Title
Epigenetic fluctuations underlie gene expression timescales and variability

Author
Lannan, Ryan

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Epigenetic fluctuations underlie gene expression timescales and variability

A dissertation submitted in partial satisfaction of the requirement for the degree of Doctor of Philosophy in Biochemistry, Molecular and Structural Biology

by

Ryan Lannan

2019

ABSTRACT OF THE DISSERTATION

Epigenetic fluctuations underlie gene expression

timescales and variability

by

Ryan Lannan

Doctor of Philosophy in Biochemistry, Molecular and Structural Biology

University of California, Los Angeles, 2019

Professor Roy Wollman, Chair

The human body is a patchwork of tissues working collectively to maintain homeostasis and achieve fitness objectives. Within these tissues are cells, the smallest units of life, which produce a tissue's morphology through emergent behavior defined by a specific pattern of gene expression. However, the observation and interpretation of this behavior clash, creating serious contradictions that need to be resolved. Isogenic populations of cells exhibit significant phenotypic heterogeneity but are still treated as monolithic entities exhibiting homogenous behavior under equilibrium. Holding this contradiction together are theories of cell behavior that posit stochastic fluctuations as responsible for heterogeneity, treating it as different snapshots of a memoryless, dynamical, process. To test this theory, we've tested the persistence of distinct cellular states, such as the allele-specific variability of an exogenous reporter system and the

intracellular calcium response to ATP, finding a lack of ergodicity. For the former we have also measured the co-fluctuation of underlying chromatin states, linking expression variability to variability in histone modifications. This work challenges existing paradigms of cellular heterogeneity, implying deeper regulation of cell state and more functionally stratified cell biology.

The dissertation of Ryan Lannan is approved.

Alexander Hoffmann

Guillaume Chanfreau

Margot Quinlan

Roy Wollman, Committee Chair

University of California, Los Angeles

2019

DEDICATION

To my mentors in secondary and undergraduate education, thank you for fanning the spark that has become a lifelong interest in science.

To Roy Wollman. Thank you for your boundless excitement about research, the philosophy that "science is fun", and your continued faith in me.

To my friends in the Wollman lab, thank you for all the fun distractions, and advice.

To my friends both near and far, for making sure I have a home wherever I go.

To my parents and family for their support of my education.

To my cat, who is a complete goof but that's why I love him.

To Tina Kantaria, for believing in, and manifesting, my best self.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Roy Wollman assisted with the conception of the projects and guided the direction of the work.

Anna Pilko assisted with the molecular biology and cell line construction for the projects.

Alok Maity assisted with the modeling and computational work in Chapter 1.

Jason Yao mentored me during the conception of Chapter 2, guiding its early work.

The members of the Wollman lab have provided advice and help through my years with them.

VITA

2014	B.S. in Biochemistry, University of Texas at Austin, Austin, TX
2014-2016	Graduate student researcher, University of California, San Diego, CA
2016	M.S. in Chemistry, University of California at San Diego, San Diego, CA
2016-2019	Graduate student researcher, University of California, Los Angeles, CA

PRESENTATIONS

Why is there so much heterogeneity in protein levels across cell populations?
Oral presentation for LMU Biology Department seminar series
Loyola Marymount University, Los Angeles, CA, September 2019

Quantifying short-term fluctuations in gene expression
Oral presentation for 2016 UCLA QCBio Retreat
UCLA QCBio Retreat 2016, Los Angeles, CA, September 2016

Contribution of epigenetic factors to protein level variability and memory
Poster presentation for 2019 qBio Conference
SFSU, San Francisco, CA, August 2019

Introduction

Since the discovery of cells as the smallest unit of life, biology has aimed to catalog and classify cells by form and function. Prior to quantitative biological measurements, early classification of human cells relied on qualitative visual differences in morphology and behavior. But the dawn of molecular biology brought with it a new understanding, defining “cell types” as groups of cells exhibiting distinct patterns of gene expression, of which morphology and behavior are emergent properties. Using this definition, researchers have identified over 210 different cell types in the human body [14]. Up until recently, these types were each treated as the limit of functional stratification within the body’s hierarchical organization. Supporting this view were the techniques available to researchers. Measurements of cell populations oftentimes captured ensemble averages, and this data was interpreted with the assumption that this reflected the biological mechanisms acting within individual cells.

The advent of technologies such as fluorescent tagging and single-cell genomics have produced a wealth of information undermining this assumption. Even within isogenic populations of cells, significant cellular heterogeneity exists, and single-cell transcriptomic studies affirm that cells occupy a large and varied landscape of possible states [4,7,9,15]. But does this heterogeneity provide meaningful biological information, i.e., is this functional heterogeneity [1]? The definition of “functional” may be vague, but it is safe to say that heterogeneity that affects cell fate, or any other irreversible phenotypic consequences, would qualify. Cellular heterogeneity in cell fate decision-making is seen in eukaryotic systems such as the selection of color photoreceptors in *Drosophila* [8] and the selection of blood cell lineages in mammalian cells [5]. Resistance to cancer drugs has been linked to epigenetic states inhabited by subpopulations of cells [10,11].

Where do we draw the line between this functional heterogeneity and the concept of cell types? One starting point may be whether this heterogeneity is stochastic or deterministic, that is, whether these cells are receiving the same molecular instructions and behaving differently due to stochastic chemical processes, or whether these cells are receiving different instructions, leading to different outcomes [12]. If there are persistent deterministic differences between isogenic cells, it would deconstruct the concept of cell types and raise doubts as to whether they are discrete entities [6]. Persistent cellular identities would also challenge studies relying on the assumption of ergodicity [2,13]. The assumption of ergodicity allows researchers to assume that snapshots of cell populations accurately represent the time averaged behaviors of small numbers of cells, as each cell is able to move through the entire state space [3]. An approach that doesn't rely on this assumption is decomposing heterogeneous distributions into more homogenous subpopulations [16]. The two-state model of gene expression relies on both the assumption of ergodicity, and the assumption that intrinsic protein-level heterogeneity is the product of stochastic RNA bursting. By testing the persistence and determinism of gene expression variability, we can better understand its emergent property, phenotypic variability.

