarly to Mean versus Variance, clones within clusters roughly segregate into distinct bands in the Mean versus CV plane. (D) Plotting log-log transformed Skew versus Kurtosis reveals a highly significant (R^2=0.96, regression p<0.001) power law relationship with $K \sim S^{1.93 \pm 0.03}$. Individual points are colored by cluster membership (1-30). Colored lines indicate linear regression slop within each cluster. Black lines indicate LOESS regression in (A) and linear regression in (C), with grey envelope indicating the point-wise 95% confidence interval.

Other scaling relationships, such as between Variance and Skew, are strongly non-linear and may be indicative of two different scaling regimes with one class along each axis. When moments were calculated across for each cluster we found quantitatively similar relationships. Specifically, we find that regression slopes for Mean~Var and Skew~Kurtosis are not significantly different from those when individual clonal distributions are considered ($p<0.05$). This suggests that clusters capture the fundamental moment scaling relationships that are observed when individual clones are examined. The significant scaling relationships observed are likely indicative of the properties of the underlying probability density function. While an exact interpretation of the mechanism that leads to the observed scalings is not clear, these findings are an important step toward resolving a kinetic model of Tat feedback that can broadly explain the observed distributions.

Subsequent to the observation of significant scaling relationships between Mean and Variance and between Skew and Kurtosis, we sought to understand whether the clones captured by each cluster partitioned into different regions within specific planes of moment space. Specifically, we examined the partitioning of clusters in the log-log transformed Mean-Variance, Mean-CV, and Skew-Kurtosis planes (Fig. 4.8). Interestingly, we find that clones from each of the 30 clusters parturition into approximately unique regions of the Mean-Variance plane and Mean-CV planes. Furthermore, although we found evidence of linear correlation between Mean and Variance, we find considerable residual curvature in log-log space (Fig. 4.8A). Additionally, the residuals for linear regression of Var~Mean are heteroskedastic, suggesting that a strictly log-log linearized model may be inappropriate. Using LOESS regression (Fig. 4.8A), we find that a non-linear relationship may more appropriate. However, without a specific mechanistic hypothesis of this relationship, candidate functional forms are uncertain. Nevertheless, the correlation and non-linear relationship are likely a property of the underlying generative process. Therefore, this finding may enable future model discrimination.

4.6 Demonstration of multi-week relaxation of expression following Tat excision by Cre recombinase

To ascertain whether we were able to efficiently excise Tat from fCLGITCh clones with Cre recombinase, we treated two clones with titrated amounts Ub-mCherry-Cre, resulting in 10-80% of cells infected. Following infection, we assayed GFP fluorescence weekly for three weeks. Surprisingly, we observed slow relaxation kinetics with apparent relaxation not observable for two to three weeks (Fig. 4.9). While relaxation of the distribution is observed, the resulting three-week distributions remain quite wide. Together, these observations may suggest that Cre recombinase is inefficiently excising Tat or that molecular memory remains after Tat is excised. Tat, by recruiting host cell chromatin modifying and remodeling complexes[1,28] to the LTR, may induce significant chromatin changes that persist long after Tat is removed. To distinguish between these possibilities we are investigating the use of a Tat-directed TALEN[29]. Slow relaxation following TALEN mediated

disabling of Tat would warrant further investigation of the molecular memory hypothesis.

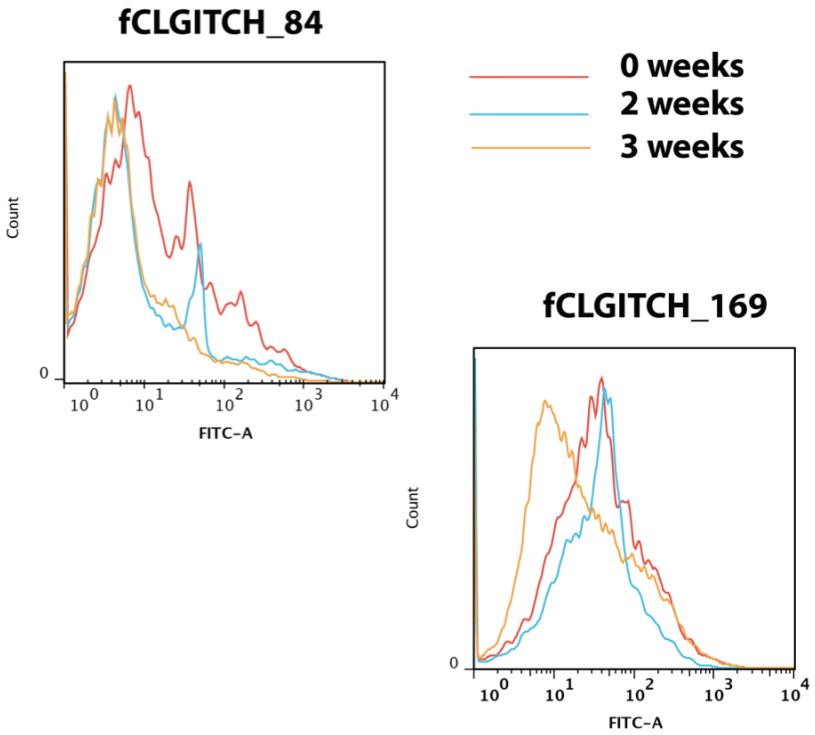


Figure 4.9| Slow multi-week clonal GFP relaxation following Cre introduction: Two fCLGITCh clones (84 and 169) were infected with a high MOI of Ub-mCherry-Cre virus. GFP expression was assayed over a three-week period following excision. We observed slow relaxation kinetics with relaxation not apparent for 2-3 weeks. This may suggest that Cre is inefficiently excising Tat or that chromatin or bound transcription factor based molecular memory persists in the absence of Tat.

4.7 Discussion

The expression output of Tat feedback circuit highly varies across integration positions. This presents significant challenges to the identification and systematic fitting of a model that can explain the observed variable dynamics. While we have not yet achieved our ultimate goal of ascertaining such a systematic model, our results here represent significant steps toward that aim. Through clustering, moment analysis, and development of an experimental system to relate basal noise to feedback dynamics, we provide a novel experimental and analytical framework that can be used to infer a novel kinetic mechanism.

Previous efforts to develop models of Tat feedback have limited their scope to bimodal clones[7,10]. These models do not generally explain the variation in Tat feedback operation across integration positions. The barriers to development of a general model have included the high degree of distribution shape heterogeneity. To reduce the size of the distribution space that is needed to infer a likely model, we

applied a clustering technique that established shape based distances between distributions. This method resolved high quality clusters that compactly describe the variation in distribution shape observed as 30 classes. Furthermore, we demonstrated the analytical validity of this clustering by comparing moment-based analysis between examining of all clones versus clusters. Specifically, we demonstrate that the relationships inferred from individual clones are statistically similar. We conjecture that by reducing the effective size of the distribution shape space, this approach will greatly simplify future modeling efforts.

Our moment based analysis revealed fundamental scaling relationships between Mean and Variance, and between Skew and Kurtosis, that can be used to discriminate between model candidates. In particular, the highly significant power-law relationship between Skew and Kurtosis is intriguing. A recent study[30] of earthquake power has resolved similar K-S power-law relationships. However, it is unclear how the dynamics of this system relates to Tat feedback. Stochastic models of gene expression (such as the two-state Markovian model used in Chapter 2), have associated probability density functions (*pdf*). Importantly, such *pdf* typically exhibit characteristic relationships between central moments. While we did not ascertain a likely mechanism suggested by the resolving scaling relationships, our resolution of these relationships is an important step toward identification of a suitable model.

An additional barrier to the development of a general model of Tat feedback has been the inability to relate basal transcriptional noise to Tat feedback distributions. Here, we develop an experimental system that excises Tat from the proviral minimal feedback circuit. In this manner, the feedback distribution can be compared to the basal distribution while controlling for the confounding effects of integration position. We demonstrate relaxation of expression following Tat excision through exogenous expression of Cre recombinase. Due to significant technical challenges, this system has not yet been applied at a large scale. However, with future application to a full subsample of clones from each of the 30 clusters, this system will likely be instrumental in overcoming the long-standing uncertainty in relating basal expression to Tat feedback expression.

4.8 Materials and Methods

4.8.1 Viral cloning

Tat was fused to mCherry through splice-overlap PCR and cloned into HIV CLGIT[7] as a BstXI-XhoI fragment to generate HIV CLGTICh. Subsequently, PCR fragments of GFP and Tat-mCherry containing terminal loxP sites were generated. GFP-loxP was inserted into HIV-CLGITCh as a BamH1-SpeI fragment using partial SpeI digestion. Tat-mCh-loxP was inserted as a BstXI-XhoI. This generated the CLG-loxP-IRES-Tat-mCherry-loxp or fCLGITCh vector analyzed in this study.

pFU-mCherry-Cre was generated through insertion of an mCherry-Cre PCR product from pLM-CMV-R-Cre[31] as an XbaI-XhoI fragment into pFU.

4.8.2 Cell Culture

Jurkat cells, used for creating clonal LGM2 cell-lines and HEK293T cells, used for packaging virus were cultured in RPMI 1640 and Isocove's DMEM, respectively. Cells were maintained at 37^0C and 5% CO_2 with the cell-media supplemented with 10% fetal bovine serum (FBS) and 100 U/mL of penicillin-streptomycin.

4.8.3 Viral Harvesting, Titering and Infections

To package the fCLGITCh and pFU-mCherry-Cre constructs, HEK293T cells were transfected with 10μg of the plasmid along with the helper plasmids pMDLg/pRRE, pcDNA3 IVS VSV-G and pRSV-Rev as described previously[32] and harvested. Harvested lentivirus was concentrated by ultracentrifugation to yield between 10^7 and 10^8 infectious units/ml. To titer, 10^5 Jurkat cells per well, were infected with a range of concentrated virus doses and six days post infection, gene expression was stimulated with 20 ng/ml tumor necrosis factor-α (TNF-α, Sigma-Aldrich). After stimulation for 18 hours, GFP expression was measured by flow cytometry, and titering curves were constructed by determining the percentages of cells that exhibited GFP fluorescence greater than background levels.

4.8.3 Clone Generation

Assuming a Poisson distribution, the well with 5% GFP infected cells was selected for expansion. This corresponds with a low MOI of ~0.05 and as previously demonstrated[7] ensures at most one integration event per infected cell within the population. The selected population was expanded for 6 days and stimulated with TNF-α as described above. Approximately 0.1*10^6 cells were sorted from the GFP^+ population on a DAKO-Cytomation MoFlo Sorter. The resulting population, which represents a polyclonal ensemble of single-integration clones, was expanded for 7 days. Single cell clones were sorted from the 'Mid' gate[9] of the bimodal bulk distribution into multi-well plates and cultured for 3-4 weeks to facilitate expansion.

4.8.5 Flow Cytometry

GFP and mCherry expression of expanded clonal cultures was assessed by flow cytometry using a BD LSRFortessa analytical cytometer. Flow cytometry data was processed using Bioconductor (Gentleman 2004) packages in custom R scripts.

4.8.6 Clustering and Statistical Analysis

All statistical analysis was performed in R using custom scripts. Clustering was performed using the GMD[27]and hclust (base R "stats") packages. Visualization was performed using lattice and ggplot2 packages.

4.9 References

1. Mahmoudi, T. *et al.* The SWI/SNF chromatin-remodeling complex is a cofactor for Tat transactivation of the HIV promoter. *The Journal of biological chemistry* **281**, 19960–19968 (2006).
2. Agbottah, E., Deng, L., Dannenberg, L. O., Pumfery, A. & Kashanchi, F. Effect of SWI/SNF chromatin remodeling complex on HIV-1 Tat activated transcription. *Retrovirology* **3**, 48 (2006).
3. Brady, J. & Kashanchi, F. Tat gets the "green" light on transcription initiation. *Retrovirology* **2**, 69 (2005).
4. Sedore, S. C. *et al.* Manipulation of P-TEFb control machinery by HIV: recruitment of P-TEFb from the large form by Tat and binding of HEXIM1 to TAR. *Nucleic acids research* **35**, 4347–4358 (2007).
5. Barboric, M. *et al.* Tat competes with HEXIM1 to increase the active pool of P-TEFb for HIV-1 transcription. *Nucleic acids research* **35**, 2003–2012 (2007).
6. Zhu, Y. *et al.* Transcription elongation factor P-TEFb is required for HIV-1 tat transactivation in vitro. *Genes Dev* **11**, 2622–2632 (1997).
7. Weinberger, L. S., Burnett, J. C., Toettcher, J. E., Arkin, A. P. & Schaffer, D. V Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell* **122**, 169–182 (2005).
8. Weinberger, L. S., Dar, R. D. & Simpson, M. L. Transient-mediated fate determination in a transcriptional circuit of HIV. *Nature genetics* **40**, 466–470 (2008).
9. Burnett, J. C., Miller-Jensen, K., Shah, P. S., Arkin, A. P. & Schaffer, D. V Control of stochastic gene expression by host factors at the HIV promoter. *PLoS Pathog* **5**, e1000260 (2009).
10. Weinberger, L. S. & Shenk, T. An HIV feedback resistor: auto-regulatory circuit deactivator and noise buffer. *PLoS Biology* **5**, e9 (2007).
11. Artyomov, M. N., Mathur, M., Samoilov, M. S. & Chakraborty, A. K. Stochastic bimodalities in deterministically monostable reversible chemical networks due to network topology reduction. *The Journal of chemical physics* **131**, 195103 (2009).
12. To, T.-L. & Maheshri, N. Noise can induce bimodality in positive transcriptional feedback loops without bistability. *Science (New York, N.Y.)* **327**, 1142–5 (2010).
13. Sneppen, K., Micheelsen, M. A. & Dodd, I. B. Ultrasensitive gene regulation by positive feedback loops in nucleosome modification. *Molecular systems biology* **4**, 182 (2008).
14. Dodd, I. B., Micheelsen, M. A., Sneppen, K. & Thon, G. Theoretical analysis of epigenetic cell memory by nucleosome modification. *Cell* **129**, 813–822 (2007).
15. Lam, F. H., Steger, D. J., O'Shea, E. K. & O'shea, E. K. Chromatin decouples promoter threshold from dynamic range. *Nature* **453**, 246–250 (2008).
16. Sneppen, K. & Dodd, I. B. A simple histone code opens many paths to epigenetics. *PLoS computational biology* **8**, e1002643 (2012).
17. Skupsky, R., Burnett, J. C., Foley, J. E., Schaffer, D. V & Arkin, A. P. HIV promoter integration site primarily modulates transcriptional burst size rather than frequency. *PLoS computational biology* **6**, 14 (2010).

18. 2012 UNAIDS Report on the Global AIDS Epidemic. at <http://www.unaids.org/en/resources/publications/2012/name,76121,en.asp>
19. Pomerantz, R. J., Bagasra, O. & Baltimore, D. Cellular latency of human immunodeficiency virus type 1. *Curr Opin Immunol* **4**, 475–480 (1992).
20. Blankson, J. N., Persaud, D. & Siliciano, R. F. The challenge of viral reservoirs in HIV-1 infection. *Annual review of medicine* **53**, 557–593 (2002).
21. Lassen, K., Han, Y., Zhou, Y., Siliciano, J. & Siliciano, R. F. The multifactorial nature of HIV-1 latency. *Trends in molecular medicine* **10**, 525–531 (2004).
22. Mizuguchi, H., Xu, Z., Ishii-Watabe, A., Uchida, E. & Hayakawa, T. IRES-dependent second gene expression is significantly lower than cap-dependent first gene expression in a bicistronic vector. *Molecular therapy : the journal of the American Society of Gene Therapy* **1**, 376–82 (2000).
23. Ji, J. P. & Loeb, L. A. Fidelity of HIV-1 reverse transcriptase copying RNA in vitro. *Biochemistry* **31**, 954–958 (1992).
24. Preston, B., Poiesz, B. & Loeb, L. Fidelity of HIV-1 reverse transcriptase. *Science* **242**, 1168–1171 (1988).
25. Yossi Rubner, C. T. The earth mover's distance as a metric for image retrieval. at <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.79.151>
26. Applegate, D., Dasu, T., Krishnan, S. & Urbanek, S. Unsupervised clustering of multidimensional distributions using earth mover distance. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11* 636 (2011).doi:10.1145/2020408.2020508
27. Zhao, X., Valen, E., Parker, B. J. & Sandelin, A. Systematic clustering of transcription start site landscapes. *PloS one* **6**, e23409 (2011).
28. He, G. & Margolis, D. M. Counterregulation of chromatin deacetylation and histone deacetylase occupancy at the integrated promoter of human immunodeficiency virus type 1 (HIV-1) by the HIV-1 repressor YY1 and HIV-1 activator Tat. *Molecular and cellular biology* **22**, 2965–2973 (2002).
29. Bedell, V. M. *et al.* In vivo genome editing using a high-efficiency TALEN system. *Nature* **491**, 114–8 (2012).
30. Cristelli, M., Zaccaria, A. & Pietronero, L. Universal relation between skewness and kurtosis in complex dynamics. *Physical Review E* **85**, 066108 (2012).
31. Papapetrou, E. P. *et al.* Genomic safe harbors permit high β-globin transgene expression in thalassemia induced pluripotent stem cells. *Nature biotechnology* **29**, 73–8 (2011).
32. Dull, T. *et al.* A third-generation lentivirus vector with a conditional packaging system. *Journal of virology* **72**, 8463–71 (1998).

Chapter 5: Modulation of expression noise by RNA export

5.1 Introduction

While the study of the transcriptional mechanisms underlying noisy gene expression has received intense interest[1–11], with analysis spanning from prokaryote to fly and mammalian systems, there has been little exploration of the roles of post-transcriptional events in modulating gene expression noise. In particular, the additional regulatory complexity of eukaryotic gene expression can entail numerous post-transcriptional step including RNA processing[12], RNA export[13], regulation by interfering and micro RNA's[14,15], translation, and ultimately degradation. Each of these processes involves chemical reactions that could impact the temporal probability of nascent mRNA being translated. By modulating this probability, post-transcriptional processes may serve to reshape protein distribution relative the mRNA distribution and thereby alter phenotypic outcomes. In particular, many genetic networks involved in multi-cellular development map protein levels to cellular decisions and phenotypic outcomes. Heterogeneity in protein levels from cell to cell can serve to determine the probability of cells committing to a particular phenotype and thus generate phenotypic diversity without explicit instruction from environmental signals [11,16,17]. Therefore, processes that modulate protein heterogeneity could have dramatic phenotypic consequences in the context of developmental circuits.

While models of stochastic gene expression that have been applied to explain observed protein and RNA distributions explicitly ignore many post-transcriptional processes, this does not discount their potential roles[18]. Mathematical models have no guarantee of uniqueness and frequently abstract underlying phenomena as lumped parameters.. Apart from some recent theoretical examinations of mRNA degradation[19] and export[20], the contributions of post-transcriptional mechanisms have primarily been overlooked. This stems in part from the lack of tools to experimentally isolate and study the effects of these processes. However, the recent development of single molecule labeling of RNA's in live[21,22] and fixed cell populations[23,24] allow tracking or localization of individual RNA. Studies of mRNA transport[13,25–27] have reported widely different kinetic rates for mRNA export from the nucleus. While such differences could result from differences in nuclear architecture and organization[28–30], it suggests that rates of export may be a regulated feature of cell biology. Importantly, the consequences of these different rates on expression noise are poorly understood.

Here, by analyzing the localization of mRNA in clonal HIV LGM2 (see Chapter 2) populations using smFISH, we observe nuclear retention of mRNA, suggesting slow RNA transport from the site of transcription through the nuclear pore. Furthermore, we observe that in clonal populations that the cytoplasmic mRNA distribution is often less noisy than the nuclear distribution, suggesting that slow mRNA transport may buffer the cytoplasm from noise arising from infrequent promoter transition and transcriptional bursts. Based on these observations, we hypothesized that a slow rate of RNA export relative to transcriptional bursting

could function as a low pass filter and thereby buffer the cytoplasm from the torrent of excursions away from the mean due to bursting and the full force of transcriptional noise. To investigate this hypothesis, we performed stochastic simulations to explore the multi-parameter regime in which this hypothesis may hold and develop an experimental system to modulate the effective rate of RNA export.