In order to study cellular heterogeneity, we developed a reporter system designed to isolate allele-specific gene expression. We imaged this system to measure the persistence of allele-specific expression identity and compared this data to *in silico* models. We then isolated distinct subpopulations from our reporter system and let them relax over long timescales. This relaxation was correlated with allele-specific chromatin states. In further work we also measure the timescale of identity of the intracellular calcium response to stress, isolating a persistent identity generated by pre-existing factors as opposed to intrinsic fluctuations. From this body of work, we propose more deterministic mechanisms underlying phenotypic and expression variability that challenge current models of gene expression and imply more granular functional stratification in multicellular systems.

References

1. Altschuler, Steven J., and Lani F. Wu. 2010. “Cellular Heterogeneity: Do Differences Make a Difference?” *Cell* 141 (4): 559–63.

2. Brenner, Naama, Erez Braun, Anna Yoney, Lee Susman, James Rotella, and Hanna Salman. 2015. “Single-Cell Protein Dynamics Reproduce Universal Fluctuations in Cell Populations.” *The European Physical Journal. E, Soft Matter* 38 (9): 102.

3. Brock, Amy, Hannah Chang, and Sui Huang. 2009. “Non-Genetic Heterogeneity--a Mutation-Independent Driving Force for the Somatic Evolution of Tumours.” *Nature Reviews. Genetics* 10 (5): 336–42.

4. Buganim, Yosef, Dina A. Faddah, Albert W. Cheng, Elena Itskovich, Styliani Markoulaki, Kibibi Ganz, Sandy L. Klemm, Alexander van Oudenaarden, and Rudolf Jaenisch. 2012. “Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase.” *Cell* 150 (6): 1209–22.

5. Chang, Hannah H., Martin Hemberg, Mauricio Barahona, Donald E. Ingber, and Sui Huang. 2008. “Transcriptome-Wide Noise Controls Lineage Choice in Mammalian Progenitor Cells.” *Nature* 453 (7194): 544–47.

6. Huang, Sui. 2009. “Non-Genetic Heterogeneity of Cells in Development: More than Just Noise.” *Development* 136 (23): 3853–62.

7. Kumar, Roshan M., Patrick Cahan, Alex K. Shalek, Rahul Satija, Ajay DaleyKeyser, Hu Li, Jin Zhang, et al. 2014. “Deconstructing Transcriptional Heterogeneity in Pluripotent Stem Cells.” *Nature* 516 (7529): 56–61.

8. Losick, Richard, and Claude Desplan. 2008. “Stochasticity and Cell Fate.” *Science* 320 (5872): 65–68.

9. Moignard, Victoria, Iain C. Macaulay, Gemma Swiers, Florian Buettner, Judith Schütte, Fernando J. Calero-Nieto, Sarah Kinston, et al. 2013. “Characterization of Transcriptional Networks in Blood Stem and Progenitor Cells Using High-Throughput Single-Cell Gene Expression Analysis.” *Nature Cell Biology* 15 (4): 363–72.

10. Shaffer, Sydney M., Margaret C. Dunagin, Stefan R. Torborg, Eduardo A. Torre, Benjamin Emert, Clemens Krepler, Marilda Beqiri, et al. 2017. “Rare Cell Variability and Drug-Induced Reprogramming as a Mode of Cancer Drug Resistance.” *Nature* 546 (7658): 431–35.

11. Sharma, Sreenath V., Diana Y. Lee, Bihua Li, Margaret P. Quinlan, Fumiyuki Takahashi, Shyamala Maheswaran, Ultan McDermott, et al. 2010. “A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations.” *Cell* 141 (1): 69–80.

12. Symmons, Orsolya, and Arjun Raj. 2016. “What’s Luck Got to Do with It: Single Cells, Multiple Fates, and Biological Nondeterminism.” *Molecular Cell* 62 (5): 788–802.

13. Thomas, Philipp. 2017. “Making Sense of Snapshot Data: Ergodic Principle for Clonal Cell Populations.” *Journal of the Royal Society, Interface / the Royal Society* 14 (136). https://doi.org/10.1098/rsif.2017.0467.

14. Trapnell, Cole. 2015. “Defining Cell Types and States with Single-Cell Genomics.” *Genome Research* 25 (10): 1491–98.

15. Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. 2014. “The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells.” *Nature Biotechnology* 32 (4): 381–86.

16. Yao, Jason, Anna Pilko, and Roy Wollman. 2016. “Distinct Cellular States Determine Calcium Signaling Response.” *Molecular Systems Biology* 12 (12): 894.

CHAPTER 1

Persistence of gene expression variability and underlying factors

Abstract

Isogenic populations of mammalian cells exhibit significant gene expression variability. This variability can be separated into two components, allele-specific and global processes. A popular theoretical model, the two-state model of gene expression, explains expression heterogeneity as arising from stochastic RNA bursting, an allele-specific process. However, this model does not leave room for cell state variability and assumes ergodicity of cell populations. We created a reporter system that isolates allele-specific variability and measures its persistence in imaging and long-term fluctuation analysis experiments. Our fluctuation analysis experiment was then used to measure the enrichment of histone modifications associated with transcriptional activation, such as H3K4me3, and discovered co-fluctuation with expression. This work challenges the two-state model, and posits a competing model that accounts for chromatin fluctuations as an upstream regulator of expression variability.

Introduction

When faced with a changing environment, cells must constantly attune their responses to match inputs and maintain homeostasis. Naively, one would expect that clonal populations faced with a homogenous environment would have an optimized, monolithic, response. However, this is not the case with single protein expression, which varies widely across clonal populations of mammalian cells. This variability may provide functionality, producing specialized

roles in complex tissues [1,15,14]. However, this variability can also have drastic consequences for cell fate and phenotype [8,25]. Determining the degree to which variability is due to defined cellular stratification or due to stochastic processes requires decomposing the contributions towards protein expression variability.

The empirical split in protein expression variability is between extrinsic sources and intrinsic sources of variability. This split has been defined in the context of covariance between reporter genes, where covarying signal is defined as extrinsic variability and the non-covarying signal is defined as intrinsic variability [16]. In a cellular context, we would refer to extrinsic variability as global factors related to the underlying cell state, and intrinsic variability as allele-specific factors. Not represented in this dichotomy are trans-regulatory elements, which can be quantified as intrinsic or extrinsic variability depending on the presence of identical regulatory elements in both genes of interest. Without controlling for this regulation, upstream allele-specific variability can influence downstream variance [35]. However, when using two reporters with identical regulatory elements, the remaining intrinsic variation has two possible contributors, cis-regulatory elements and stochastic RNA bursting.

The current model of gene expression, the two-state model, assumes that intrinsic variability is defined by stochastic RNA bursting, creating an over-dispersed, non-Poisson distribution of expression [10,11,36]. The initial evidence for bursting protein expression was heterogeneous responses to stimuli [24], and was only directly observed thanks to live imaging technologies nearly two decades later [41]. According to the two-state model, these “bursts” are caused by stochastic transitions between a gene’s ON and OFF states [23]. When ON, a gene produces a burst of mRNA transcripts, which subsequently causes a sharp increase in translation further downstream. Intuitively, we understand this model because of the low copy

number of DNA and other molecules regulating expression, making transcription dependent on stochastic chemical processes [16,30].

The two-state model accounts for the high level of variability seen across cell populations, intuitively follows our understanding of transcription, and is consistent with observations tracking transcript number in single cells [28]. Yet this model makes the implicit assumption that cells are identical and interchangeable once extrinsic factors are removed [38]. This assumption is made not due to evidence of absence of distinguishing features, but due to a lack of high resolution data. As resolution in single-cell data increases, the possibility that a "hidden variable" explaining this heterogeneity will be found also increases. The local chromatin environment has been implicated in impacting protein expression at different genomic positions [13]. Furthermore, specific transcription start site (TSS) associated histone modifications such as H3K4me3 and H3K79me2 have been shown to have impacts on gene activity [31] and expression variability [5] respectively. If the impacts of these variables have any heterogeneity across cell populations, then they could act as potential "hidden variables" creating a deterministic distribution of expression states.

Another consequence of the assumptions inherent to the two-state model is that intrinsic protein expression must have a rapid mixing timescale. The stochastic transitions of genes between ON and OFF states are memoryless in the two-state model, so the timescale of mixing must follow the timescale of protein dilution and degradation. This makes any expression identity persisting longer than cell cycle suspect, as it would seem to contradict the two-state model. Sigal et al., 2006 saw persistent expression identity but attributed it to upstream trans-factors [35]. Phillips et al., 2019 saw correlations in transcriptional activity between sister cells [29], but couldn't determine whether inherited factors or cellular microenvironment were

responsible for this phenomenon. One possible refutation of the two-state model would be the identification of a long term expression identity that is directly attributed to a specific cis-regulatory mechanism.

Using imaging on a two reporter system, we have determined that allele-specific variability decays at a timescale longer than predicted by the two-state model. A new model that incorporates this finding would need to both explain the breadth of allele-specific variability while also providing a mechanism that accounts for the slow fluctuations in expression identity. Our model incorporates a modular transcription rate (*Km(t)*) that is heterogeneous across a cell population and proposes that this modularity is tuned through the presence of different levels of histone modifications. To test this model, we implemented a fluctuation analysis technique that allowed for the measurement of allele-specific gene expression across ten days. This technique was then utilized to measure the levels of histone modifications using ChIP-qPCR. This fluctuation analysis demonstrates the co-fluctuation of allele-specific gene expression variability and the histone modification H3K4me3 at the promoter region, demonstrating the incompleteness of the two-state model. This cis-fluctuating component may act as a “hidden variable” creating a distribution of expression states within a population, implying more defined cellular stratification within clonal populations than previously predicted.

Results

In order to accurately quantify the contribution of allele-specific variability, we need to decompose the contributions towards protein expression variability using a multiple reporter system [16]. In a two reporter system, covariance between the reporters is defined as extrinsic variability while non-correlated noise maps to allele-specific variability for each reporter. Given

our interest in quantifying allele-specific variability for one reporter, but not the whole system, our design incorporates multiple copies of our co-varying reporter. This eliminated the allele-specific component for our co-varying reporter by diluting the cis-effects affecting each copy, leaving a signal more defined by global fluctuations.

These global fluctuations can affect system dynamics and stochasticity, creating statistical dependence between intrinsic and extrinsic noise [33]. Given this we represented the relationship between intrinsic and extrinsic variability using the definition of variance and the double expectation theorem, shown in Equation 1. Here, A represents a random variable, such as our singly-integrated reporter, while C represents the extrinsic factors influencing A, empirically defined by our multiply-integrated covarying reporters.

$$\underset{total}{Var(X)} = \underset{intrinsic}{E\big(Var(A|C)\big)} + \underset{extrinsic}{Var(E(A|C))} \qquad \text{(Equation 1)}$$

We applied this equation to single cell data through binning on C and calculating appropriate expectation values and variability, isolating a conditioned A signal (Supplementary Fig. 1.1). This data was gathered on a system designed in K562, a suspension cell line derived from chronic myelogenous leukemia. We performed long term imaging (41 hours) on our system in order to quantify the timescale of allele-specific variability. The signal of our singly integrated reporter was conditioned on our multiply integrated co-varying reporters, removing extrinsic factors including trans-regulatory components, and isolating its allele-specific components (Fig. 1.1a). We then calculated the autocorrelation of these conditioned cell traces over the imaged time span (Fig. 1.1d).