5.2 Nuclear retention of transcripts suggests slow export

Infrequent transcriptional bursts result in sharp increases in the nuclear RNA copy number, with average burst sizes ranging from 25-150% of clonal population means (see Chapter 2). We observed that across a number of clones with and without visual evidence of active or recent transcription, there appeared to be substantial retention of mRNA within the nucleus (Fig. 5.1). This suggested that the rate of transport of mRNA from the site of transcription to the cytoplasm may be slow. We hypothesized that this slow export may result from the lack of splicing in our model system. Recent studies have demonstrated the coupling of transcription, end processing, pre-mRNA splicing, and mRNA export[31–33]. The assembly of messenger ribonucleoprotein (mRNP) complexes through these coordinated processes may enable cells to regulate the rate of mRNA export.

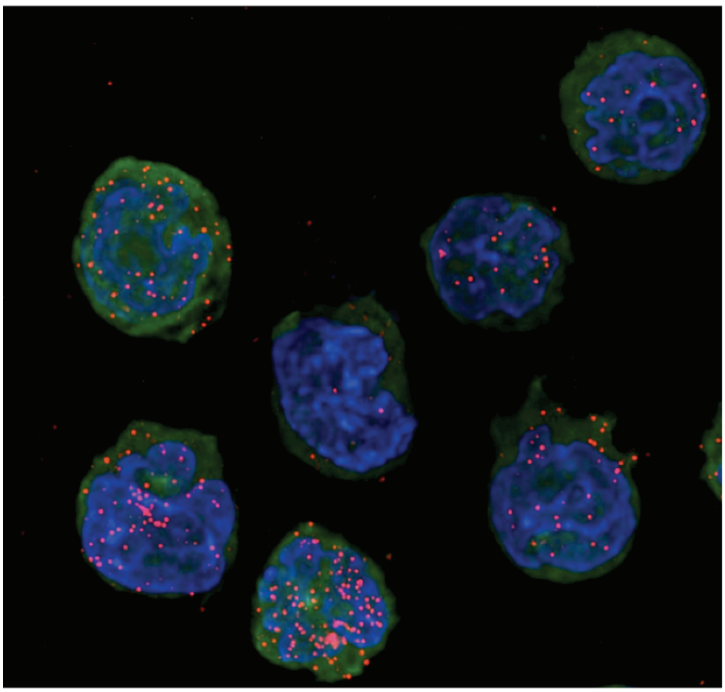


Fig. 5.1| Nuclear retention of mRNA suggests slow export: Individual mRNA within LGM2 clone BC6 were labeled via smFISH, with nuclei counterstained with DAPI. We observed apparent retention of mRNA in the nucleus in cells without active transcription (bright foci in lower left cells). This suggested the transcripts remain in the nucleus for long time periods after being transcribed. We hypothesized that this may be due to slow mRNA export.

5.3 Comparison of mean and noise of localized mRNA suggest slow export may buffer cytoplasm from noisy transcriptional output

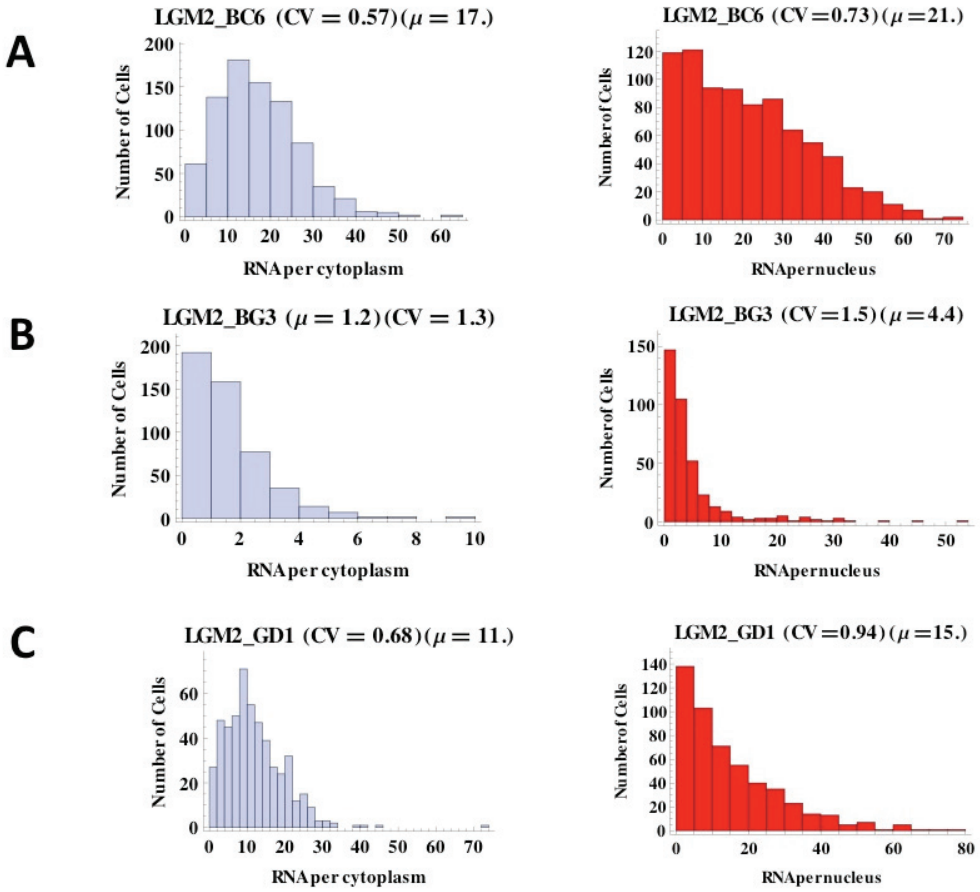


Fig. 5.2| Cytoplasmic distributions of mRNA copy number exhibit lower mean and noise than nuclear distributions: To investigate the potential role of slow mRNA export in determining expression noise, we examined the localization of mRNA in three LGM2 clones BG6(A), BG3(B) and GD1(C). In all three clones we found that the nuclear compartment exhibited a higher mean and CV or noise than the cytoplasmic compartment. This suggested that slow mRNA export could function to buffer the cytoplasm from noisy transcriptional bursting. Specifically, first-order chemical reactions can function as low-pass filters and reshape output relative to the input signal. Therefore, we hypothesized that slow export, by slowly transferring the product of bursts to the cytoplasm, could reduce cytoplasmic noise

Following our qualitative observation of nuclear mRNA retention in LGM2 clones, we developed a localization algorithm to determine the localization to either the nucleus or cytoplasm. Briefly, the nucleus is segmented from the cell volume through morphological processing of the DAPI channel of the input stack of images. Interior holes are filled in and the nuclear boundary determined through morphological opening with an elliptical structuring element. This generates a stack of binary masks that are then used to mask FISH objects and determine their appropriate assignment to either the nucleus or cytoplasm. Using this approach, we determined the distribution of LGM2 mRNA in the nucleus and cytoplasm within three clones (Fig. 5.2). Interestingly, within these three clones we found that the nuclear compartment exhibited higher mean and CV or noise relative to the cytoplasm. Furthermore, we found that the nuclear distributions are highly skewed with long tails, whereas cytoplasmic distributions are less skewed. This suggests a reshaping or filtering of the noisy output of transcriptional bursting. We reasoned that first-order processes like mRNA export could function as low-pass filters and remove higher frequency components from input signals. Furthermore, we hypothesized that slow RNA export, by functioning as a low-pass filter, could buffer the cytoplasm from transcriptional noise. Through this mechanism, the degree of buffering would be determined by the rate of export, with slower rates resulting in stronger filtering and greater buffering.

5.4 Stochastic simulations demonstrate slow export can reduce cytoplasmic noise through filtering of transcriptional bursting

Following our preliminary observations, we used stochastic simulations to further investigate our hypothesis of noise filtering by slow nuclear export. Specifically, we extended the two-state gene expression model discussed in Chapter 2 by explicitly modeling nuclear export as a first-order kinetic process. This assumes that linear transport kinetics without explicit consideration of queuing effects. We simulated this model using the Gillespie SSA[34]. Initially, we sought to isolate the effects of the rate of RNA export on cytoplasmic RNA noise. Therefore, we kept the promoter on and off rates, burst size, and rate of RNA degradation constant. We used values similar to those inferred from clones in Chapter 2 with a burst size of 20 and promoter off rate of 0.3 hr^{-1}. We simulated mRNA distributions over physiologically relevant[26] rates of RNA export expressed as multiples of the rate of RNA degradation ranging of 0.1 to 5. From averaging over 10,000 independent realizations, we find that cytoplasmic CV is an approximately hyperbolic function of the relative rate of RNA export (Fig. 5.3A), with the 50% rise occurring where the rate of export is equal to the rate of degradation and saturation occurring near 5 fold the degradation rate. We find that over a 20-fold increase in RNA export relative to RNA degradation, cytoplasmic CV increases by 200%. Consistent with an export mediated modulation of cytoplasmic noise, we find that the fraction of total mRNA that are cytoplasmic is highly correlated to Cytoplasmic CV (Fig. 5.3B).

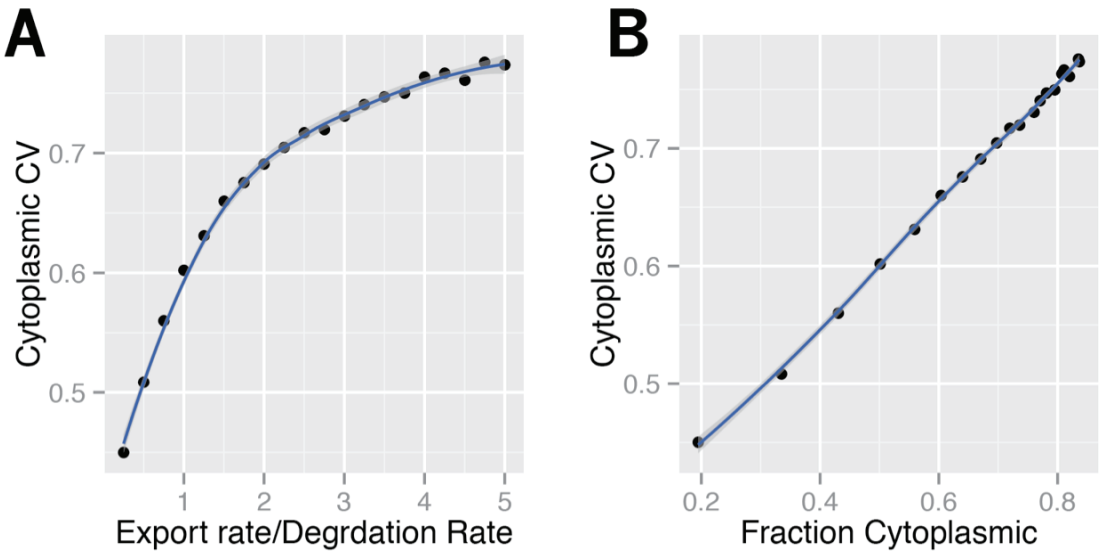


Fig. 5.3| Stochastic simulations demonstrate buffering of cytoplasmic noise by slow RNA export:(A) We extended the basic two-state mode of stochastic gene expression to explicitly include nuclear and cytoplasmic compartments with RNA transport modeled as a first-order kinetic process. We explored the effect of rates of RNA export ranging from 0.25-5 fold the rate of RNA degradation. All other model parameters were kept constant to values similar to those inferred in Chapter 2. We found that rates of RNA export less than 150% of the RNA degradation rate yield Cytoplasmic CV less than 80% of the saturating value with fast RNA export (~ 5-fold the rate of RNA degradation). Rates of export less than 2-4 fold lower than degradation yield significantly lower (62.5-40% of saturating CV) cytoplasmic noise. (B) Consistent with an export mediated buffering of cytoplasmic noise, we found that marginal (with respect to the rate of RNA export) fraction of RNA that are localized in the cytoplasm is highly correlated to the marginal cytoplasmic CV. Each point in the above plots represents average quantities computed from the final values of 10,000 realizations of the simulation run until stable values for distribution mean and variance were reached.

Our simulation results suggested that slow RNA export could buffer cytoplasmic noise. To investigate whether this might be due to low-pass filtering of transcriptional bursting, we examined the time-courses of individual runs of the simulation under slow (Export=0.5*Degradation) and fast (Export=5*Degadation) export rates. For slow rates of export, we found while the nuclear time courses undergo large excursions corresponding to the timing of bursts, the cytoplasmic time courses are smoother with less frequent and lower amplitude excursions. In contrast, with fast export we found that both the nucleus and cytoplasm are exposed to high frequency 'chatter'. This is consistent with filtering out of the higher frequency components in the temporal transcriptional output. Specifically, bursts represent sharp increases in the nuclear concentration of mRNA and excursions from the time-averaged mean. Viewed as a contiguous time signal, transcriptional bursts represent sharp edges in amplitude space and correspondingly consist of high-frequency components in the frequency domain.

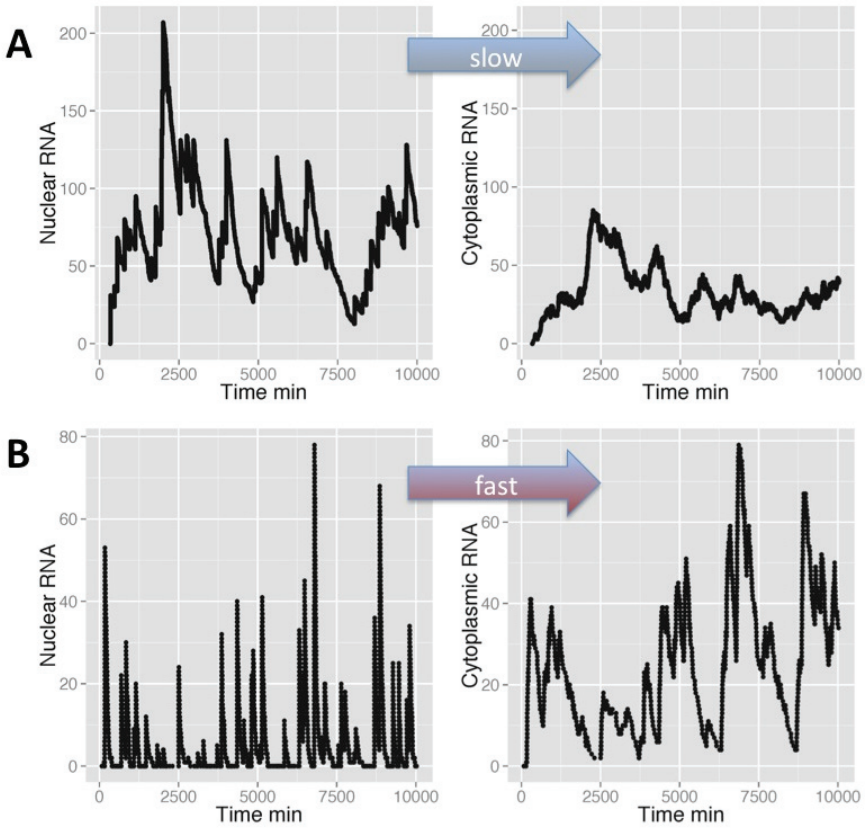


Fig. 5.4| Comparison of simulation time courses for slow and fast export suggest filtering of high frequency components arising from transcriptional bursting: We examined the time courses of RNA copy number in the Nucleus and Cytoplasm under conditions of a slow (A) rate of RNA export with $k_{export}=0.5*k_{deg}$ and a fast rate of export with $k_{export}=5*k_{deg}$.. Under conditions of slow to moderate rates of RNA export, we find that the impulse like high frequency components of the nuclear signal are filtered and smoothed by export, yielding less frequent and lower amplitude excursions in the cytoplasm. In contrast, under conditions of fast export both compartments are exposed to high frequency 'chatter'.

5.5 Differential localization suggests effective transport rates may differ across larger set of clones

While our preliminary results and simulations support our hypothesis, we sought a deeper understanding of how widespread this phenomenon is. Therefore, we examined the localization of RNA across 28 LGM2 clones that we have previously (Chapter 2 and 3 additional clones) analyzed using smFISH (Fig. 5.5). Of these we find that 18 clones exhibited ratios of Nuclear CV to Cytoplasmic CV greater than 1.1, with ratios ranging from 1.1 to 1.5. This suggests that many clones exhibit apparent buffering of cytoplasmic noise. While the buffering is somewhat modest, even modest differences in CV can exert dramatic impacts on phenotype when coupled with feedback loops. However, the observed differences in the apparent buffering phenomena across clones are unclear. We hypothesized that the promoter on frequency, the size of transcriptional bursts and the export rate together may determine a parameter regime where noise buffering occurs. Therefore, only clones

with parameterizations within this regime would exhibit buffering. Alternatively, our simulation results suggest that the ratio of export rate to degradation rate may determine the degree of buffering that occurs. Therefore, differential buffering across clones could arise from different effective rates of export between clones. Such differences could for instance result from the specific 3D location of the site of integration[30]. Sites closer on average to the nuclear envelope may exhibit faster effective rates due to a shorter mean path length[25].

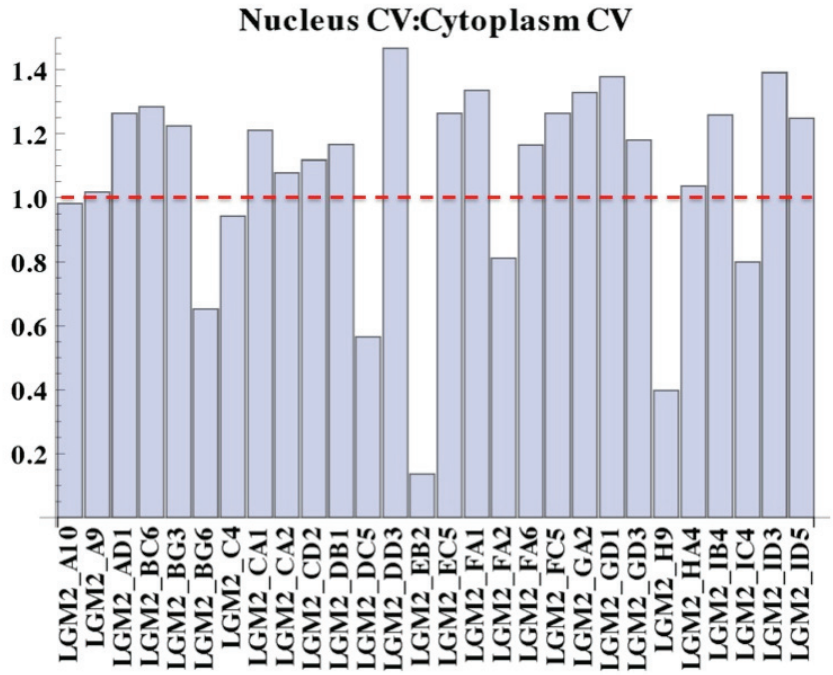


Fig. 5.5| Variable nuclear CV to cytoplasmic CV ratio across clones suggests differences in apparent buffering: To determine whether cytoplasmic CV is broadly buffered in noise relative to nuclear CV we examined the ratio of N CV to C CV across 28 LGM2 clones. 18 clones exhibit ratios larger than 1.1, which we established as the minimum biological significant difference based on studies of the sensitivity of developmental circuits to changes to noise. Clones EB2 and H9 are very low expressing clones with large a large mode of cells with 0 RNA per nuclei and few cells with 1 or 2. This inflates nuclear CV, leading to the observed depressed ratios.