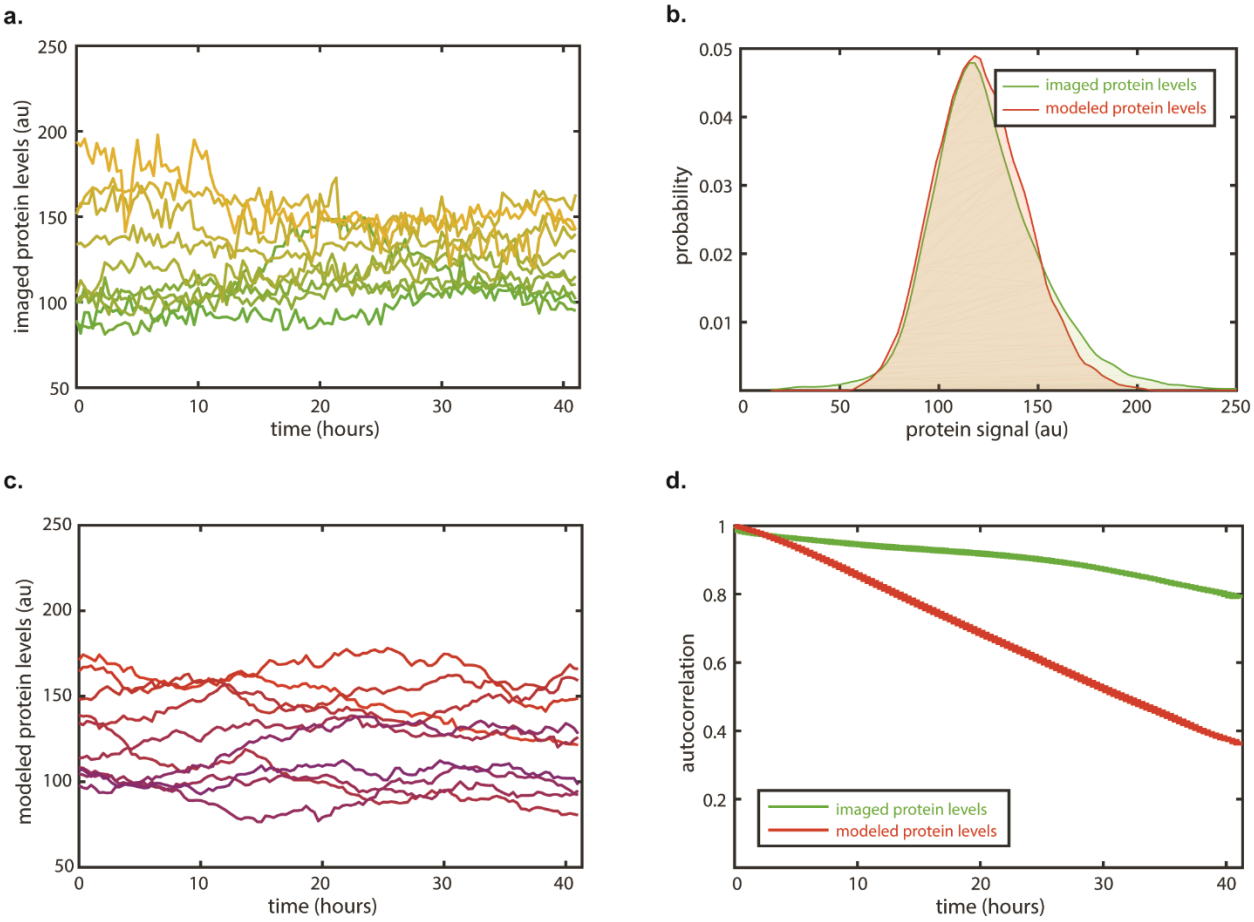


Figure 1.1: *In vitro* imaging shows persistent expression identity longer than cell cycle

(a) Cell traces of singly-integrated fluorescent reporter (mVenus) conditioned on multiply-integrated reporter signal (mTurquoise2) in K562. Both reporters are regulated by the ubiquitin promoter. Fluorescence imaging was performed for 41 hours. (b) Distributions of *in vitro* system and *in silico* two-state model traces at the first time point. Cost function for simulated annealing was evaluated on the absolute difference between the curves when binning on protein signal. (c) Cell traces of *in silico* two-state model fit to distribution of *in vitro* system using simulated annealing. (d) Autocorrelation plots of *in vitro* system traces (mVenus) versus *in silico* two-state model traces over 41 hours.

To contextualize this autocorrelation against the two-state model we developed a simulation comprised of seven first order biochemical reactions that capture its structural assumptions (Supplementary Table 1.1). From our starting parameter values we applied a Gillespie stochastic simulation [18] to generate single cell protein level traces (Fig. 1.1c). The distribution of these values was compared against the distribution of experimental values to generate a cost function (Fig. 1.1b) Simulated annealing was used to find an optimal parameter space for five parameters until our cost function showed less than 10% difference between experimental and simulated distributions (Fig 1b) (final values in Supplementary Table 1.1). The autocorrelation of the optimized traces was then compared to imaging data (Fig. 1.1d) showing a more persistent identity exists *in vitro* than in our *in silico* theoretical assumptions.

This slow timescale of expression identity could originate from an upstream component (Sigal, 2006). As our reporter system shares trans-regulatory elements, they cannot be responsible for the slow timescale. However, TSS-associated histone modifications such as H3K4me3 or H3K79me2 remain as possibilities. A distribution in enrichment of these modifications could lead to impacts in downstream gene expression, contributing both to the breadth of expression variability as well as its long relaxation timescale. We developed a simulation exemplifying this hypothesis (Fig. 1.2b), altering the original model (Fig. 1.2a) by adding an upstream component influencing a time-dependent transcription rate, translating into variability in burst size. Synthesis and decay processes were incorporated into our model for the upstream component, representing enrichment and removal of histone modifications. These random birth/death events create a time-dependent transcription rate for any given cell (Fig. 1.2c), which has precedence in previously published models [12,21,34].

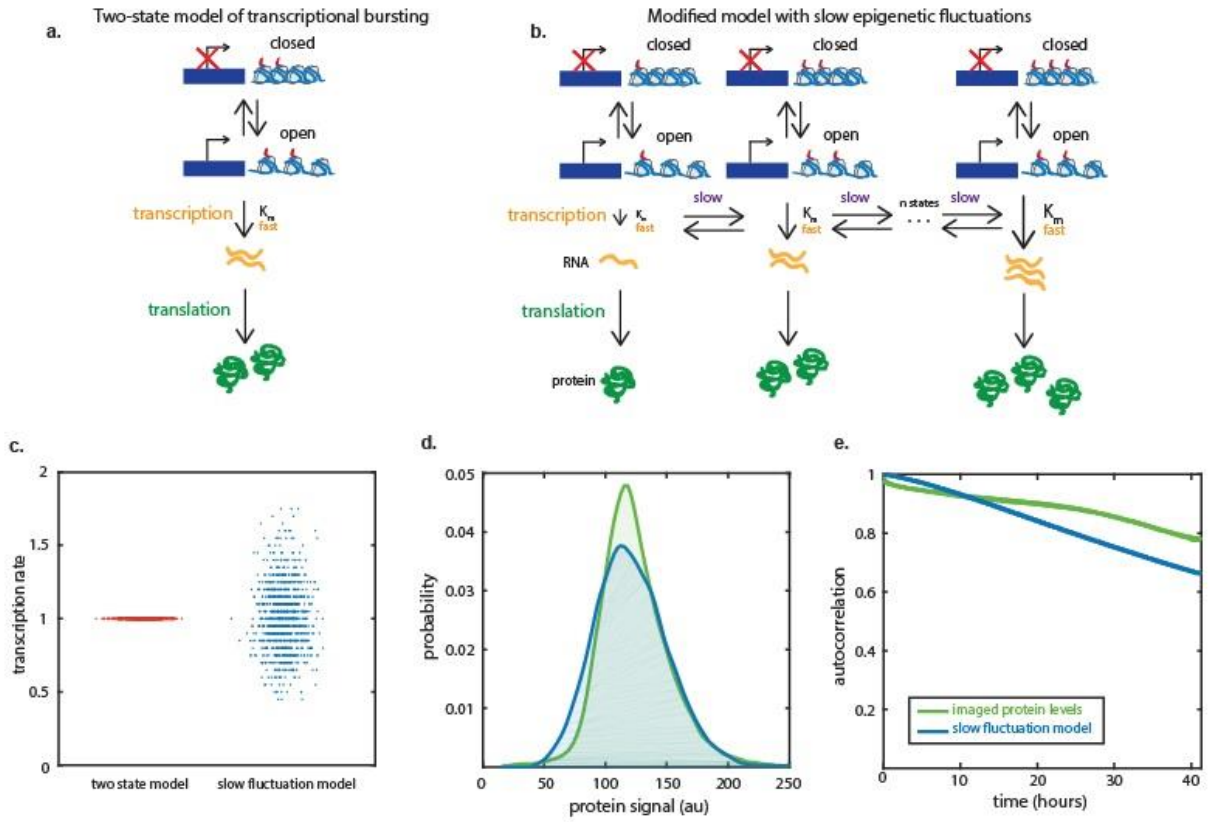


Figure 1.2: Slow fluctuating epigenetic factors can explain persistent timescale

(a) Representation of the two-state model of gene expression. (b) Representation of our slow-fluctuation model, with a spectrum of states holding different transcription rates. (c) Plot showing distribution of transcription rates for both the two-state model and our slow fluctuation model which was fed into their respective simulations. (d) Distributions of *in vitro* system and our *in silico* slow fluctuation model traces at the first time point. Cost function for simulated annealing was evaluated on the absolute difference between the curves when binning on protein signal. (e) Autocorrelation plots of *in vitro* system traces (mVenus) versus *in silico* slow fluctuation model traces over 41 hours.

We tested our slow cis-fluctuation model in a parallel approach to the one used to validate the two-state model. Our new model consists of nine first order biochemical reactions that can be subjected to a Gillespie stochastic simulation in order to generate single cell protein traces (Supplementary Table 1.2). The distribution of these traces was also compared against experimental data to generate a cost function (Fig. 1.2d). Simulated annealing was used to find an optimal parameter space for five parameters while the remaining four parameters were preset (final values in Supplementary Table 1.2). The autocorrelation of the optimized traces was compared to imaging data (Fig. 1.2e), demonstrating the slow cis-fluctuation model's confluence with our *in vitro* study. This demonstrates the plausibility of an upstream component influencing a time-dependent transcription rate as an underlying mechanism contributing to both the distribution of allele-specific variability and its long timescale.

Verifying our slow cis-fluctuation model requires further characterization of this timescale and exploration of its mechanistic underpinnings. Isolating this slow timescale can be accomplished by multi-day tracking, removing the influence of short-term promoter fluctuations. Exploring its mechanism requires a form of tracking that provides for batch analysis. To cover both we developed a fluctuation analysis protocol that isolates discrete gene expression states, namely the top, middle, and bottom quintile of the allele-specific expression distribution, and tracks their expression across ten days using flow cytometry (Fig. 1.3a). Isolated quintiles were conditioned on covarying signals through modified fluorescence-activated cell sorting (FACS) (Fig. 1.3b). BD FACSDiva files were written by MATLAB scripts, creating customizable gates. Internal controls were introduced by permanent lentiviral fluorescent labelling of the original cell population and spiking isolated quintiles with these labeled subpopulations, allowing for the removal of well-to-well variability by normalizing quintile signal with the full labeled population (Fig. 1.3c).