We first examined whether there might be a relationship between promoter on rate and noise buffering or between burst size and noise buffering. Burst size represents the average number of transcripts produced while the promoter is in the ON state and effectively determines the amplitude of excursions in RNA copy number. Relatedly, the promoter on rate determines the frequency of such excursions. Together they determine how 'spiky' and impulse-like transcriptional output is. Therefore, we hypothesized that either or both of these kinetic parameters could couple with the rate of RNA export to determine a regime where the nuclear signal contains more high frequency components that would be strongly filtered by slow RNA export and result in buffering. Through different parameterizations, individual clones could exhibit buffering to varying degrees. To

investigate this possibility, we compared burst size and burst frequency to cytoplasmic buffering (Fig. 5.6A,B). We used the ratio of nuclear noise to cytoplasmic noise as a measure of the degree of buffering. Similarly, to ascertain whether burst size or burst frequency affected the relative distribution of RNA between the two compartments, we compared the inferred values of these parameters to the fraction of total transcripts that are cytoplasmic (Fig. 5.6C,D). While we found no specific correlation between these quantities, we found that clones with a lack of apparent buffering (N CV:C CV≤1) tend to be dim clones with burst sizes less than five (Fig. 5.6A). However, neither burst size or burst frequency is explanatory of the degree of noise buffering. Therefore, we find no evidence for some clones occupying a specific parameter regime that explains the observed differential buffering.

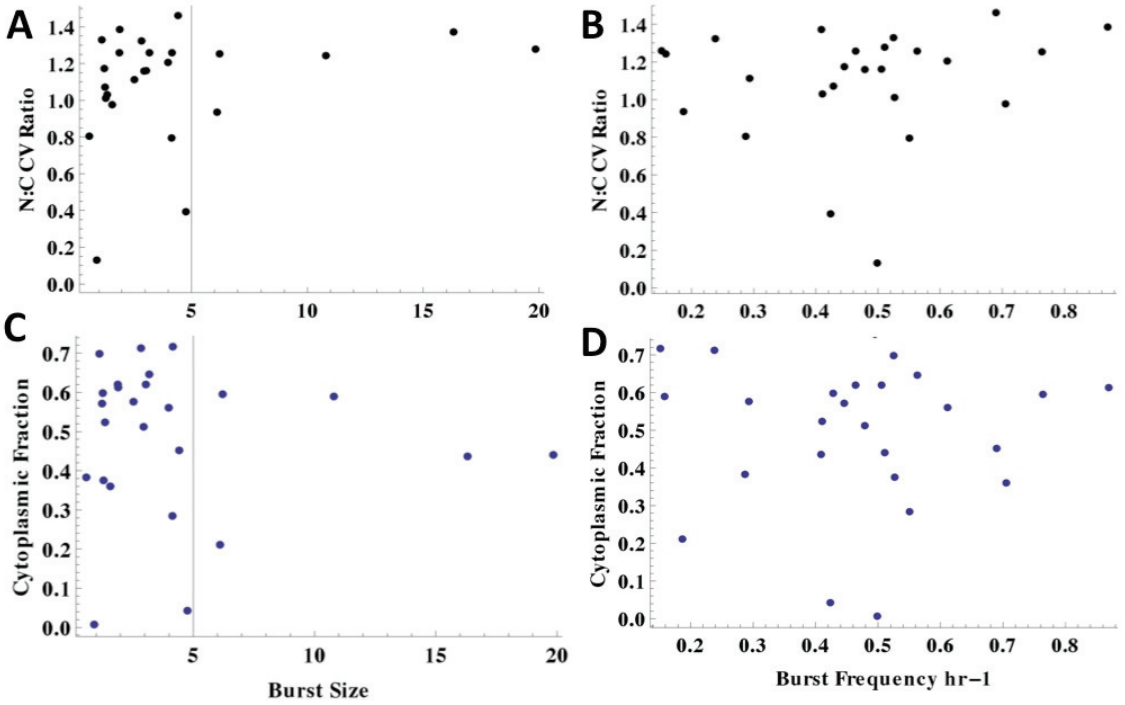


Fig. 5.6| No significant correlations between bursting parameters and apparent buffering or cytoplasmic fraction: We found no apparent relationship between burst size(A) or burst frequency (B) the ratio of compartmental CV. Similarly, we found no significant relationship between the promoter on rate ("burst frequency") and the fraction of transcripts localized in the cytoplasm (C). However, clones that exhibit N CV: C CV ratios <1 tend to be dim with burst sizes<5.

We next investigated whether different rates of export across clones could explain the observed differences in apparent buffering. Under conditions of linear first-order mRNA export, we expect the cytoplasmic fraction to vary with respect to export rate. Across the clones examined we observed a range of cytoplasmic fractions, ranging from ~10-70% of total transcripts. Interestingly, we find the cytoplasmic fraction is moderately explanatory of the nuclear CV to cytoplasmic CV

ratio (Fig.5.7A, R^2=0.66). This suggests that the nuclear compartment becomes noisier as a greater fraction of RNA is localized in the cytoplasm. As the rate of RNA export increases, we expect both the nucleus and the cytoplasm to become noisier. We expect the nucleus to become noisier because as the time-averaged mean tends towards zero bursts represent increasingly large excursions. However, contrary to this expectation we found that the cytoplasmic CV is uncorrelated to the cytoplasmic fraction. As we previously inferred in Chapter 2, each clone is specified a burst size and frequency. We would expect the cytoplasmic CV to be a function of these bursting parameters as well as the rate of RNA export. Therefore, our observation that cytoplasmic CV is uncorrelated from the cytoplasmic fraction doesn't directly imply that the rate of RNA export isn't related to cytoplasmic CV. Rather, that the interplay between bursting parameters and export needs to be further resolved. Further experiments are required to measure or infer export rates and directly investigate these relationships.

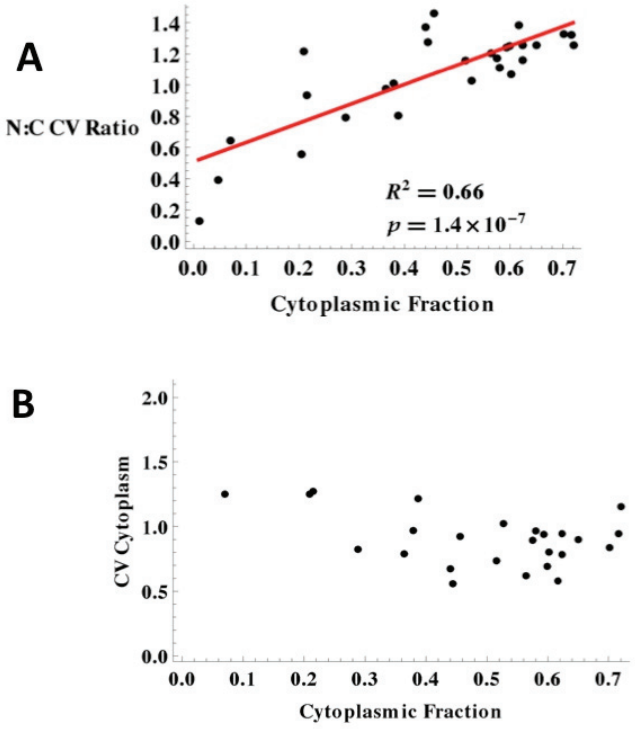


Fig.5.7| Differences in fraction of cytoplasmic transcripts suggest different transport rates: (A) We found that the fraction of transcript localized in the cytoplasm is moderately explanatory of observed variation in the N:C CV ratio with a roughly linear relationship. This is consistent with our expectation that over physiological export rates, nuclear CV would linearly with export rate while cytosplasmic CV would increase hyperbolically to saturation. Furthermore, we would expect to see differences in the cytoplasmic fraction if the export rate varied between clones (B) We find no correlation between cytoplasmic CV and the fraction of transcripts that are cytoplasmic. Cytoplasmic CV is likely determined both bursting kinetics and the export rate, both of which may vary between clones.

The relationship that we inferred between the cytoplasmic fraction and nuclear CV to cytoplasmic CV ratio suggests that there is a regime where functional

buffering may occur. We define functional buffering as a ratio between the saturating value for CV under fast export relative to the buffered CV greater than 1.1. Specifically, our simulation results suggest that functional buffering occurs when $k_{export}<2*k_{degrdation}$ (Fig. 5.3A). This corresponds to less than 60% of transcripts being localized to the cytoplasm. In this regime, we would expect that an increase in the rate of RNA export would lead to a ~10-100% increase in cytoplasmic noise. While modest, such increases in noise can significantly alter phenotypic outcomes in developmental circuits[16,35,36]. We expect that for $k_{export}>2*k_{degrdation}$, cytoplasmic CV would approach saturation, and a significant fraction of transcripts is localized in the cytoplasm. Currently, we are unable to directly infer the rate of export. However, we reasoned that clones with cytoplasmic fractions less than 60% and N:C CV ratios greater than 1.1 likely exhibit functional buffering. For these clones, we would expect that shifting nuclear transcripts into the cytoplasm through faster export would increase cytoplasmic noise. We found that 10 clones appear to occupy this regime (Fig. 5.8, red dashed box). In these clones a significant fraction (>40%) of transcripts are retained in the nucleus, suggesting slow export. Furthermore, these clones exhibit noisier nuclear compartments. Together, these suggest that export may be sufficiently slow in some clones to filter noisy transcriptional output and buffer cytoplasmic noise.

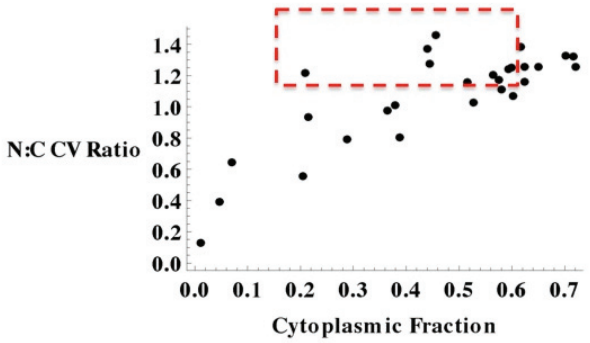


Fig.5.8| Subset of clones exhibit evidence of functional noise buffering: Our simulation results indicated that under conditions where the cytoplasmic fraction is less than 60%, the cytoplasmic CV is less than 80% of its saturating value. Furthermore, we define functional buffering where the N:C CV ratio is greater than 1.1. We found that subset of 10 clones appears to operate in this regime. We hypothesize that increasing the rate of RNA export in these clones would lead to an increase in cytoplasmic CV.

5.6 Development of experimental system to directly relate the rate of RNA export to expression noise

Across LGM2 clones we observed that a subset (10 of 28) of the clones examined exhibit evidence of slow RNA export and buffering of cytoplasmic noise. Furthermore, our simulation results suggest that slow RNA export can function as a low-pass filter to reduce the exposure of the cytoplasm to transcriptional noise arising from infrequent bursts. Additionally, previous studies[16,35–38] have shown that small changes in the noise of master transcription factors in central decision and

development circuits can have dramatic impacts on expression outcomes and the distribution of phenotypes within a population. Therefore, we sought to develop an experimental system that would enable demonstration of the modulation of cytoplasmic noise by the rate of RNA export.

Nucleocytoplasmic transport is a highly regulated process orchestrated through the interactions between cargo, protein importins and exportins, the Ran GTPAse, and the nuclear pore complex[39,40]. Through these interactions, access to the genome and the timing of the delivery of gene products are dynamically regulated. Importantly, while multiple mRNA export pathways exist[41–43], the kinetics associated with these pathways is largely unknown. Nominally, unspliced intronless mRNAs such as our model LGM2 transcripts are exported via the TAP dependent pathway. Due to the reported coupling[33,44] between splicing and export, unspliced intronless mRNA may be export inefficiently. The HIV viral factor Rev is a Ran exportin that binds a structured RNA element (Rev-Response Element, RRE) to facilitate the efficient export of unspliced (genomic) and singly spliced viral RNA via the CRM-1 exportin[45] pathway. Importantly, it has been demonstrated that Rev enhances the rate of unspliced genomic RNA 3-12 fold[46]. Conveniently, our LGM2 transcripts contain the RRE in their 5' leader, which facilitates export during viral packaging[47].

We reasoned that through titration of Rev in *trans*, we could increase the effective rate of LGM2 export and thereby increase cytoplasmic noise. Therefore, we developed a lentiviral vector that delivers Rev with a mCherry reporter constitutively expressed by the Ubiquitin promoter (HIV CS Ub-Rev-IRES-mCherry, Ub-RIM). Unfortunately, while we have titrated Ub-RIM into a number of LGM2 clones, we have faced technical challenges that have prevented full validation of this approach. Specifically, we have found that LGM2 clones analyzed here and in Chapter 2 are not stable across freeze-thaw cycles, with frequent occurrences of silencing that alter the gene expression distribution. Therefore, while we can infer clones that were likely operating in the regime of functional noise buffering, their distributions and parameterizations irrevocably change upon silencing. However, our findings suggest that clones with burst sizes larger than five and cytoplasmic fractions less than 60% (Fig.4.6 A,C) are most likely to exhibit slow export and noise buffering (N:C CV>1.1). We conjecture that this finding coupled with further simulations or model analysis will greatly aid in the selection of candidate clones for Ub-RIM titration in future studies.

5.7 Discussion

Export of mRNA from the site of transcription within the 3D volume of the nucleus to the cytoplasm where translation occurs is a fundamental process in eukaryotic cell biology. Owing to their role as the intermediate between the information storage of the genome and functional protein production, the production of mRNAs is highly regulated at every step in their lifecycle. In particular, the export of mRNA and associated proteins as mRNP is highly regulated with multiple pathways for different classes of mRNA. These pathways may regulate

the kinetics and efficiency of mRNA transport across functional classes of mRNA or on a per gene basis. The wide variation[21,25,27,28,48] in reported mRNA export rates may be a consequence of such regulation or simply arise from differences in gene positioning within the dynamic nuclear environment. Here, by leveraging the powerful localization information revealed by smFISH and computational modeling, we demonstrate that slow RNA export may function as a low-pass filter to buffer cytoplasmic noise. Furthermore, by analyzing simulations and localization across a large set of clones we indicate a likely kinetic regime where these phenomena may occur.

Cell to cell heterogeneity in gene expression under homogenous environments is an unavoidable consequence of stochastic biochemical kinetics. In the context of cell biology, such expression noise can be deleterious. Others[49–51] and we[10] (see also Chapter 2) have demonstrated that transcriptional bursting is a major source of gene expression noise. Such noise presents a barrier to deterministic cellular function and behavior. Therefore, what mechanisms may cells invoke to buffer gene expression from the noisy effects of transcriptional bursting? Eukaryotic cells have numerous post-transcriptional processes as their disposal including mRNA export[20], translation[52], mRNA degradation[18], and interaction with small regulatory RNAs[53] that could serve to shield cellular phenotypes from expression noise. The effects of such post-transcriptional processes on noisy expression are poorly understood. Here, our simulation results indicate that slow RNA export can function as a low pass first-order and buffer the cytoplasm from the 'chatter' of burst like transcription. Furthermore, we find that such buffering occurs when export is less than 2-fold the rate of RNA degradation. In this regime, slow export significantly reshapes the distribution of RNA per cytoplasm with a cytoplasmic CV less than 80% of the saturating value under fast transport. Therefore, by regulating the rate of mRNA export over specific classes of genes, cells or tissues could efficiently limit the exposure of expression to noisy transcription.

Genes are positioned along chromosomes, and chromosomes are positioned within the 3D nuclear volume, in non-random manner with potential functional consequences on the regulation gene expression[13,29,30,44,54]. Retroviruses integrate semi-randomly[55] across much of the genome. The precise 3D location of each of these sites of integration within the nuclear volume and mean distance from the nuclear pore may contribute to the effective rate of viral mRNA export. By analyzing the localization of HIV LGM2 mRNA across 28 clones, we found differences in the fraction of transcripts localized in the cytoplasm. This suggests that the rate of export may differ between clones. While this could be a result of extrinsic variation in export factors across clones, it is likely that the different sites of integration may impose differing mean path lengths to the nuclear pore. The mean path length would determine the average time mRNP take to diffuse across the nuclear volume to be captured and exported[26,48]. We found that a subset (10 of 28) of clones exhibited evidence of nuclear retention and slow export. Furthermore, we found evidence that this slow export reduces the noise of the cytoplasm relative to the nucleus. While further direct investigation of the effective rate of export is required, these findings indicate that slow export may play a biologically significant role in reducing expression noise.

Recent studies in yeast [51]and mammalian systems[50] suggest that many genes exhibit burst-like transcription. Importantly, the precise timing and size of bursts appears to differ from gene to gene. Furthermore, as we demonstrate in Chapter 2, different genomic locations appear to confer unique burst kinetics. By comparing localization and relative compartmental CV to inferred burst parameters (size and frequency) we sought to infer whether there is a regime where slow export buffers the high-frequency components of infrequent bursts. While we found no systematic relationship, we find that clones with burst sizes greater than 5 are more likely to exhibit evidence of slow export and buffering. This makes intuitive sense in that larger bursts are more impulse signal like and therefore are expected to be more strongly filtered by the low pass filtering of slow first-order export. Importantly, we cannot currently resolve whether different rates of transcription or promoter off rates underlie inferred differences in burst size between clones. Alteration of the promoter off rate across clones will modulate the lifetime of the burst and change the effective pulse width. Longer burst lifetimes with lower transcription rates may be less strongly filtered, whereas shorter burst lifetimes with higher transcription rates are more impulse like and therefore may be more strongly filtered. Further analysis of these potential effects is necessary to precisely understand over what regimes of burst kinetics slow export may function to buffer cytoplasmic noise.

Our combined experimental and computational results represent novel observations and are consistent with our hypothesis. However, further experiments using the Rev titration system we developed will be necessary to firmly establish a causal relationship. Specifically, we expect that by titrating Rev into a clone with a moderate to large burst size that exhibits significant nuclear retention of transcripts, cytoplasmic noise will be increased from 50-100%. We envision further extending these basic studies of the functional relationship between export rate and expression noise to the Oct4-Nanog fate switch. Recent studies have indicated that differences in expression noise in Oct4 underlie bimodal Nanog expression. Bimodal Nanog expression serves as critical stem-cell fate switch that bifurcates pluripotent embryonic stem cell populations between self-renewing and differentiable states. Specifically it was demonstrated that an increase in Oct4 CV of 20% is sufficient to induce excursions to the high Nanog state. Therefore, we foresee this as an ideal system to use Rev regulated expression rate as a noise rheostat to modulate the bifurcation propensity of the ESC Oct4-Nanog circuit. This would demonstrate a highly biologically significant role for RNA export in modulating expression noise and cellular decisions or phenotypic diversification.

5.8 Materials and Methods

5.8.1 Cell lines and smFISH

The generation of LGM2 clonal cells lines, culture conditions, and analysis with smFISH is described in Chapter 2.

5.8.2 Computational localization of transcripts

Each slice of the stack of images corresponding to the DAPI channel was subjected to ISODATA segmentation. Internal holes were filled using the morphological DIPImage (http://www.diplib.org) *imfill* algorithm. The edges of each slice were then smoothed using morphological opening. The resulting smoothed reconstruction of the nuclear volume was then used to mask single molecule RNA objects. The assignment of objects on the border was determined through pixel voting, where more nuclear masked pixels determined nuclear classification. This was compared to a greedy assignment. While there was minimal difference in the primary findings reported here with greedy assignment, we found the pixel-weighted algorithm to correspond best to manual assignment.