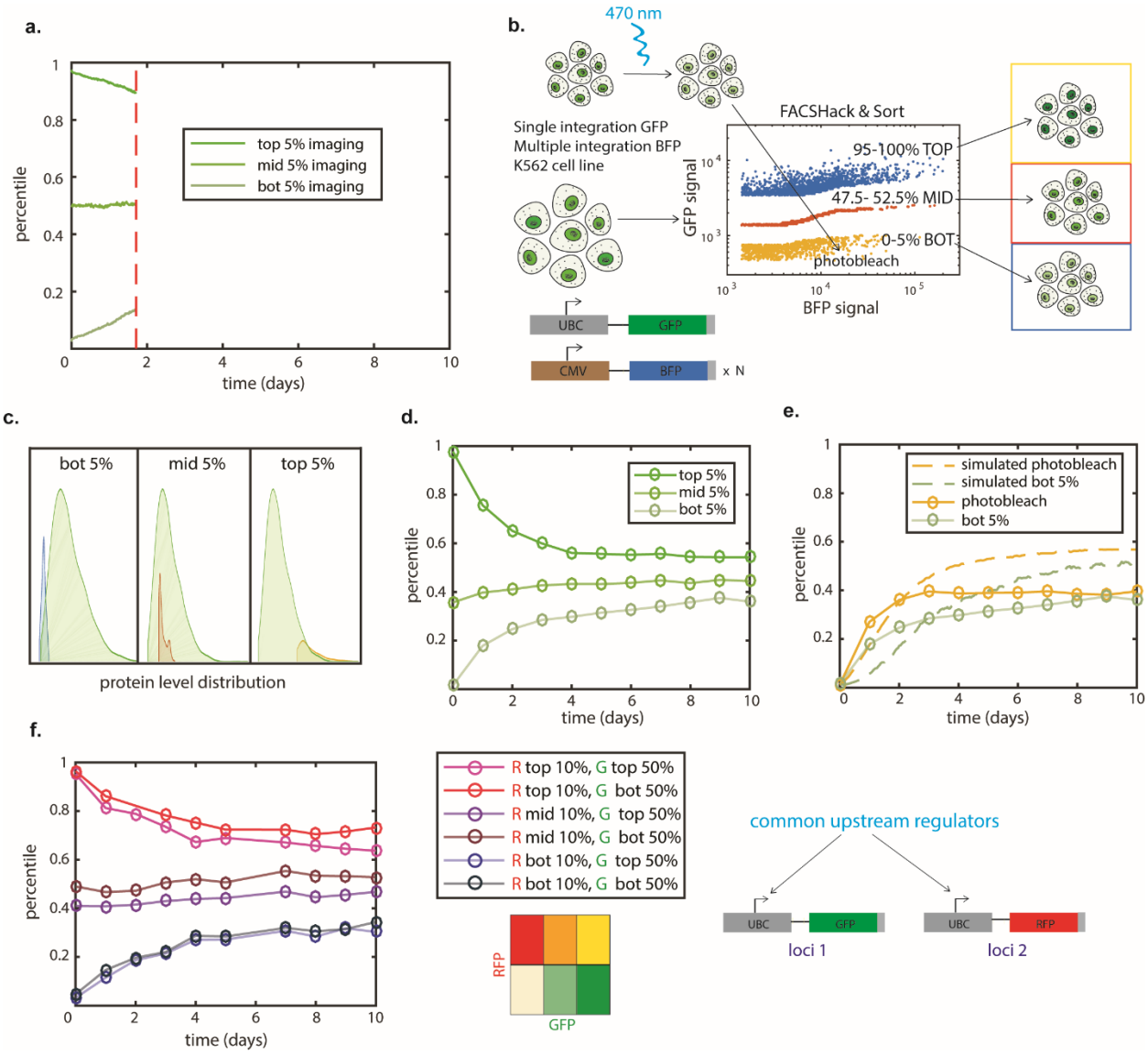


Figure 1.3: Multi-day expression identity not due to upstream transcription factors or stochastic fluctuations

(a) Cell traces of imaging (mVenus conditioned on mTurquoise2 signal) separated into the bottom, middle, and top quintiles and tracked by percentile over 41 hours. **(b)** Diagram of multi-day fluctuation analysis protocol using FACS. Example input population is a K562 system containing singly-integrated EGFP reporter and multiply-integrated TagBFP reporters, each regulated distinctly (ubiquitin and CMV respectively). Bottom, middle and top quintiles of EGFP expression as conditioned on TagBFP signal using our FACSHack were isolated. Photobleach control population was isolated simultaneously on same gate as bottom quintile. **(c)** Internal control (lentivirally labeled) populations are shown as green histograms with the isolated populations (no labelling) they were spiked into for normalization of inter-well variability and percentile measurement. **(d)** Relaxation of EGFP signal after conditional isolation by FACS. Initial timepoint based on FACS isolation data. Subsequent data gathered from daily flow

experiments, with percentiles being determined by normalization against spiked population. **(e)** Relaxation of EGFP signal for photobleach control experiment. Control generated by artificially lowering fluorescence signal of full reporter system population using 474 nm light too approximately lowest quintile signal. Relaxation compared against relaxation of naturally isolated bottom quintile. *In silico* populations also shown, simulated using slow fluctuating model and recapitulating similar dynamics. **(f)** Relaxation of EGFP signal for control utilizing shared upstream regulation. Input system contains singly-integrated EGFP reporter, singly-integrated tdTomato reporter and multiply-integrated TagBFP reporters. Both EGFP and tdTomato share regulation by ubiquitin promoter. Input isolation used standard conditioning on TagBFP signal as well as differential isolation on all permutations of top half/bottom half tdTomato as well as top decile, middle decile and bottom decile EGFP.

Several reporter systems were evaluated using our fluctuation analysis protocol. Figure 1.3d depicts the analysis of a system with a single reporter regulated by the ubiquitin reporter and covarying multiple reporters regulated by the CMV reporter. These relaxation curves demonstrate multi-day timescales, with the top quintile demonstrating a half-life of 0.95 days and the bottom quintile demonstrating a half-life of 1.40 days. However, this system doesn't adequately remove trans-regulation and doesn't quantify the relative timescale of allele-specific promoter fluctuations. To exclude the possibility of trans-regulation, a second system (Fig. 1.3f) demonstrates the lack of influence of a co-regulated reporter. Discrete expression states were isolated through two conditioned gates, the first isolating the top half and bottom half of expression for a RFP reporter and then separated into the top, middle and bottom deciles for an eGFP reporter, both of which are co-regulated by the ubiquitin reporter. The lack of significant change in timescale or relaxation for any given eGFP decile given different RFP expression levels demonstrates a lack of correlation and implies that the slow timescale is due to an allele-specific factor, not upstream trans-regulation.

Assuming that the distribution of allele-specific variability is due to both the heterogeneity generated by promoter fluctuations and a cis-fluctuating component, by isolating discrete expression states we would also enrich for specific states within the distribution of this component. However, the full expression distribution would average out these states, so if this

distribution was artificially depressed, its relaxation to the mean wouldn't be defined by the slow component, it would be defined by protein birth/death rates and promoter fluctuation timescales. To create an artificially depressed state, we used photobleaching prior to FACS sorting, altering the reporter signal to be fluorescently identical with the bottom quintile of the expression distribution, then isolated the cells with the same gate as the bottom quintile (Fig. 1.3b). Relaxation of this artificially depressed population is more rapid than the bottom quintile of allele-specific variability, suggesting a fundamental difference between these populations (Fig. 1.3e). This confirms the presence of a cis-fluctuating component with a multi-day timescale that impacts gene expression variability.

Identifying the mechanism of our cis-fluctuating component requires the linking of an epigenetic and allele-specific factor to gene expression variability. TSS-associated histone modifications like H3K4me3 or H3K79me2 are productive avenues of exploration due to their known impact on gene expression across different loci [5,31]. Moving down this avenue, our approach was using our fluctuation analysis protocol to gather large numbers of cells from discrete expression states and assess their enrichment for likely histone modifications immediately upon isolation, and after ten days of relaxation (Fig. 1.4a). Isolation was performed on the bottom, middle and top deciles of ubiquitin-regulated EGFP signal as conditioned on forward scatter. Assessment was performed by ChIP-qPCR at the transcription start site of the ubiquitin promoter. An immunoglobulin G (IgG) negative control showed minimal background, validating further antibody pulldowns (Fig. 1.4b). Enrichment levels of H3K79me2 remained constant across discrete expression states and across time, showing no significant pattern that can be linked to our cis-fluctuating component (Fig. 1.4c). However, enrichment of H3K4me3 is heterogeneous after initial isolation, showing a higher enrichment for the top decile of expression relative to the middle and bottom deciles (Fig. 1.4d). This pattern relaxes by day 11, showing lessened enrichment (Fig. 1.4d), which is confluent with our observations in Figure 1.3.