Table 5.1 Localization Statistics

clone	C fraction	CV N : CV C	<C>	<N>	CV C	CV N
LGM2_A10	0.36	0.98	3.10	5.43	0.80	0.78
LGM2_A9	0.38	1.02	1.86	3.07	0.98	0.99
LGM2_AD1	0.72	1.26	3.65	1.42	1.16	1.47
LGM2_BC6	0.44	1.28	16.70	20.99	0.57	0.73
LGM2_BG3	0.21	1.22	1.15	4.39	1.26	1.54
LGM2_BG6	0.07	0.65	1.15	15.57	1.26	0.82
LGM2_C4	0.21	0.94	2.00	7.34	1.28	1.21
LGM2_CA1	0.56	1.21	9.82	7.61	0.63	0.76
LGM2_CA2	0.60	1.08	4.28	2.84	0.81	0.87
LGM2_CD2	0.58	1.12	3.50	2.54	0.98	1.09
LGM2_DB1	0.62	1.17	3.97	2.41	0.95	1.11
LGM2_DC5	0.20	0.56	0.29	1.15	2.23	1.26
LGM2_DD3	0.46	1.47	3.56	4.26	0.93	1.37
LGM2_EB2	0.01	0.14	0.04	3.86	5.99	0.82
LGM2_EC5	0.65	1.26	4.69	2.53	0.91	1.15
LGM2_FA1	0.70	1.33	3.25	1.39	0.85	1.13
LGM2_FA2	0.39	0.81	1.37	2.17	1.22	0.99
LGM2_FA6	0.51	1.17	5.88	5.54	0.74	0.87
LGM2_FC5	0.62	1.26	4.41	2.67	0.79	1.00
LGM2_GA2	0.72	1.33	4.62	1.84	0.95	1.27
LGM2_GD1	0.44	1.38	11.42	14.59	0.68	0.94
LGM2_GD3	0.57	1.18	2.86	2.12	0.90	1.06
LGM2_H9	0.05	0.40	0.43	9.02	2.65	1.06
LGM2_HA4	0.53	1.04	2.42	2.18	1.03	1.07
LGM2_IB4	0.60	1.26	5.39	3.62	0.70	0.88
LGM2_IC4	0.29	0.80	5.47	13.58	0.83	0.67
LGM2_ID3	0.62	1.39	7.68	4.79	0.59	0.82
LGM2_ID5	0.59	1.25	8.27	5.69	0.95	1.18

5.8.3 Model and stochastic simulations

We extended the two-state model presented in Chapter 2 to explicitly model RNA transport as a linear first-order reaction with rate k_{export}. We implemented a compartmental Gillespie SSA in MATLAB. The data points of Fig.4.3 represent 10,000 runs over 2800 minutes. The final values RNA copy numbers in each compartment were recorded and the average variance and mean across simulation end-points used to compute CV for each export rate. The export rate was varied from 0.25 to 5 fold the RNA degradation rate, which was set to the experimentally determined value from Chapter 2. A burst size of 20 was used with the transcription rate set to 400 per hour with a promoter off rate of 20 per hour and promoter on rate of 0.3 per hour.

5.8.4 Data analysis and statistics

All data analysis was performed in R or Mathematica 8 using custom scripts.

5.9 References

1. Arkin, A., Ross, J. & McAdams, H. H. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. *Genetics* **149**, 1633–1648 (1998).
2. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic Gene Expression in a Single Cell. *Science* **297**, 1183–1186 (2002).
3. Shahrezaei, V. & Swain, P. S. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 17256–17261 (2008).
4. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & Van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nature genetics* **31**, 69–73 (2002).
5. Becskei, A., Kaufmann, B. B. & Van Oudenaarden, A. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nature genetics* **37**, 937–944 (2005).
6. Blake, W. *et al.* Phenotypic Consequences of Promoter-Mediated Transcriptional Noise. *Molecular cell* **24**, 853–865 (2006).
7. Kaern, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics* **6**, 451–464 (2005).
8. Samoilov, M. S., Price, G. & Arkin, A. P. From fluctuations to phenotypes: the physiology of noise. *Sci STKE* **2006**, re17 (2006).
9. Weinberger, L. S., Burnett, J. C., Toettcher, J. E., Arkin, A. P. & Schaffer, D. V Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell* **122**, 169–182 (2005).
10. Skupsky, R., Burnett, J. C., Foley, J. E., Schaffer, D. V & Arkin, A. P. HIV promoter integration site primarily modulates transcriptional burst size rather than frequency. *PLoS computational biology* **6**, 14 (2010).

11. Wernet, M. F. *et al.* Stochastic spineless expression creates the retinal mosaic for colour vision. *Nature* **440**, 174–180 (2006).
12. Wilhelm, B. T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239–1243 (2008).
13. Lionnet, T., Wu, B., Grünwald, D., Singer, R. H. & Larson, D. R. Nuclear Physics: Quantitative Single-Cell Approaches to Nuclear Organization and Gene Expression. *Cold Spring Harbor Symposia on Quantitative Biology* **75**, 113–26 (2011).
14. Cui, Q., Yu, Z., Purisima, E. O. & Wang, E. MicroRNA regulation and interspecific variation of gene expression. *Trends Genet* **23**, 372–375 (2007).
15. Ebert, M. S. & Sharp, P. A. Roles for microRNAs in conferring robustness to biological processes. *Cell* **149**, 515–24 (2012).
16. Feinerman, O., Veiga, J., Dorfman, J. R., Germain, R. N. & Altan-Bonnet, G. Variability and robustness in T cell activation from regulated heterogeneity in protein levels. *Science* **321**, 1081–1084 (2008).
17. Lewis, J. From signals to patterns: space, time, and mathematics in developmental biology. *Science* **322**, 399–403 (2008).
18. Pedraza, J. M. & Paulsson, J. Effects of molecular memory and bursting on fluctuations in gene expression. *Science (New York, NY)* **319**, 339–343 (2008).
19. Kuwahara, H. & Schwartz, R. Stochastic steady state gain in a gene expression process with mRNA degradation control. *Journal of the Royal Society, Interface / the Royal Society* **9**, 1589–98 (2012).
20. Singh, A. & Bokes, P. Consequences of mRNA transport on stochastic variability in protein levels. *Biophysical journal* **103**, 1087–96 (2012).
21. Shav-Tal, Y. *et al.* Dynamics of single mRNPs in nuclei of living cells. *Science (New York, N.Y.)* **304**, 1797–800 (2004).
22. Larson, D. R., Zenklusen, D., Wu, B., Chao, J. A. & Singer, R. H. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science (New York, N.Y.)* **332**, 475–8 (2011).
23. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA Synthesis in Mammalian Cells. *PLoS Biology* **4**, e309 (2006).
24. Raj, A., Van den Bogaard, P., Rifkin, S. A., Van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* **5**, 877–879 (2008).
25. Vargas, D. Y., Raj, A., Marras, S. A., Kramer, F. R. & Tyagi, S. Mechanism of mRNA transport in the nucleus. *Proc Natl Acad Sci U S A* **102**, 17008–17013 (2005).
26. Mor, A. *et al.* Dynamics of single mRNP nucleocytoplasmic transport and export through the nuclear pore in living cells. *Nature cell biology* **12**, 543–52 (2010).
27. Mor, A. & Shav-Tal, Y. Dynamics and kinetics of nucleo-cytoplasmic mRNA export. *Wiley interdisciplinary reviews. RNA* **1**, 388–401
28. Shav-Tal, Y., Darzacq, X. & Singer, R. H. Gene expression within a dynamic nuclear landscape. *The EMBO journal* **25**, 3469–79 (2006).
29. Kumaran, R. I., Thakar, R. & Spector, D. L. Chromatin dynamics and gene positioning. *Cell* **132**, 929–934 (2008).

30. Lanctôt, C., Cheutin, T., Cremer, M., Cavalli, G. & Cremer, T. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nature reviews. Genetics* **8**, 104–15 (2007).
31. Luo, M. J. & Reed, R. Splicing is required for rapid and efficient mRNA export in metazoans. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 14937–42 (1999).
32. Valencia, P., Dias, A. P. & Reed, R. Splicing promotes rapid and efficient mRNA export in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 3386–91 (2008).
33. Reed, R. Coupling transcription, splicing and mRNA export. *Current opinion in cell biology* **15**, 326–31 (2003).
34. GILLESPIE, D. T. EXACT STOCHASTIC SIMULATION OF COUPLED CHEMICAL-REACTIONS. *J Phys Chem-us* **81**, 2340–2361 (1977).
35. Colman-Lerner, A. *et al.* Regulated cell-to-cell variation in a cell-fate decision system. *Nature* **437**, 699–706 (2005).
36. Kalmar, T. *et al.* Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS biology* **7**, e1000149 (2009).
37. Samoilov, M. S. & Arkin, A. P. Deviant effects in molecular reaction pathways. *Nature biotechnology* **24**, 1235–1240 (2006).
38. Artyomov, M. N., Das, J., Kardar, M. & Chakraborty, A. K. Purely stochastic binary decisions in cell signaling models without underlying deterministic bistabilities. *Proc Natl Acad Sci U S A* **104**, 18958–18963 (2007).
39. Weis, K. Regulating access to the genome: nucleocytoplasmic transport throughout the cell cycle. *Cell* **112**, 441–51 (2003).
40. Cook, A., Bono, F., Jinek, M. & Conti, E. Structural biology of nucleocytoplasmic transport. *Annual review of biochemistry* **76**, 647–71 (2007).
41. Cullen, B. R. Nuclear RNA Export Pathways. *Molecular and Cellular Biology* **20**, 4181–4187 (2000).
42. Cullen, B. R. Nuclear RNA export. *Journal of cell science* **116**, 587–97 (2003).
43. Jarmolowski, A., Boelens, W. C., Izaurralde, E. & Mattaj, I. W. Nuclear export of different classes of RNA is mediated by specific factors. *The Journal of cell biology* **124**, 627–35 (1994).
44. Komili, S. & Silver, P. A. Coupling and coordination in gene expression processes: a systems biology view. *Nat Rev Genet* **9**, 38–48 (2008).
45. Hutten, S. & Kehlenbach, R. H. CRM1-mediated nuclear export: to the pore and beyond. *Trends in cell biology* **17**, 193–201 (2007).
46. Blissenbach, M., Grewe, B., Hoffmann, B., Brandt, S. & Uberla, K. Nuclear RNA export and packaging functions of HIV-1 Rev revisited. *Journal of virology* **84**, 6598–604 (2010).
47. Dull, T. *et al.* A third-generation lentivirus vector with a conditional packaging system. *Journal of virology* **72**, 8463–71 (1998).
48. Vargas, D. Y., Raj, A., Marras, S. A. E., Kramer, F. R. & Tyagi, S. Mechanism of mRNA transport in the nucleus. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 17008–13 (2005).

49. Singh, A., Razooky, B., Cox, C. D., Simpson, M. L. & Weinberger, L. S. Transcriptional bursting from the HIV-1 promoter is a significant source of stochastic noise in HIV-1 gene expression. *Biophysical journal* **98**, L32–4 (2010).
50. Suter, D. M. *et al.* Mammalian genes are transcribed with widely different bursting kinetics. *Science (New York, N.Y.)* **332**, 472–4 (2011).
51. Zenklusen, D., Larson, D. R. & Singer, R. H. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural & Molecular Biology* **15**, 1263–1271 (2008).
52. Komorowski, M., Miekisz, J. & Kierzek, A. M. Translational repression contributes greater noise to gene expression than transcriptional repression. *Biophys J* **96**, 372–384 (2009).
53. Levine, E., McHale, P. & Levine, H. Small regulatory RNAs may sharpen spatial expression patterns. *PLoS computational biology* **3**, e233 (2007).
54. Batada, N. N. & Hurst, L. D. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat Genet* **39**, 945–949 (2007).
55. Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C. & Bushman, F. D. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res* **17**, 1186–1194 (2007).

Appendix A

Table of Contents

I Gating of Flow Cytometry Data

Previous studies in yeast[1] have suggested a strong dependence between cell size (Forward Scatter, FSC) and noise as measured by the Coefficient of Variation (CV) or CV^2. As previously suggested[2] we find an extremely narrow gate to be unnecessary to limit the correlation between cell size (FSC) and GFP. However, to minimize any errors due to manual gating we used a data driven gating strategy provided by the *norm2h* filter of the Bioconductor flowViz package in R (Fig. A1). To ascertain whether there was any residual correlation between FSC and GFP we examined both linear (Fig. A2) and non-linear correlations and found that for all clones there is no significant correlation between FSC and GFP (Fig. A3).

Furthermore, as previously suggested[1] we do not find that size gating has any quantitative effect on the scaling between GFP mean and GFP variance or between GFP mean and GFP CV. Specifically, we examined the relationship between expression noise (CV) as a function of expression before and after gating and found that both exhibited uncorrelated relationships (Fig.A4A).Additionally, we examined whether the relationship between variance and mean was strongly modulated by gate size (Fig.A4B). Interestingly, we found that neither the regression slope (Fig.A4C) nor the extent of correlation (Fig.A4D) was strongly affected by gate size. In contrast to findings in yeast, these results suggest that cell size does not have a significant impact on either expression mean or noise.

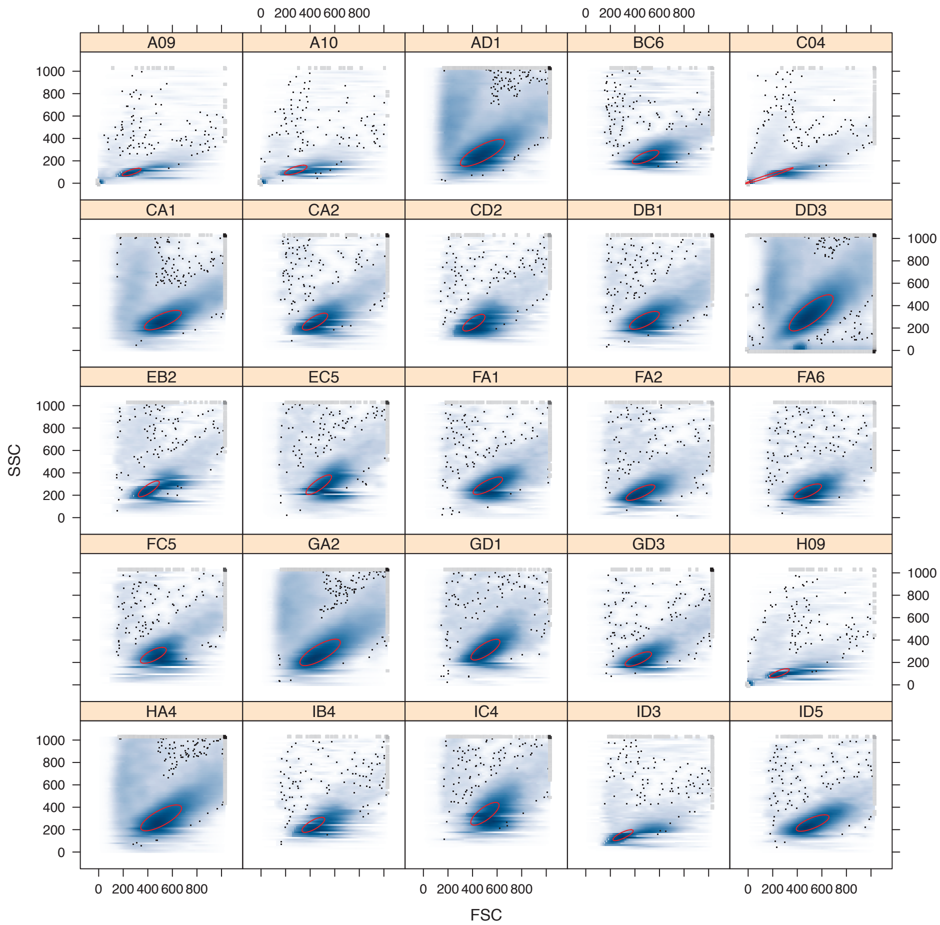


Figure A1-Representation of gating strategy for subset of clones
Raw flow cytometry channel data for all 227 LGM2 clones was gated in Side Scatter (SSC) and Forward Scatter (FSC) space using the *norm2h* filter provided by the BioCondutor flowViz package. This filter fits a 2D Gaussian around the mode of the data, which results in a small ellipse. The resulting gate for the 25 clones used for FISH analysis is depicted (red ellipse) overlaid on a smoothed FSC v SSC scatterplot.

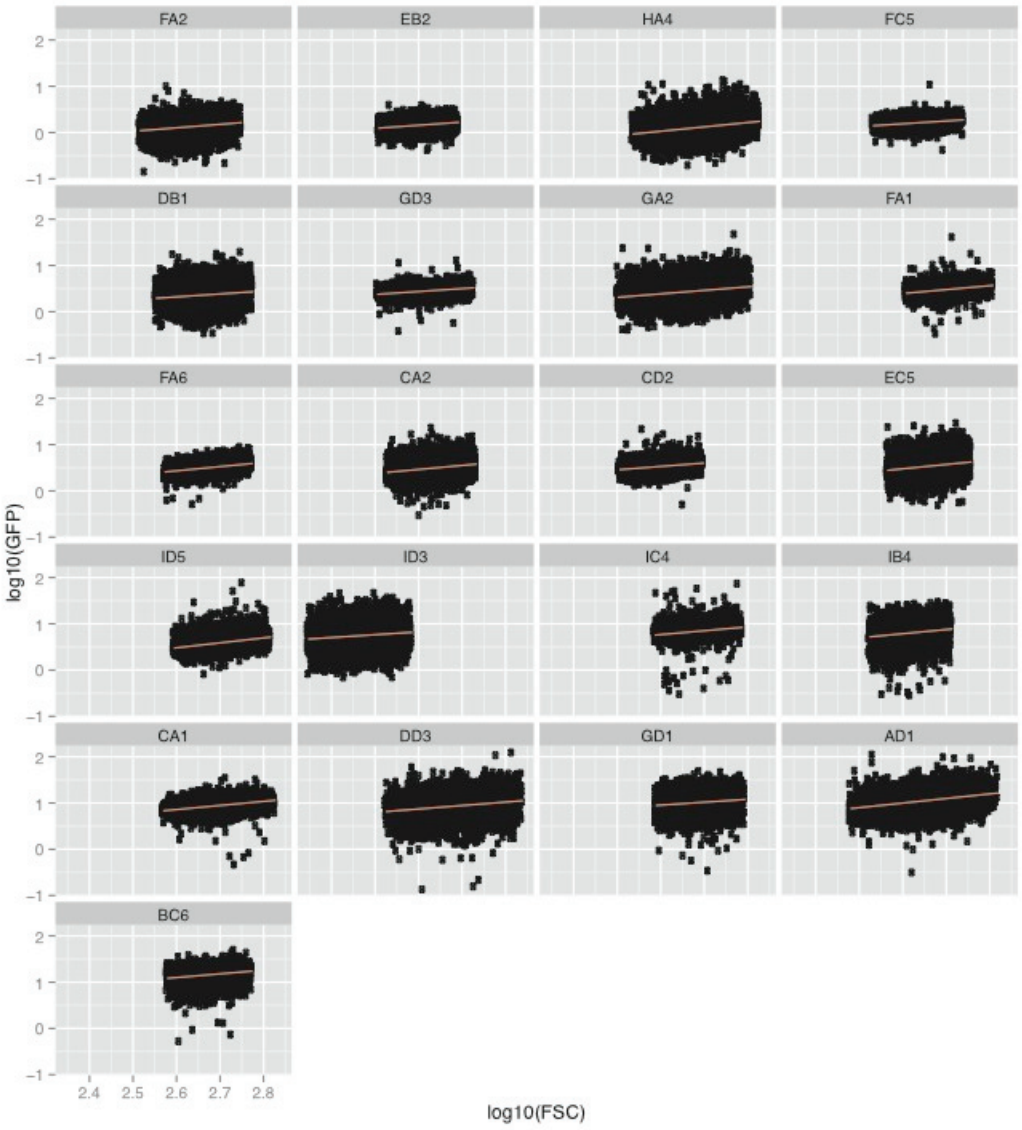


Figure A2- Cell Size and GFP are not linearly correlated post-sorting As an initial assessment of whether any relationship remained between cell size (FSC) and GFP we examined scatterplots and performed linear fits (red line) for the clones used for FISH analysis. There is no apparent relationship and all linear regressions do not have slopes significantly different from zero. Furthermore, the scatterplots suggests very little difference (the range of GFP is independent of FSC) in the CV in GFP as a function of FSC.