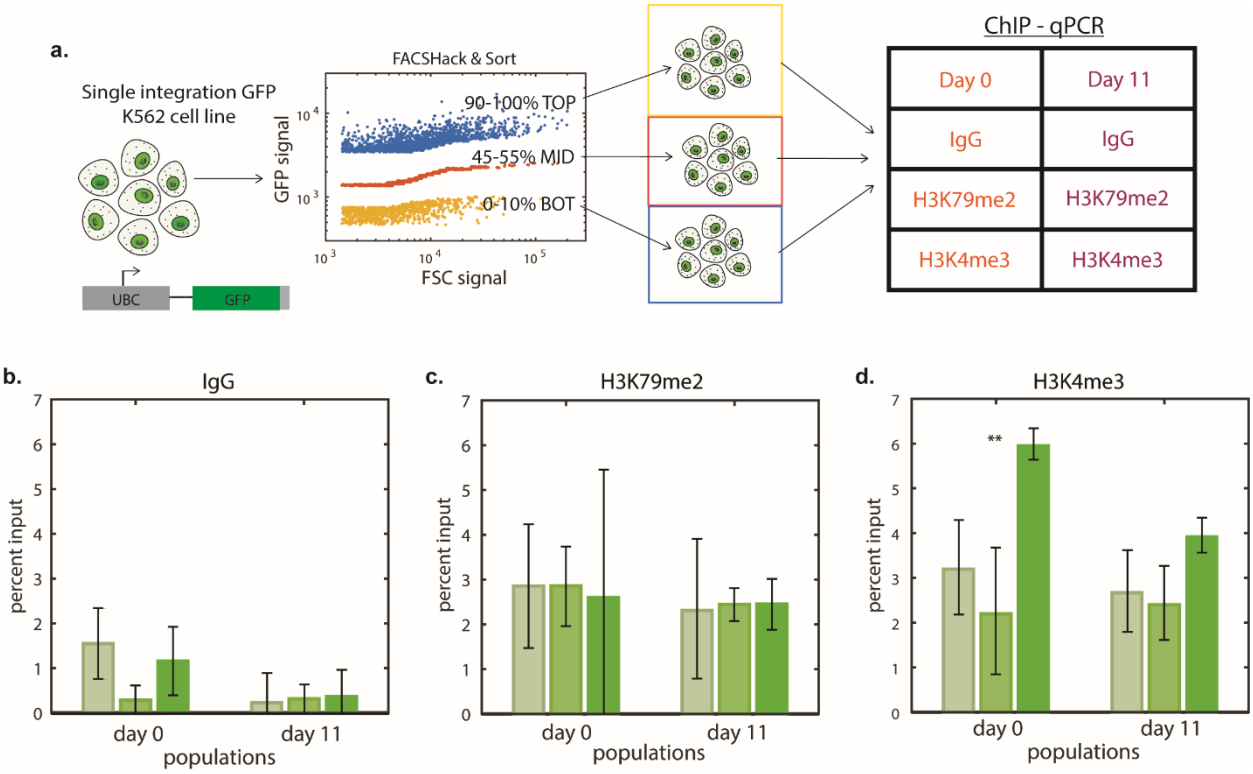


Figure 1.4: H3K4me3 enrichment co-fluctuates with gene expression variability

(a) Diagram of expression isolation protocol for ChIP analysis using FACS. Example input population is a K562 system containing singly-integrated EGFP reporter. Bottom, middle and top deciles of EGFP expression as conditioned on forward-scatter signal using our FACSHack were isolated. Post-isolation, populations were separated for analysis immediately following isolation as well as analysis ten days after isolation. Populations were immunoprecipitated by IgG negative control, anti-H3K79me2, and anti-H3K4me3 antibodies. **(b)** Bar graph showing relative enrichment as measured by qPCR signal (percent input as metric) with the negative control IgG antibody **(c)** Bar graph showing relative enrichment as measured by qPCR signal (percent input as metric) with the anti-H3K79me2 antibody. (d) Bar graph showing relative enrichment as measured by qPCR signal (percent input as metric) with the anti-H3K4me3 antibody. Day 0 shows significant enrichment ($p = 0.0087$) for the top 10% population. Day 11 shows a relaxation of this enrichment ($p = 0.16268$).

This links our slow cis-fluctuations in allele specific variability with H3K4me3 enrichment, implying that a hidden distribution of epigenetic states underlies the breadth of gene expression variability.

Discussion

In this work we isolate a cis-fluctuating component of gene expression which we linked to the histone modification H3K4me3. Our approach was to compare the timescale of relaxation of *in vitro* allele-specific variability against both a model representing the assumptions of the two-state model and an experimental control that approximates the timescale of relaxation due to stochastic RNA bursting. In both of these cases, a long term allele-specific component persisted beyond the expectations of the two-state model. This component was then found to relax on a similar timescale as enrichment of the histone modification H3K4me3, which is associated with high levels of expression at the loci we measured. This demonstrates cellular heterogeneity in underlying chromatin states that influence gene expression, disproving the assumption that expression variability is a snapshot of identical processes stochastically fluctuating between different states.

This finding seems overdue given the breadth of work implicating positional effects and histone modification as impacting transcription regulation [4,30]. The acknowledgement that chromatin regulation impacts transcriptional activity seems to already imply a distribution of epigenetic states underlying expression variability. The likelihood of uniform enrichment levels across isogenic cell populations is low due to the same molecular stochasticity cited by the two-state model, and any deterministic effects of heterogeneous enrichment levels would be carried downstream. Despite this intuition, the two-state model is popular in the literature, with multiple papers modeling the distribution of expression variability using its assumptions [13,25,27,36].

Moving forward this approach seems incomplete, as it is clear that upstream epigenetic components can also play a role in downstream heterogeneity, but as genomic position is a large determinant of the nature of chromatin regulation at a given loci, this is not an absolute across all cases.

The persistence of gene expression identities in select populations of cells also challenges the ergodic hypothesis; that a subset of a population of cells can replicate the behavior of the full population given time. Invalidating this assumption has consequences for interpretation of previous work [6,39], including measurements of information capacity in signaling networks [9,32]. Acknowledging that isogenic populations of cells have a spectrum of long-term identities also has implications for our understanding of developmental biology, reinterpreting the deep wells of Waddington's landscape as shallow and multistable. From a systems interpretation of cell state, you could say that multiple cell states exist within isogenic populations [17].