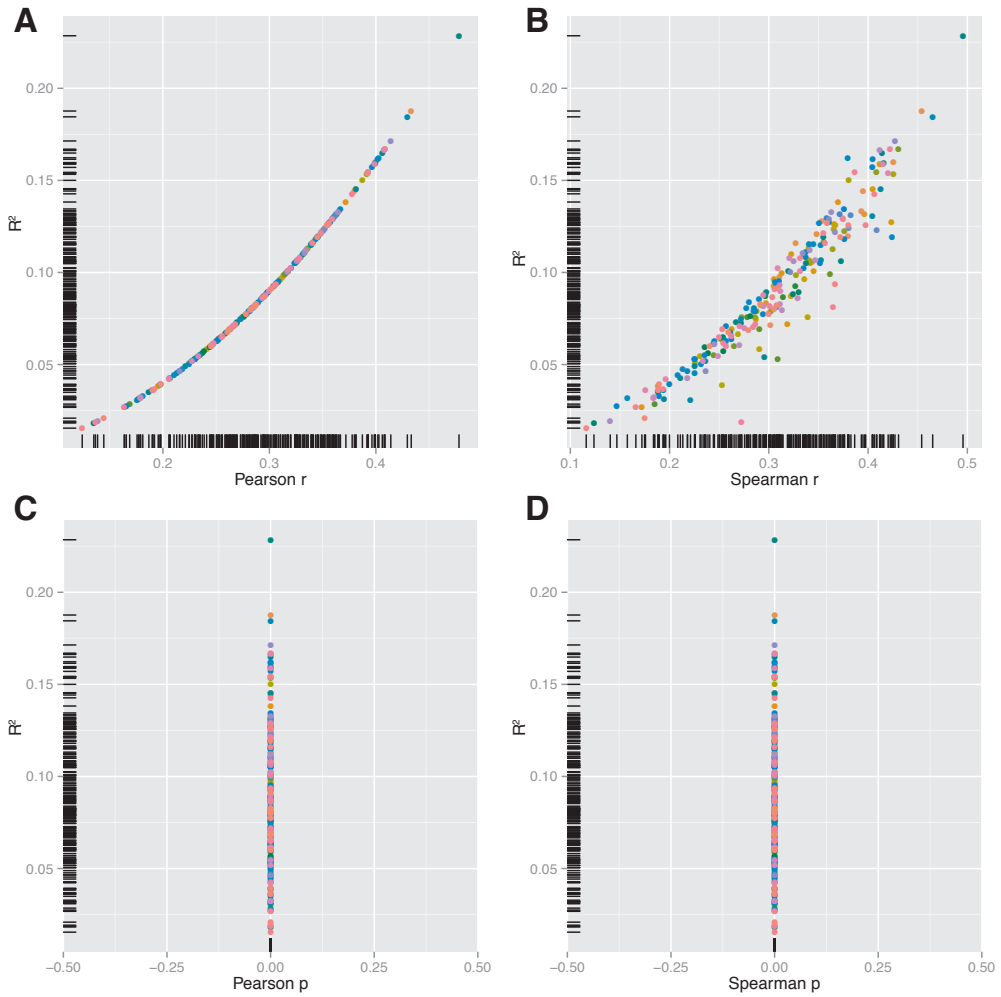


Figure A3- No significant linear or non-linear relationship between cell size and GFP expression observed across all clones: To further verify our gating strategy and analyze whether cell size correlated to GFP expression across all 227 LGM2 clones we examined both linear and non-linear correlations and performed statistical testing **(A) Minimal variance in GFP explained by FSC:** A plot of FSC v. GFP R^2 versus the Pearson correlation coefficient for each clone reveals that very little observed variance (R^2<0.2,Pearson r<0.4). The marginal rug plots indicate the relative density for the majority of clones centers around R^2~0.1 and Pearson r~0.3. **(B) Minimal non-parametric relationship between FSC and GFP:** To further address whether any monotonic relationship between FSC and GFP existed post-gating, we generated a plot of of FSC v. GFP R^2 versus the Spearman correlation coefficient. Similar to linear correlation, this plot suggests very limited monotonic relationships exist for all the clones (Spearman r<0.5). **(C,D) Minimal correlations not significantly different from zero:** Statistical testing of the correlations does not find significant support for either linear or non-linear correlations.

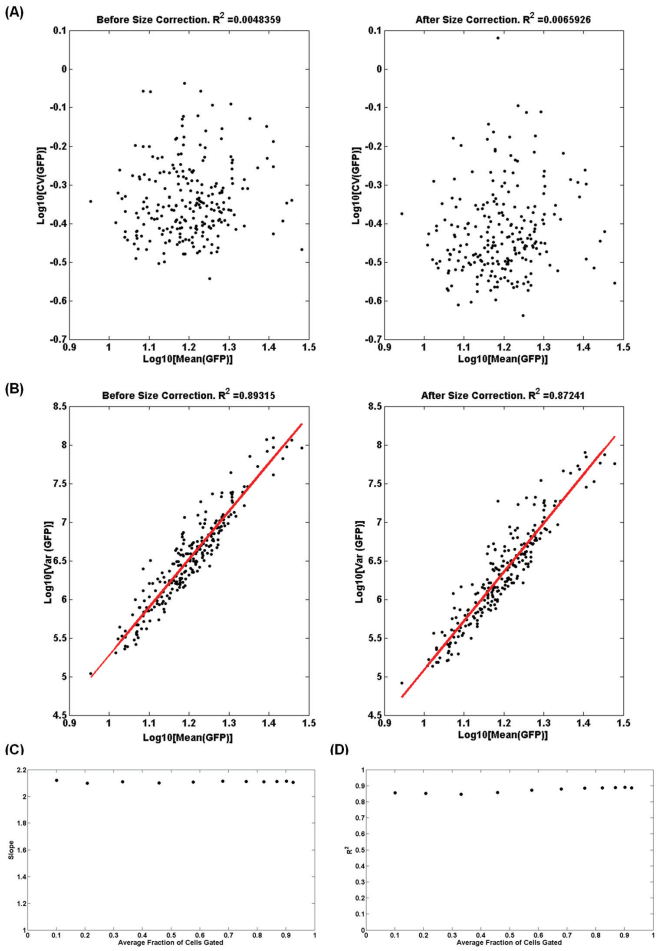


Figure A4-Cell size does not significantly underlie moment based relationships: (A) Uncorrelated relationship between mean and noise not a function of cell size: Cells were first gated using a broad gate capturing all live cells. Subsequent cell size correction by gating in a small region of SSC/FSC space does not alter the observed uncorrelated relationship between expression mean and expression noise ($R^2 \sim 0$). **(B) Observed telegraph-like scaling between mean and variance not a function of cell size:** We examined the relationship between mean and variance prior to and after cell size correction and find that both exhibit scaling not significantly different from 2 and have highly similar R^2 values. **(C,D) Inference of slope and correlation not a function of gate size:** The regression slope and correlation between variance and mean were examined following gating with various size square gates centered about the density mode in SSC/FSC space that capture from 10% to 90% of the total live cells. We do not find a significant modulation of the inferred relationship between mean and variance as a function of gate size.

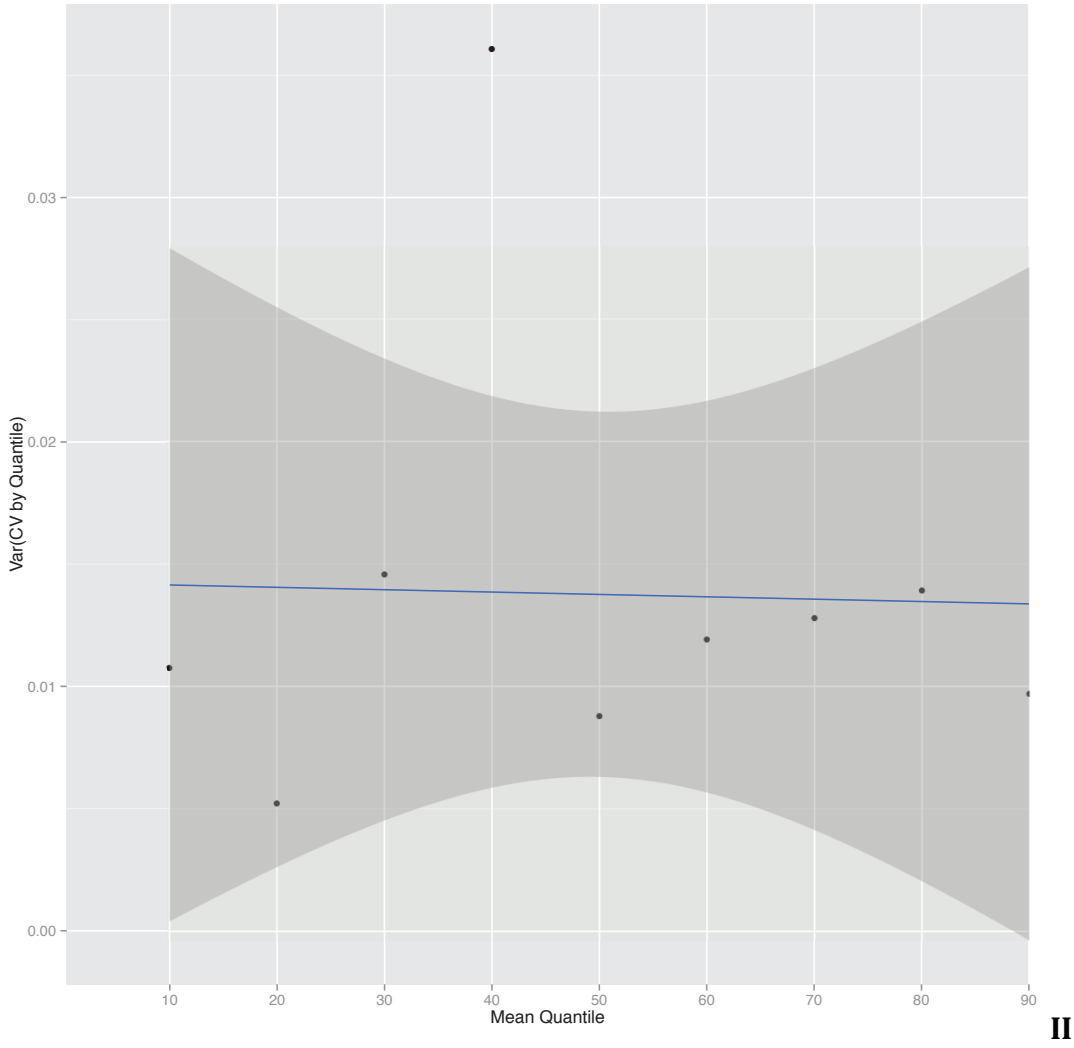


Figure A5- Variance of CV is highly similar as a function of Mean: To demonstrate that expression noise is constant as a function of Mean, the Variance of CV as a function of Mean Quantiles (10%) was determined. No correlation between Mean Quantile and Var(CV) is observed and values fall within a narrow band.

Independent Clone Generation Replicate

Despite the large number of clones generated, we entertained the possibility that the relationships inferred from this large set of clones may be an artifact of the particular clones generated or the particular culture conditions. Therefore, to address this possibility we generated an independent set of 191 clones. These clones were derived under identical conditions as the main set from an independent infection of uninfected Jurkat T cells with LGM2 virus. We found that the relationships between Mean and Variance (regression slope ~2) and Mean and CV (no significant correlation) were not significantly different from the main set of clones used in this study (Fig.A6) Therefore, we do not find evidence that suggests

out findings are dependent on the particular set of clones generated or culture conditions.

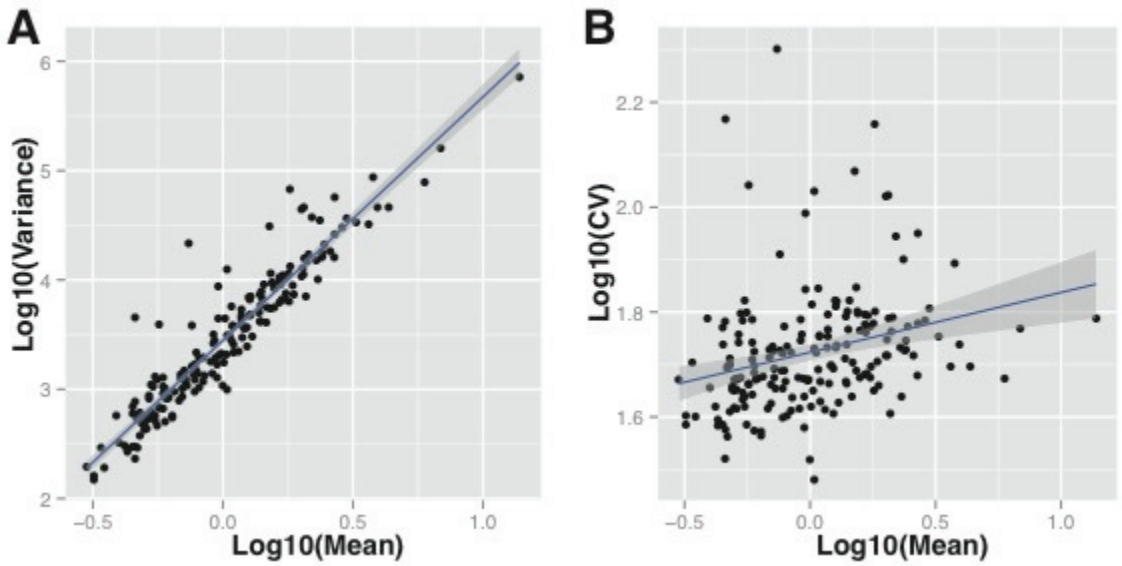


Figure A6- Repeat clone generation and analysis reveals no significant difference from main clones. A set of 191 clones was independently derived using the same protocol described for the main set of clones used in this study and scaling between moments was examined. **(A) No significant difference in mean v. variance relationship:** Regression analaysis of the mean v. Variance relationship of the independent set of clones reveals a high degree of correlation(R^2=0.89) and no significant difference in slope ($p>0.1$) from our analysis in the main text. **(B) Uncorrelated Mean and CV:** In agreement with the main set of clones, we find there to be no significant correlation between expression noise (CV) as a function of mean (R^2=0.08)

III Clone Selection and Stability

Our initial observation of an uncorrelated relationship between expression noise and mean led us to devise a clone selection method that would: a) capture the range of CV and mean observed and b) permit deeper analysis of the regulation of expression noise for a given mean. Toward this aim, we performed hierarchical clustering and analyzed the mean silhouette value as a function of cluster number (Fig. A7A, left). We found that the mean Silhouette is only weakly dependent on the cluster number with four clusters yielding the highest value. Examining the intra-cluster Silhouette value for each clone (Fig. A7A, right) reveals those 4 clusters result in broad shoulder of clones with similar Silhouette values and very few clones with negative Silhouette values. Clustering primarily places clones with similar mean expression into the same cluster (Fig. A7B). For further analysis, pairs of clones were selected from each cluster that had similar mean expression but large differences in their CV (Fig. A7C). A subset of 25 representative clones that were chosen for further analysis to capture the span of expression means and variances observed in the full set of clones.

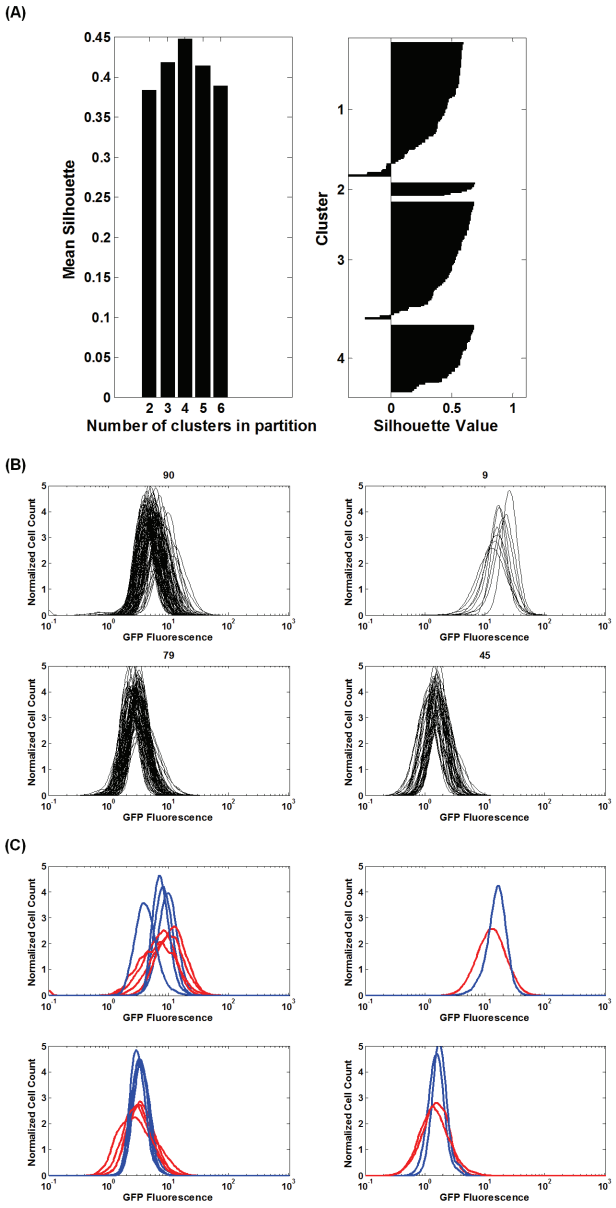


Figure A7- Hierarchical clustering of clones results in four clusters with similar intra-cluster mean but differing CV. Normalized GFP distributions of 223 clones were clustered into different number of clusters using a hierarchical clustering algorithm. (A) *Left* - The mean silhouette shows that the clones are optimally clustered into 4 groups. *Right* - The silhouette value of each clone when clustered into 4 groups. Most of the clones within a group have similar silhouette values with very few clones having negative silhouette values. (B) Each sub-figure represents the normalized GFP distributions of clones that fall within that cluster. The clones get segregated into different clusters primarily on the basis of their mean expression level. The number above each sub-figure represents the number of clones within that cluster. (C) Pairs of clones were chosen from each cluster to have similar mean but widely differing CV. In this figure, each pair is represented by a blue and red distribution. The GFP distributions in blue indicate clones with low CV whereas the GFP distributions in red represent clones with high CV.

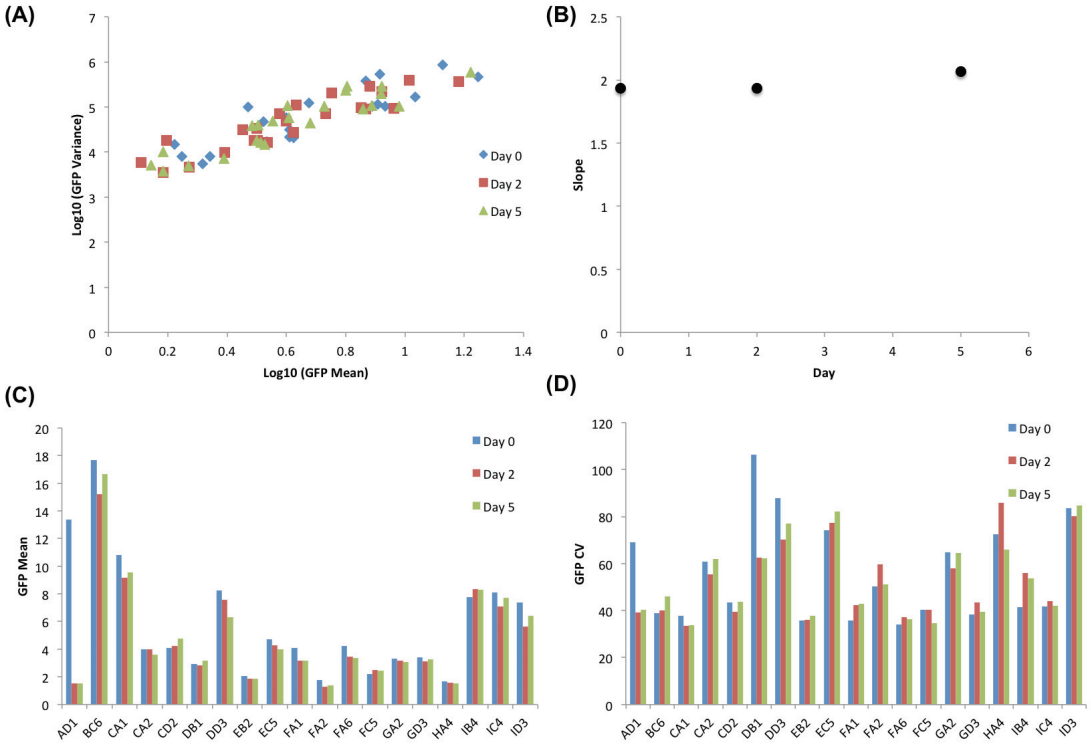


Figure A8- Clones selected for smFISH display stability in mean and noise over 5 days. Cell sorting and expansion produce stable clones that achieve steady state gene expression. GFP distribution for clones that were chosen for smFISH were monitored over 5 days. (A) Each color represents GFP variance *vs.* mean for all the clones on one particular day. (B) The slope of the GFP variance *vs.* mean plot remains unchanged over the 5 days with a value of ~2. (C) GFP Mean and (D) GFP CV for individual clones remain unchanged over the 5 days, indicating that they have achieved steady state gene expression.

IV Computational Analysis of Microscopy Images

smFISH enables elegant visualization of single mRNA molecules in fixed populations of cells. However, its application in high-throughput has been hampered in part by the lack of automated software tools. Therefore, to enable the application of smFISH across many clonal populations and over fifteen thousand single cells, we implemented primarily automated cell and smFISH segmentation software in MATLAB (R2011a, Mathworks Inc.) using the freely available DIPImage toolbox (v. 2.2, Linux). DIPImage provides a rich set of basic morphological, intensity and feature based image processing and measurement functions. Importantly, all functions in the toolbox scale to stacks of images such as those resulting from wide-

field deconvolution microscopy. We broke down high-throughput image processing into three tasks: 1) deconvolution to enhance signal-noise and correct geometry 2) cell segmentation and 3) smFISH signal segmentation.

High-throughput Deconvolution:

The optical properties of wide-field epifluorescent microscopes lead to blurring from out of focus light and geometric distortion from the point-spread-function (PSF) of the light path. We found that deconvolution using a PSF measured using 100nm fluorescent beads (Invitrogen Inc.) significantly enhances the automation of downstream processing. In particular, morphological operations and shape-based classification greatly benefit from the higher signal-to-noise ratio and corrected geometry resulting from deconvolution. Therefore, all 84 fields from each of the 25 LGM2 clones analyzed were first deconvolved using Huygens Core (v.2.x, SVI) on a 10-node dual-processor Linux cluster with jobs managed by Silicon Grid Enginer . Custom scripts in Tcl and Bash issued deconvolution commands and ran jobs. Deconvolution parameters were empirically determined to maximize final image quality.

Cell Segmentation:
Denconvolved image stacks undergo a multi-step process (Fig. A9) to segment whole cells in the field. Each raw image stack (Fig. A9A) is first pre-processed using a homomorphic filter (Fig. A9B) to remove the slowly varying luminance component from the image. Specifically, each image stack is first log-transformed and a heavily blurred (Guassian kernel with large sigma) copy of this log-transformed image is subtracted. The resulting stack is then transformed back to linear intensity space by exponentiation and linearly stretched to 8-bit gray scale. Subsequently, to remove interior peaks in intensity the image stack is then morphologically reconstructed (Fig. A9C) with an H-dome (Luc Vincent, 1993), which results in flatter intensity profiles that are less susceptible to over-segmentation. To generate a binary mask, the reconstructed image stack is then thresholded using an empirically determined multiplier of the *otsu* threshold. Then to separate touching objects, a Euclidean distance transform is applied (Fig. A9D) and then thresholded using a different multiplier of the *otsu* threshold. This results in unique seeds within each cell (Fig. A9E). Importantly, to prevent analysis of partial cells, seeds intersecting the image boundary are rejected. We found this method of seeding to be more robust than computing seeds from the DAPI stained nuclei. Specifically, nuclei frequently have interior holes and discontinuities that frequently result in over-segmentation. The resulting seeds were then first grown on the Euclidean transformed image, which effectively results in larger seeds that were then grown on the reconstructed image to arrive at final segmentation of the field (Fig. A8F). The resulting objects are rejected on the basis of perimeter to area (P2A) and Podczeck circularity shape descriptors, which identify aberrant objects resulting from under or over segmentation. Fields with rejected objects are retained for manual inspection. Retained fields typically comprise less than 5% of total fields. Segmentation problems in these fields can be overcome through a combination of threshold

adjustment. Segmentation object boundaries accurately determine cell boundaries (Fig. A9G). Furthermore, the sizes of the cell objects are not significantly different from clone to clone (Fig. A10A) and individual cell size distributions are well described by normal distributions (Fig. A10B). Together, these suggest that this segmentation method is robust across the clones studied and that cell size is not a significant factor underlying our analysis.

smFISH segmentation:

Each segmented cell object then undergoes a multistep filtering and classification process to identify both single molecule RNA signals and burst-like features. To remove the majority of slowly varying background remaining in deconvolved images (Fig. A11A), a morphological TopHat filter with an elliptical structuring element sized (7 pixels) to pass FISH signals but reject background is applied. To further enhance spherical signals a Laplacian of Gaussian transform is applied[3], resulting in a stack where almost all the background has been removed (Fig. A11B). What remains are true FISH signals, which are highly spherical, and irregular blobs. For each clone, a 'ballpark' threshold, which provides a reasonable mask of FISH signals and blobs, is manually determined by assessing performance across ten fields chosen at random. A previously reported method relied on setting a precise threshold for each cell[3]. This method does not scale efficiently to tens of thousands of cells and may be susceptible to user bias. Rather than relying strictly on a threshold, we found that classifying objects on the basis of size, circularity and intra-object grey level intensity standard deviation.
We empirically determined thresholds on these features that yielded classification in good agreement with manual classification. The combination of these three features efficiently distinguishes aberrant blobs from FISH signals. Specifically, to segment FISH signals connected component analysis is performed on the thresholded image. Very small (<30 total pixels) and very large (>300) objects are excluded in this step, which provides an initial classification filter. Subsequently, objects are only accepted if their perimeter to area (p2a) values are highly consistent with a sphere (p2a between 0.96 and 1.02). Secondly, background blobs have very uniform pixel intensity across the object while FISH signals have radially decreasing intensity from the center of the diffraction limited spot. This is captured most succinctly in the pixel grey level standard deviation within each object. Therefore, objects with a grey level standard deviation below an empirical threshold are also rejected. In practice, this classification provide counts <5% different from manual counts and is highly scalable. The combination of a 'ballpark' threshold with a three-feature classification scheme allowed us to scale semi-automated FISH analysis to tens of thousands of single cells.

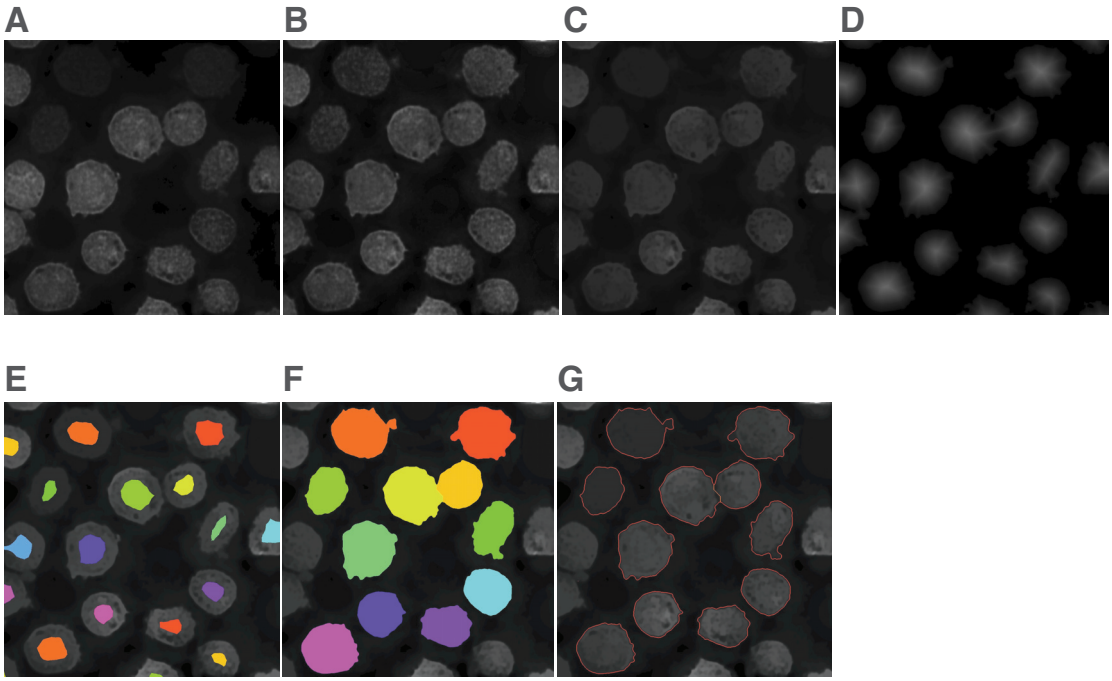


Figure A9- Unsupervised segmentation of cells in fields accurately identifies single cell boundaries To normalize intensity and reduce intensity noise within cells, deconvolved fields **(A)** are first pre-processed with a homomorphic filter **(B)** and then reconstructed with an H-dome structuring element **(C)**. The reconstructed image is then thresholded and a Euclidean distance transform applied **(D)**. Subsequently, the transformed image is thresholded and seeds identified using connected components analysis **(E)**. These seeds are then expanded through a seeded region-growing algorithm to arrive at final segmentation **(F)**, which provides robust identification of single cells and accurately identifies cell boundaries. All image are maximum intensity projections, however processing is performed on the entire image stack.

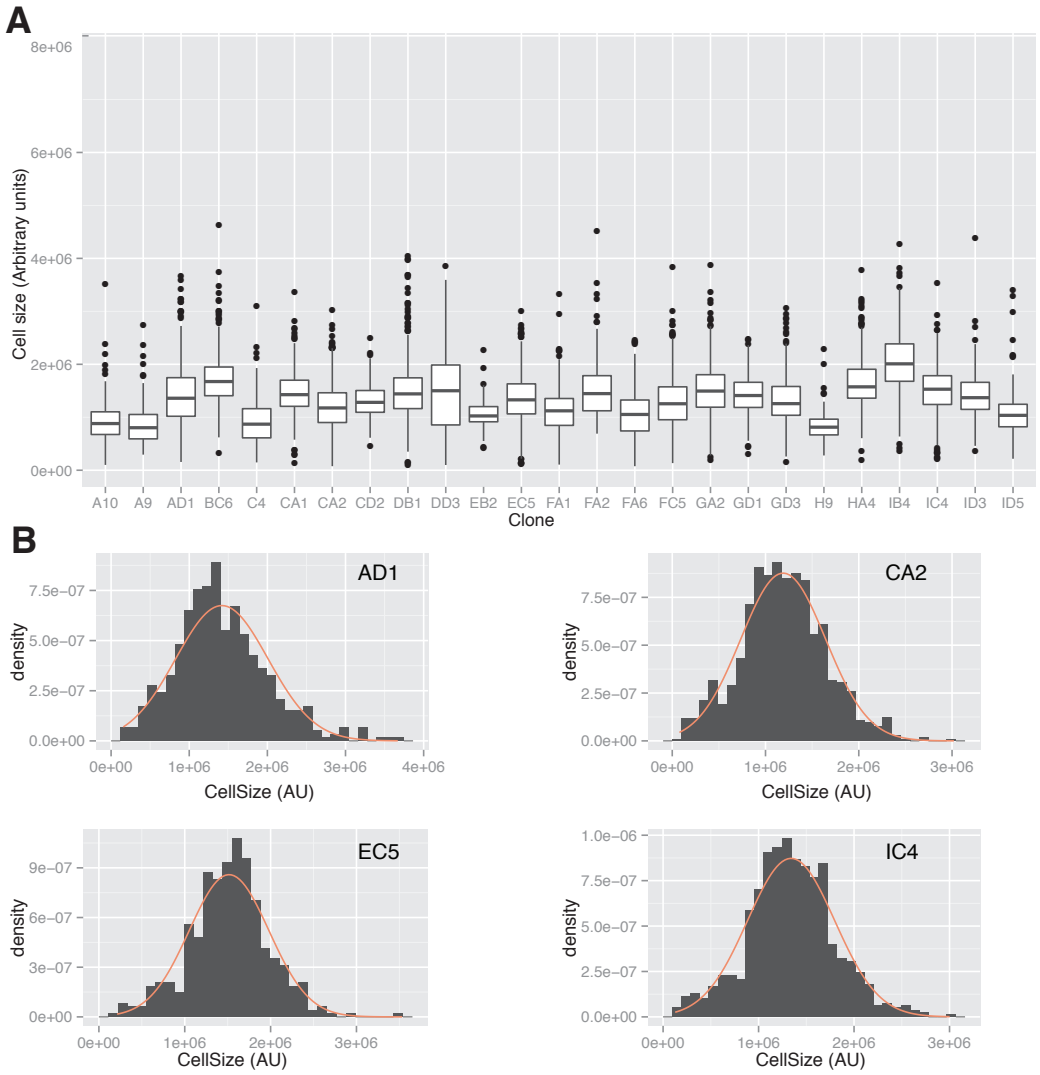


Figure A10- Cell sizes are well described by Normal distributions and mean cell size exhibits low variation across clones As an initial assessment of the quality of segmented objects, we examined the cell size (sum of segmented object pixels) distribution for each clone. **(A) Box and Whisker plots of clone cell size:** Examination of Box and Whisker plots of cell size for each of the clones we analyzed reveals little variation (<2) in mean cell size and no apparent trend between either mean RNA copy number or noise. **(B) Normal Fits of Cell Size Distributions:** The cell size distributions of all clones are well described by a Normal distribution. A subset of clones with Normal fits overlaid on histograms of cell size is shown.

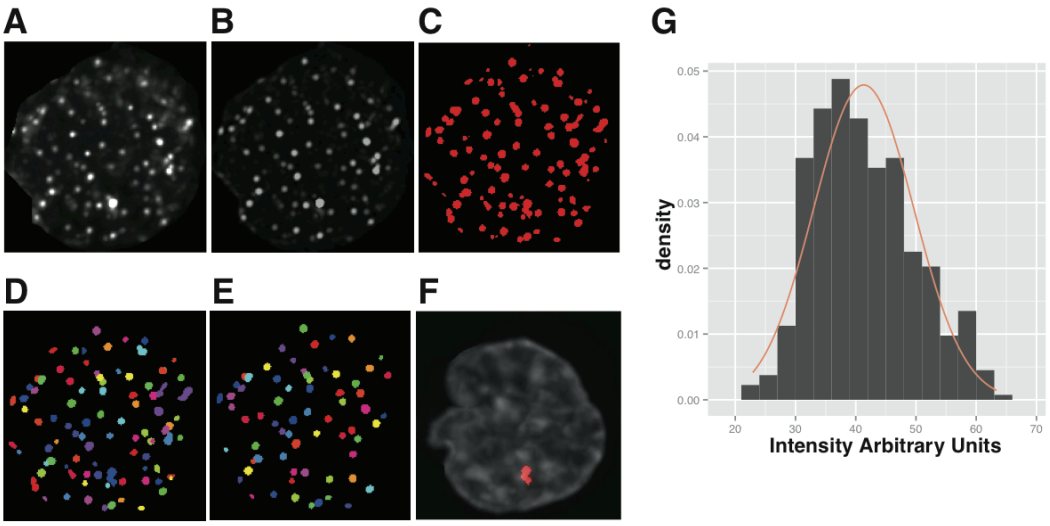


Figure A11-A combination of thresholding and feature based classification permits high-throughput identification of smFISH signals.
(A-E) Multistep filtering and segmentation of FISH signals: To remove background and amplify spherical signals in the deconvolved images(A), a TopHat filter and Laplacian of Gaussian transform are sequentially applied(B). A 'ballpark' threshold determined for each clone is used to threshold the filtered and transformed image (C) Connected components analysis identifies all particles between 30 and 300 pixels in size (D). Perimeter to area (P2A) and pixel grey-level standard deviation are used to classify true FISH signals (E) from background objects. **(F) Burst identification:** On the basis of size, summed pixel intensity and pixel grey-level standard deviation active transcriptional centers ('Bursts') can also be identified. **(G) Normally distributed particle intensity:** The intensity of the identified smFISH signals are well described by a single Gaussian (red) distribution, suggesting predominantly single molecule signals. The histogram shown represents thousands of particles scaled according to empirical density to permit comparison to Gaussian density.

V Measurement of RNA Degradation

Basal mRNA degradation rates were determined through the measurement of GFP mRNA abundance in a polyclonal population of HIV-1 LGM2 infected Jurkat T cells treated with 50μM α-amanitin (Sigma A2263). Total RNA samples were obtained by collecting 100,000 cells in TRIzol (Invitrogen 15596-018) at 0, 1, 2, 4, 6, and 8 hours post α-amanitin treatment and then extracted by repeated phenol-chloroform extractions. RNA concentration was quantified with a NanoDrop 1000 spectrophotometer. GFP mRNA abundance was determined by RT-qPCR in triplicate using a QIAGEN QuantiTect SYBR Green RT-PCR Kit (QIAGEN 204243) with a Bio-Rad iQ5 system. 50ng of total RNA and 0.25μM of GFP forward and reverse primers were reacted according to the supplied instructions at an annealing temperature of 62°C for 45 cycles. To assess the reliability of our measurement, β-Actin mRNA degradation was also measured as a control for normal mRNA degradation rates. We consistently observed an increase in the apparent LGM2 mRNA level at the 2hr time-point, which is likely due to RNA export phenomena as

previously reported (Raj, 2006). Furthermore, cells are markedly sick at the 8hr time-point. Therefore, we only fit the 2hr, 4hr and 6hr time-points. Best-fit estimates of RNA degradation rates were obtained by performing linear regression on semi-log transformed data. Based on this, we find the LGM2 degradation rate to be approximately -0.34 hr^{-1} (Supplementary Fig. 11A, regression R^2=0.92) and the β-Actin degradation rate (Supplementary Fig. 11B, regression R^2=0.73) to be approximately -0.14 hr^{-1}. Our estimated β-Actin mRNA degradation rate was found to be highly similar to previously published values[2]. Primer sequences used are displayed below.

Primer	Sequence (5' to 3')
β-Actin forward	CCTGGCACCCAGCACAAT[1]
β-Actin reverse	GCCGATCCACACGGAGTACT[1]
GFP forward	AGCAAAGACCCCAACGAGAA
GFP reverse	CGTCCATGCCGAGAGTGAT

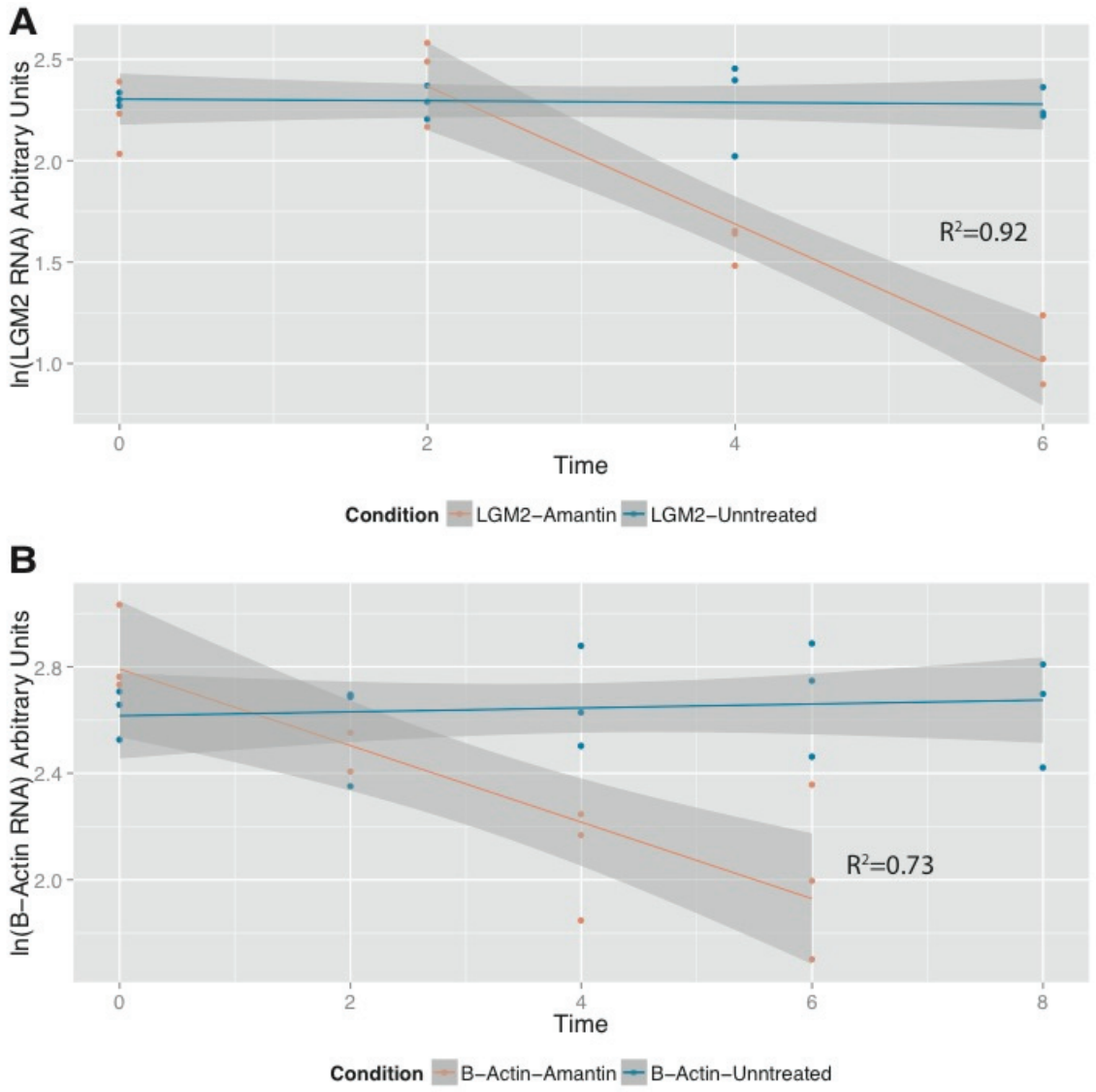


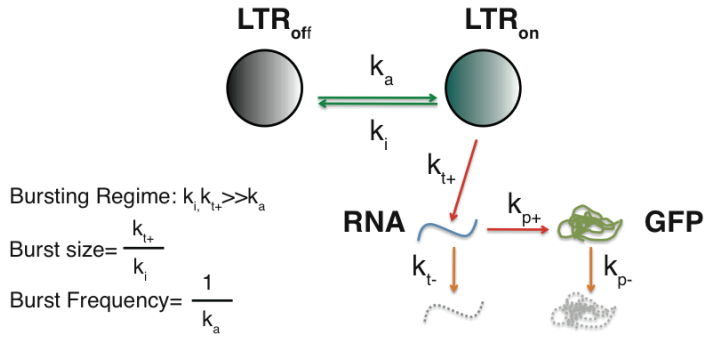
Figure A12- Experimental determination of mRNA degradation rates for LGM2 and β -Actin. (A) LGM2 mRNA degradation: To halt RNA Polymerase II transcription, a polyclonal LGM2 population was treated with 50μM α -Amanitin. Relative RNA levels were determined at 0-6 hours post-treatment for both Amantin treated (orange) and untreated (blue) populations through qPCR. Due to a consistently observed increase in LGM2 level at 2hrs, the 0 time-point was excluded from estimation of the degradation rate. The degradation rate was estimated through log-linear regression of the 2, 4, 6 hr timepoints with 3 qPCR replicates per time-point. The best-fit slope is -0.34 hr^{-1} (R^2=0.92), representing a half-life of ~2 hrs. **(B)** To ascertain whether our system for measuring degradation rate provided a reasonable estimate, we also measured β -Actin levels following treatment identical to (A). Using Log-linear regression (orange line) we found the best-fit rate to be -0.14 hr^{-1} (R^2=0.73), which is highly consistent with a previously published rate. Shading around the best-fit regression represents the point-wise 95% confidence interval.

VI Model Fitting

Maximum likelihood estimation (MLE) of the promoter ON rate (k_a) and average burst size (k_{t+}/k_r) for each clone was performed against the full analytical

probability density function (*pdf)* of the standard 2-state model (Peccoud and Ycart, 1995,Raj 2006) of gene expression (Supplementary Fig. 12). Despite it's abstractness, this model has received widespread use due to its comprehensive ability to reveal kinetic mechanisms underlying burst-like transcription. MLE was implemented through specification of the log-likelihood function of a custom distribution in Mathematica 8 (Wolfram Inc.). Parameters were estimated by numerically minimizing the negative log-likelihood of the two-state *pdf* given the experimentally determined RNA distribution for each clone. As discussed, we experimentally measured the LGM2 degradation rate and therefore k_{t-} was fixed at this value. In addition, as previously reported, RNA distributions are insufficient to separately determine the promoter off rate and the transcription rate. As we and others have done previously, (.....) we held the transcription rate in the ON state constant across clones. We used our previously reported value of 60 hr-1 (Skupsky 2010). Values ±50% of this fixed value did not significantly change estimates of k_a or Burst sizes, suggesting that our results are highly independent of the particular value chosen.

A



B

$$P(n\ \ RNA)=\frac{\left(\frac{k_{t+}}{k_{t-}}\right)^{n}\Gamma\left(\frac{k_a}{k_{t-}}+\frac{k_i}{k_{t-}}\right)\Gamma\left(\frac{k_a}{k_{t-}}+n\right)\,{}_1F_1\left(\frac{k_a}{k_{t-}}+n;\frac{k_a}{k_{t-}}+n+\frac{k_i}{k_{t-}};-\frac{k_{t+}}{k_{t-}}\right)}{\Gamma(n+1)\,\Gamma\left(\frac{k_a}{k_{t-}}\right)\Gamma\left(\frac{k_a}{k_{t-}}+n+\frac{k_i}{k_{t-}}\right)}$$

Figure A13-A two-state stochastic model of gene expression:
(A) Model configuration: The LTR is allowed to transition between a non-productive 'OFF' state and a productive 'ON' state with forward and reverse rates k_a and k_r, respectively. RNA is produced in the 'ON' state at rate k_t and is degraded by a first-order process at rate k_{t-}. Similarly, protein is produced and degraded at rates k_p and k_{p-}, respectively. Kinetic rates are taken to be transition state probabilities. Under conditions where infrequent transitions are made to the 'ON' state, the lifetime of the 'ON' state is short, and transcription is fast relative to the frequency of turning 'ON', transcription is 'burst-like'. The steady-state RNA distribution resulting from burst-like transcription is well described by the frequency of turning on ('Burst frequency') and the average number of transcripts produced in the 'ON' state ('Burst size') **(B) Steady-state RNA probability density function:** Maximum-likelihood estimation of model parameters was performed against the full RNA

probability density function, which represents the steady-state solution to the chemical master equation representing promoter and RNA copy number states but not protein.

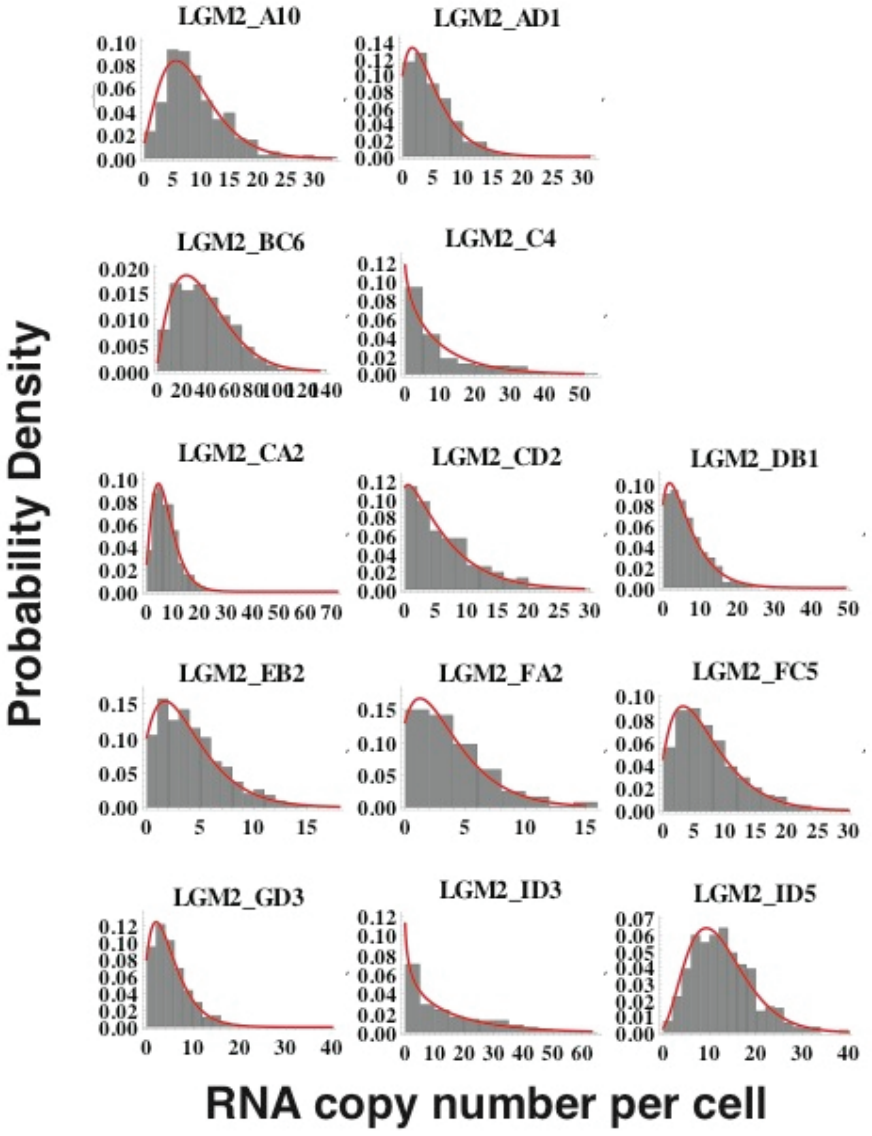


Figure A14-Maximum-Likelihood Estimation (MLE) of two-state stochastic model quantitatively describes LGM2 distributions: As detailed in Chapter 2, MLE was performed using an experimentally determined rate of RNA degradation and holding the rate of transcription constant. Estimates of the burst frequency and size quantitatively describe all 25 LGM2 distributions.

VII Summary of Clone Moments and Model Parameters

See Table A2

VIII Additional Fit Parameter Analysis

In this study we reveal both observational and mechanistic othorgonality in the control of expression mean and expression noise (CV) across integration positions. Specifically, we find expression mean and noise to be uncorrelated with burst size primarily explaining mean and promoter ON rate primarily explaining noise. Despite the lack of significant cross-correlations between promoter on rate and mean, and burst size and noise, some possibility of co-modulation remains. Therefore, we directly examined burst size as a function of promoter on rate (Fig.A14) and found no evidence of significant correlation ((R^2=0.08, p>0.1), (r_s=-0.38,p>0.05)). This suggests that burst size and ka respectively provide orthogonal explanations for mean and noise.

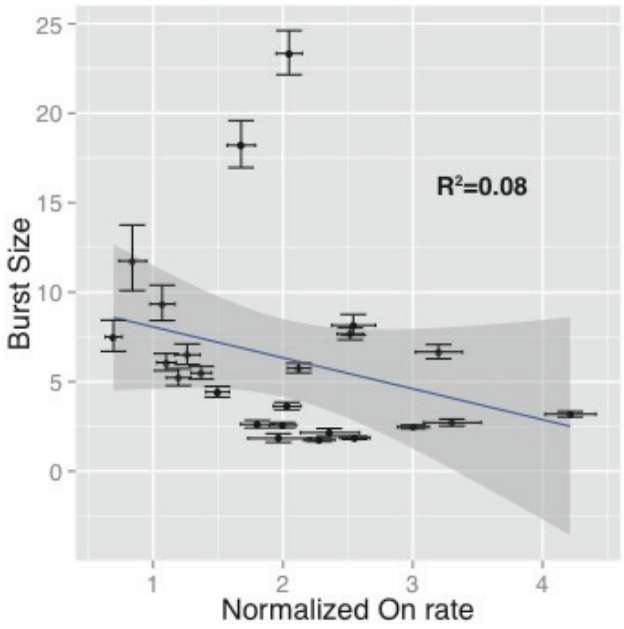


Figure A15 Burst size is not a significant function of the promoter ON rate: To resolve whether burst-size and promoter ON rate respectively provide orthogonal explanations of expression mean and expression noise, we examined both linear regression (blue line) and non-parametric Spearman correlation. Neither linear regression (R^2=0.08, p>0.1) or the Spearman correlation coefficient (r_s=-0.38,p>0.05) suggest that there is a significant correlation. Shading around the regression line indicates the point-wise 95% confidence interval.

IX Additional Nucleosome Sensitivity Analysis

In this study we established that for similar mean levels of expression, noisier clones have more inaccessible chromatin at the promoter. Closer analysis along different regions of the promoter revealed that the HSS consistently has the highest ratio of chromatin inaccessibility between high- and low-noise clone pairs. This arises from much lower chromatin inaccessibility values for low-noise clones compared to other sites in the promoter (Supplemental Fig. 14). We also established that chromatin inaccessibility at Nuc-1, unlike HSS or Nuc-0 (Supplemental Fig. 15) is the best predictor of the promoter ON rate. This is further reinforced by performing a principal components analysis (Supplemental Fig. 16).

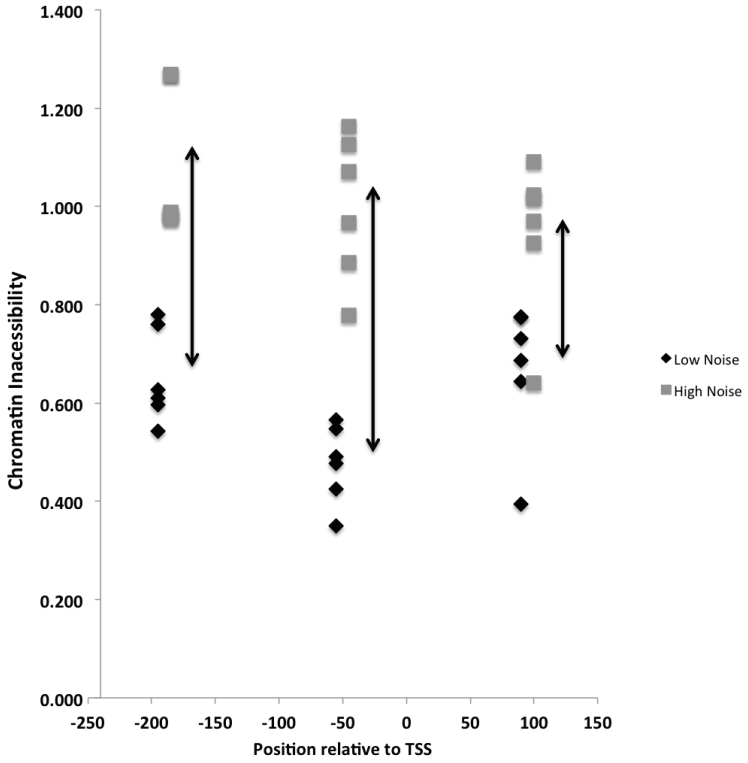


Figure A16 – Chromatin inaccessibility for all clones at three positions along the HIV-1 LTR. The figure shows that for all clones and across all positions within the HIV-1 promoter, noisier clones have more inaccessible chromatin. Interestingly, while the mean levels of chromatin remains unchanged across the promoter for the noisier clones, low-noise clones tend to have more open chromatin at the HSS compared to other regions of the promoter. The x-axis corresponds to the midpoint of the amplicons used in qPCR.

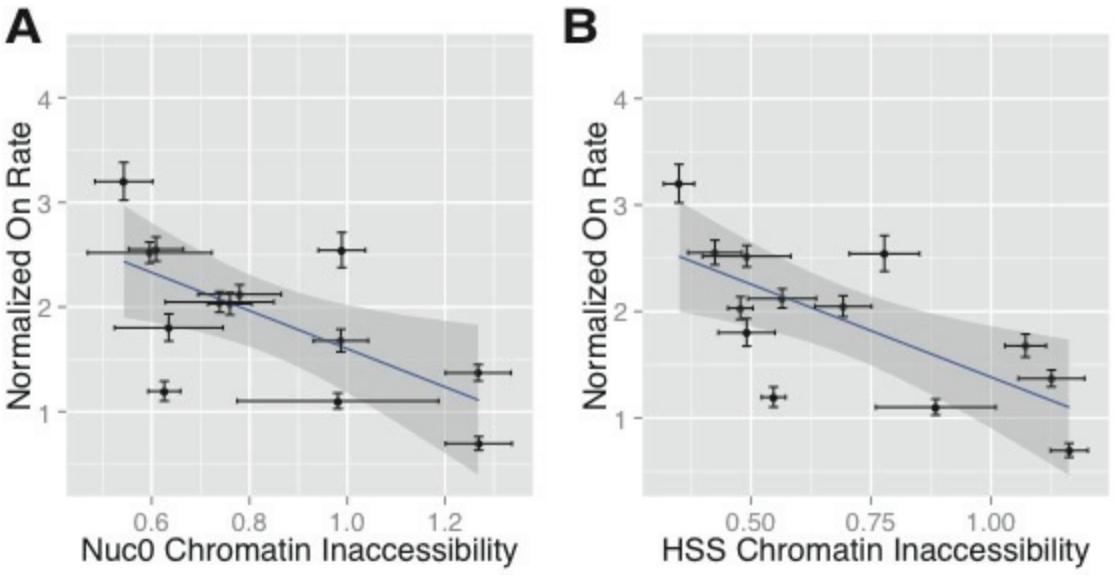


Figure A17 –The promoter activation rate does not correlate strongly with the chromatin state at other regions of the HIV-1 promoter. Chromatin inaccessibility at (A) the hypersensitive site (HSS) and (B) the Nuc-0 site does not correlate strongly with the promoter activation rate. This suggests that Nuc-1 plays the most important role in setting the rate of promoter transition from the inactive to active state. All qPCR was performed in triplicate and error bars reflect the standard deviation from the mean. Error bars for the maximum likelihood estimates of k_a represent 95% confidence intervals.

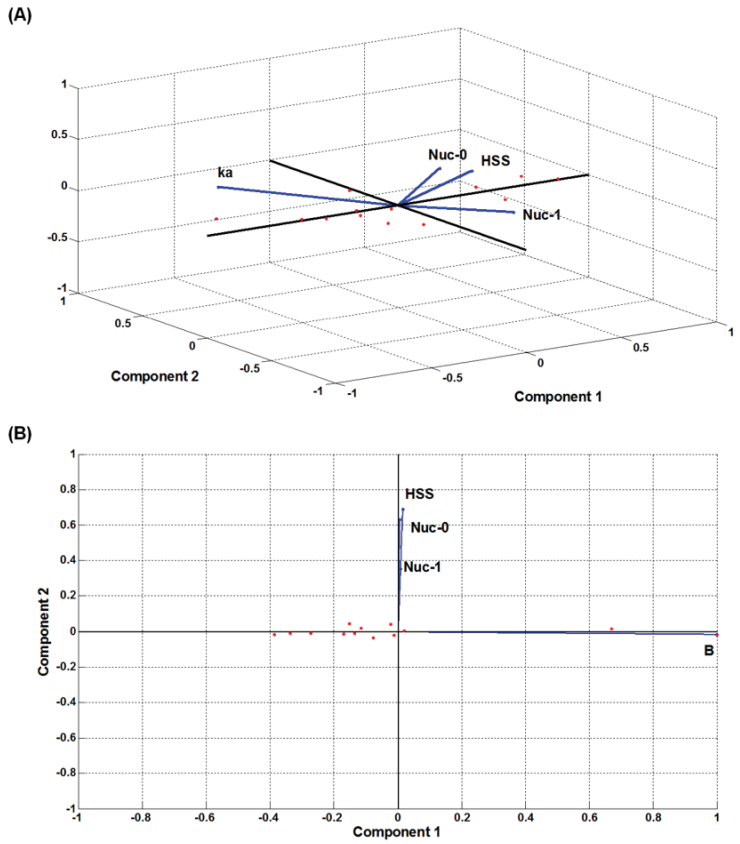


Figure A18- Chromatin inaccessibility at Nuc-1 is the best predictor of the promoter ON rate while the burst size is not correlated to the chromatin state at the promoter. (A) Principal components analysis (PCA) for the promoter ON rate and the chromatin state at the promoter shows that most of the data lies within the 2D space of the first two principal components. The axis for chromatin inaccessibility at Nuc-1 is almost negatively correlated to the promoter ON rate whereas HSS and Nuc-0 are almost orthogonal to the promoter ON rate. Therefore, chromatin inaccessibility at Nuc-1 appears to be the best predictor of k_a. (B) PCA for the burst size and the chromatin state at the promoter shows that most of the data lies along the 1D axis of the first principal component and that the burst size is independent of the three chromatin inaccessibility measurements along the promoter.

X Summary of Statistical Analysis

Table A1: Summary of linear regression analysis

	explanatory	response	intercept	slope	slope95CI	R^2	$r_{spearman}$	p
1	log10(all.Mean)	log10(all.Var)	-0.92	2.10	0.11	0.86	0.92	0.00
2	log10(all.Mean)	log10(all.CV)	-0.46	0.05	0.05	0.01	0.17	0.01
3	log10(subset.Mean)	log10(subset.Var)	-0.73	2.04	0.37	0.85	0.93	0.00
4	log10(subset.Mean)	log10(subset.CV)	-0.37	0.02	0.18	0.00	0.05	0.81
5	log10(RNA.Mean)	log10(RNA.CV)	0.09	-0.20	0.13	0.31	-0.43	0.03
6	log10(RNA.Mean)	log10(RNA.Var)	0.17	1.61	0.25	0.88	0.93	0.00
7	log10(RNA.Mean)	log10(GFP.Mean)	-0.24	0.88	0.20	0.79	0.86	0.00
8	log10(RNA.Var)	log10(GFP.Var)	-1.39	1.17	0.31	0.73	0.84	0.00
9	B	RNA.Mean	1.96	1.34	0.34	0.75	0.79	0.00
10	ka	RNA.Mean	6.78	5.26	12.16	0.04	0.20	0.35
11	B	RNA.CV	0.83	0.00	0.01	0.01	0.03	0.89
12	ka	RNA.CV	1.16	-0.51	0.13	0.75	-0.89	0.00
13	Nuc1	ka	1.53	-1.11	0.50	0.69	-0.85	0.00
14	Nuc1	RNA.CV	0.31	0.59	0.51	0.38	0.62	0.02
15	ka	B	9.79	-5.08	7.66	0.08	-0.38	0.07

XI Investigation of Nuc-1 occupancy and noise regulation by BAF250a

While our results presented in Chapter 2 provide strong inferential evidence of the relationship between Nuc-1 occupancy and noise, they do not provide a direct casual relationship. A recent study suggests that in a model HIV system that included Tat, a SWI/SNF complex containing BAF250a is responsible for moving a nucleosome from its lowest energy position to the repressive Nuc-1 position[4]. Our findings strongly indicate a role for Nuc-1 occupancy in regulating the rate of stochastic LTR transitions and the resulting level of gene expression noise across integration positions. We hypothesized that differential recruitment of BAF250a containing complexes across genomic sites of viral integration may underlie our observation. We reasoned that by knocking down the expression of BAF250a with a

synthetic siRNA, we could decrease Nuc-1 occupancy, increase the rate of LTR activation and thereby decrease expression noise for a given integration position.

To facilitate quantitative analysis relating the extent of BAF250a knock-down with changes in GFP expression, we first optimized staining of BAF250a using a Jurkat cell line containing single integrations of a minimal Tat feedback vector. BAF250a is a nuclear resident transcription factor and therefore requires permeabilization of the nuclear envelope prior to antibody staining. We found that detergent based protocols involving Triton X-100 at percentages ranging from 0.1%-0.3% resulted in significant cell lysis. Therefore, we employed sequential 4% Formaldehyde and 100% cold Methanol to fix and permeabilize Jurkat cells[6]. While a commercial monoclonal anti-BAF250a antibody (Santa Cruz Biotechnology, sc-3276) failed to yield staining under these conditions, a rabbit BAF250a anti-sera (obtained as a kind gift from the NIH laboratory of Dr. Weidong Wang) yielded staining well discriminated from background (Fig. A19).

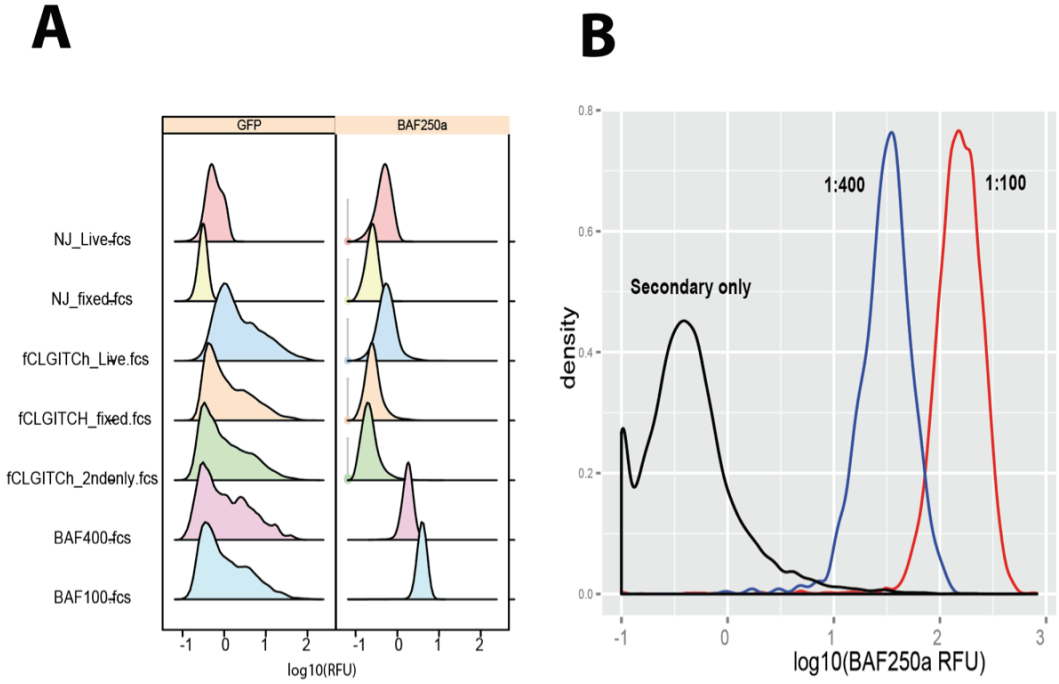


Figure A19-Optimization of BAF staining: (A) Following sequential 4% Formaldehyde, 100% Cold Methanol fixation and permeabilization, we stained GFP expressing Jurkat T cells with rabbit anti-BAF250a sera at dilutions of 1:100 and 1:400 followed by staining with an Alexa647 Goat anti-rabbit IgG secondary antibody. GFP and Far-Red ('BAF250a') fluorescence where measured by flow cytomerty. As controls we examined both live and fixed GFP negative Jurkat's (NJ), both live and fixed GFP expressing Jurkat's (fCLGITCh) and GFP expressing Jurkat's treated with only secondary antibody. Following fixation, we observed reduction in both Autofluorescence and GFP fluorescence. Further, we found that only the 1:100 anti-sera dilution is significantly separated from the secondary only control. (B) Density overlay of just the BAF250a channel for secondary only, and anti-sera dilutions demonstrates that the staining is well discriminated from background.

Following optimization of BAF250a staining, we examined the effects of BAF250a knockdown on LGM2 GFP expression. A pooled siRNA targeting BAF250a (Dharmacon, SmartPool ON-TARGETplus ARID1A siRNA) was introduced into three LGM2 clones through nucleofection (Amaxa Nucleofector II instrument with Nucleofector kit V), which facilitates direct nuclear delivery of siRNA. 48 hrs. after nucleofection, cells where fixed and stained for BAF250a. Unfortunately, there was only partial knockdown of ~30% of cells. We compared the GFP CV for BAF250a- population to the BAF250+ population and found minimal difference. There appeared to be a slight decrease in distribution width, however this was only observed in one clone (Fig. A20). It is unclear if this is due to inefficiency of knockdown or that GFP measurements are indirect and out of steady-state. This result may also to be to cytotoxicity resulting from nucleofection.

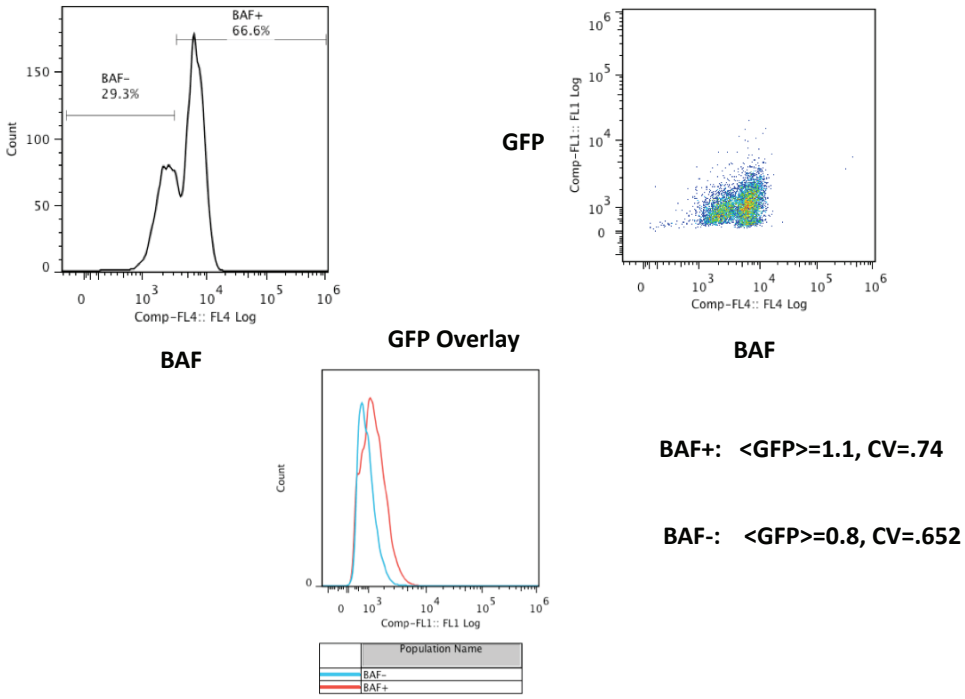


Figure A20-BAF250a siRNA nucleofection yields partial knockdown with minimal reduction in GFP CV: Three LGM2 clones were nucleofected with BAF250a siRNA and GFP fluorescence determined 48 hrs. post-nuclefection. BAF250a expression was determined through immunofluorescence using a rabbit anti-sera. BAF250a exhibited minimal apparent knockdown (~30%). Comparing the BAF250- subpopulation to the BAF250+ subpopulation, we find minmal differences in GFP CV (0.652 versus 0.74 respectively). The GFP mean also decreased and may suggest cell stress.

Table A2: Summary statistics, model fit parameters and NSA measurements for LGM2 clones analyzed in Chapter 2

Clone	N	<RNA>	RNA Var	RNA CV	RNA Sk	RNA Fano	ka	B	<GFP>	GFP Var	GFP Fano	GFP CV	Nuc1	HSS	Nuc0	LD Primers
LGM2_A09	236	4.93	15.77	0.81	1.39	3.20	0.80	2.15	3.41	1.01	0.30	0.29				
LGM2_A10	344	8.52	30.14	0.64	1.23	3.54	1.12	2.71	2.58	0.61	0.24	0.30				
LGM2_AD1	498	4.60	16.53	0.88	1.73	3.59	0.61	2.62	1.48	0.30	0.20	0.37	0.86	0.49	0.63	0.14
LGM2_BC6	824	37.69	511.59	0.60	0.67	13.57	0.70	23.34	16.60	55.48	3.34	0.45	0.85	0.69	0.74	0.14
LGM2_C04	208	9.33	110.74	1.13	1.57	11.87	0.29	11.74	1.58	1.86	1.18	0.86				
LGM2_CA1	1027	17.43	111.00	0.60	0.72	6.37	0.86	7.67	9.40	8.43	0.90	0.31	0.77	0.49	0.60	0.18
LGM2_CA2	1058	7.13	25.99	0.72	2.82	3.65	1.02	2.47	3.42	4.26	1.24	0.60	1.09	0.97	0.97	0.20
LGM2_CD2	482	6.04	30.63	0.92	1.16	5.07	0.41	5.23	4.65	3.58	0.77	0.41	0.78	0.55	0.63	0.05
LGM2_DB1	715	6.38	33.97	0.91	2.02	5.33	0.51	4.43	3.14	3.66	1.17	0.61				
LGM2_DD3	464	7.82	65.04	1.03	2.58	8.31	0.43	6.50	5.95	19.23	3.23	0.74				
LGM2_EB2	887	3.90	9.99	0.81	1.09	2.56	0.77	1.75	1.85	0.45	0.24	0.36				
LGM2_EC5	816	7.22	41.61	0.89	1.54	5.77	0.47	5.49	3.90	9.89	2.54	0.81	1.02	1.13	1.27	0.44
LGM2_FA1	1050	4.64	13.58	0.80	1.58	2.93	0.87	1.87	3.18	1.80	0.57	0.42	0.64	0.42	0.61	0.06
LGM2_FA2	180	3.54	9.59	0.88	1.23	2.71	0.67	1.84	1.37	0.47	0.34	0.50				
LGM2_FA6	1061	11.43	67.68	0.72	1.33	5.92	0.72	5.76	3.33	1.24	0.37	0.33	0.69	0.57	0.78	0.07
LGM2_FC5	782	7.08	29.48	0.77	1.15	4.16	0.69	3.63	2.32	0.43	0.19	0.28	0.73	0.48	0.76	0.07
LGM2_GA2	679	6.45	37.51	0.95	1.42	5.81	0.37	6.07	2.88	3.31	1.15	0.63	1.02	0.89	0.98	0.20
LGM2_GD1	513	26.01	353.84	0.72	1.16	13.60	0.57	18.21	8.32	28.01	3.37	0.64	0.97	1.07	0.99	0.25
LGM2_GD3	881	4.99	17.50	0.84	1.76	3.51	0.68	2.57	3.15	1.13	0.36	0.34				
LGM2_H09	366	9.45	89.33	1.00	2.05	9.45	0.36	9.34	3.35	3.61	1.08	0.57				
LGM2_HA4	469	5.08	33.70	1.14	1.44	6.64	0.24	7.51	1.42	0.81	0.57	0.63	0.92	1.16	1.27	0.65
LGM2_IB4	372	18.54	135.07	0.63	1.52	7.28	0.86	8.15	8.26	18.93	2.29	0.53	0.64	0.78	0.99	0.05
LGM2_IC4	435	19.04	104.31	0.54	0.48	5.48	1.09	6.66	7.78	10.86	1.40	0.42	0.39	0.35	0.54	0.02
LGM2_ID3	546	13.96	188.84	0.98	0.97	13.52	0.24	21.29	5.38	18.23	3.39	0.79				
LGM2_ID5	563	12.47	45.95	0.54	0.89	3.68	1.43	3.19	3.91	3.85	0.98	0.50				

Appendix B

Synthetic U3 generation protocol

A. PCR extension of degenerate pre-parts

1. Parts are pooled at 1nmol/ul total concentration
2. Setup 10 50ul PCR reactions with the following components:
    a. 1nmol ssLTR parts
    b. 1.25nmol Bt-extend primer (a 5' Biotinylated primer for primer extension)
    c. 2ul 10mM dNTPs
    d. 1ul Vent (New England Biolabs)polymerase (add this on the PCR block at 95C or keep reactions on ice)
    e. ThermoPol Buffer (10X)

3. Use the following PCR program
    a. 95C 5 minutes → add Vent to tubes here
    b. 60C 2 minutes
    c. 72c 70minutes

B. Cleanup of extended parts

1. Pool PCR reactions and Phenol:Chloraform extract with 200ul P:C
2. Remove top layer, avoiding interfacial layer (you should be able to recover ~400-450ul)
3. Ethanol precipitate with 1:10 vol 3M NaOAc pH 5.2 and 2.5 volumes cold 100% EtOH
4. Mix well and incubate on ice for 30 minutes
5. Spin in 4C microcentrifuge (roon 277) for 15 minutes at 4C
6. Large white pellet should be visible, remove supernatant and wash the a few hundred ul of cold 70% EtOH. If pellet lifts off the side of the tube, you can re-spin briefly to re-pellet. Remove pellet and let air-dry for 10 minutes with the tube cap open
7. Resuspend in 200ul dH20. Quantitate DNA, if above 1ug/ul dilute to 1ug/ul, label tube "Extended LTR Parts".

C. MlyI digestion

1. Setup 15 30ul digests with 10ug extended parts, NEB #4 Buffer,BSA and 2ul MlyI (New England Biolabs).
2. Digest for 3hrs at 37C

D. Steptavidin Column cleanup

1. Pool digests and P:C extract with 150ul Phenol:Chloraform
2. Remove top layer to a clean tube

3. Aliquot 2x500ul of Invitrogen M280 Streptavidin beads
4. Wash 3-4X in 1X B+W (see manufacturer's instructions for binding DNA) buffer using a magnetic separator or centrifuge
5. Dilute P/C extracted digests from (2) with equal volume of 2X B+W Buffer.
6. Apply half of this mixture of each of the two tubes of beads
7. Place on tube rotator for at least 30 minutes

E. Concentration and Desalting

1. Pellet beads on magnetic separator or centrifuge and remove supernatant to clean tube. Centrifuge tube to remove remaining beads. Separate bead-free supernatant into two Amicon Ultra 0.5ml 3K spin concentrators. Spin at 14,000 g for 30 minutes. Discard flow-through and invert membrane in clean collection tube. Spin for 2 minutes at 1000g to collect concentrated solution

1. Prepare GE G50 spin desalting column by vortexing column for 30 sec, twisting off end-cap, vent top cap but leave on and spin for 1 minutes at 750g. Discard flow-through
2. Apply ~30-50ul of concentrated parts to center of column bed. Spin for 2 min at 750g to desalt sample

F. Phosphorylation of cut LTR parts

1. Quantitate cut and desalted parts (typicall 200-400ng/ul in a total volume of 40ul)
2. Setup PNK reaction with all the concentrated part solution (50-60ul total reaction volume) with 1X final T4 Ligase Buffer (source of ATP for PNK) and 2ul PNK
3. Incubate at 37C for 1hr (60 min or PCR block or 70 minutes in 37C incubator)
4. Calculate concentration of parts in PNK reaction based on input concentration and dilution factor in reaction

G. Part assembly reaction

1. Setup multiple ligation reactions with 1.5ug of cut parts, 20ul 50% PEG, (0.5ug-2ug HPcap), 1X final T4 Ligase Buffer, 1ul T4 Ligase, dH20 to a 50ul total reaction volume. Slowly mix by pipetting until you can low longer notice swirls
2. The HPCap is 5' phosphorylated and kept in aliquots at -20C. Just prior to use, 200ng/ul in NEB Buffer #2. HPCap is folded through kinetic trapping by heating to 95C for 4 minutes followed by snap-cooling on ice.

H. Gel analysis and product recovery

1. Prepare a 2% agarose gel and cast with a single large tooth comb (with the thick comb to maximize well volume)
2. Run 1kb and Low MW ladder, the extended parts (1ug is plenty), HP (again 1ug is plenty), HP only control, an empty lane, ligation reaction with parts

3. For initial diagnostic tests, run for 24-30 minutes to fully resolve product size. For preparative applications, run for ~15 to provide minimal separation. This limits the amount of gel material that must be processed.
4. Choose the reactions (lanes) that provide the desired size distribution and make horizontal cuts across selected lanes to bracket products into three different size ranges (~200-500bp, 500bp-1.5kb, 1.5-4kb).
5. Gel extract fragments using standard gel extraction. Pool slices according to original size brackets.
6. P/C extract gel extraction eluant.