Alternatively, you could treat this as confluent with recent reports of broad gene expression continuums within cell types [7,20,40]. That selective forces haven't removed this defined heterogeneity implies functionality. Extremes of functional expression continuums could serve as specialist types within a spatially stratified tissue [1], or allow for population-level computation [14]. Lastly, traditionally defined cell types could contain snapshots of "pre-differentiation" states that are actively undergoing novel canalisation.

Novel canalisation of the epigenetic landscape seems to match data from our fluctuation analysis of allele-specific gene expression in our system (Fig. 3d, Fig 3f). The lack of ergodicity of our discrete expression states implies semi-permanent expression identities that could map

to “pre-differentiation” subpopulations of cells. This provides a new context to questions of cell fate as well as a new approach to cancer cell reprogramming [19,37]. However, we’ve only linked distinct expression identities with histone modifications for the top quintile of allele-specific expression, not the bottom quintile. That H3K4me3 doesn’t have depressed enrichment for the bottom quintile of expression (Fig. 4d) may imply bistable chromatin regulation poised between the middle and top quintiles of expression. This asymmetry of cell fate may be only part of the bigger picture, as this doesn’t discount different overlapping loci-specific mechanisms determining expression variability between the bottom and middle quintiles.

While the mechanisms we’ve discovered don’t produce “functional” variability at our contrived loci [2], these same mechanisms are loose in the genomic environment. This means their influence is likely felt at other loci which could generate functional diversity. While protein network architecture can act as an effective regulator of expression noise [22], tightly regulated distributions of chromatin states across a cell population could act as a secondary constraint on variability for housekeeping genes, possibly loosening given ageing [3]. And at other loci where variability is desirable in order to respond more effectively to fluctuating environments, distributions in histone modifications could act as diversity generating in order to stimulate cellular stratification within isogenic populations.

Recent advances in single-cell measurements will allow for greater resolution into underlying cell states. The development of technologies such as single-cell ChIP may make batch methods like the ones used in this work no longer necessary for measuring histone modification levels effectively, allowing for far more nuance and throughput in population level heterogeneity measurements. It is an obvious epistemic bias that the importance of cellular heterogeneity has become evident now, just as more single-cell measurements become

available. However, the mental models researchers use to assess this data aren't contingent on its quality. The understanding that functional stratification of isogenic cell populations exists is key to interpreting population level data and moving towards less reductive models of gene expression that don't rely on RNA bursting to explain over-dispersed variability. The introduction of contingent position effects as time-dependent upstream factors can better account for the persistence of expression identity and connects better to the nascent literature on enhancers, histone modifications, and other cis-regulated elements.

Acknowledgements

Thank you to Anna Pilko for her help with cell line construction, Alok Maity for his efforts on modeling work and Roy Wollman for assistance with conception and direction.

Lead contact and materials availability

Further information and requests for resources and reagents should be directed to, and will be fulfilled by the corresponding author Roy Wollman (rwollman@ucla.edu).

Materials and methods

K562 cell culture

A K562 suspension cell line provided by Sigma-Aldrich was grown at 37 °C in RPMI 1640 medium (Gibco) supplemented with 10% FBS (Gibco), 1% penicillin-streptomycin (Gibco) and 1% GlutaMAX (100x) (Gibco) under 5% CO2 atmosphere.

Reporter plasmids

The following elements were included in the base plasmid in order to allow for viral packaging and integration. HIV-1 truncated 5' LTR, HIV-1 packaging signal, HIV-1 Rev response element (RRE), HIV-1 truncated 3' LTR and Central polypurine tract (cPPT). Five distinct transcription elements were used as expression reporters. Ubiquitin promoter (ubi) driving mVenus, ubi-mTurquoise2, ubi-EGFP, ubi-TdTomato and cytomegalovirus promoter (CMV) driving expression of TagBFP.

Lentiviral production and cell line generation

Reporter and third generation lentiviral packaging plasmids were transfected into HEK 293T cells to generate reporter vectors. Transfected HEK293T culture supernatant was collected and concentrated by Lenti-X-concentrator (Takara) 48 hours post transfection. K562 cells were transduced with reporter-containing lentivirus in media supplemented with 5 µg/ml polybrene and 20mM HEPES for 2 hours of spinoculation and left to incubate for 24 hours. To generate singly-integrated cell lines, an MOI of 0.01 was used to ensure that the majority of transduced cells integrated with a single reporter copy. To generate multiply-integrated cell lines, multiple rounds of transduction were performed at saturating MOI. Founder cells were then singly sorted by fluorescence-activated cell sorting (FACS) at 72 hours post-transduction to generate unique cell lines.

Time-lapse microscopy

Imaging was performed using a Nikon Plan Apo λ 10×/0.45 objective with a 0.7× demagnifier and Nikon Eclipse Ti microscope with a sCMOS Zyla camera. All imaging was done using custom automated software written using MATLAB and Micro-Manager (Edelstein et al., 2010). Image analysis was accomplished using a custom MATLAB code published previously (Selimkhanov et al., 2014).

Gillespie stochastic simulation and simulated annealing

Parameter values for first order biochemical reactions were optimized by a combined approach of simulated annealing and stochastic simulation algorithms (SA+SSA). A parameter set of rate constants was generated in each SA iteration and fed into the SSA to generate cell traces. The difference of the generated distribution to the experimental distribution was compared to create our cost function:

$$Cost = \sum_{i=1}^{k} |Xexp(P_i) - Xsim(P_i)|$$

$X_{exp}(P_i)$ and $X_{sim}(P_i)$ are the probabilities of protein level at the ith point for experimental and simulation respectively, whereas *k* is the total number of grid points in the distribution of protein levels. Cost was minimized with each SA step, with a termination criteria of Cost < 0.1.

No boundary constraints were applied on parameter values during optimization. Perturbation does not exceed 5% of the previous value, following the below equation:

$$Param_{n+1} = Param_n + Param_n \; x \; (-1)^m x \; \delta \; x \; rand$$

Param represents a single parameter value, *n* represents the iteration number, *m* is a random integer [0 or 1], δ is the magnitude of perturbation (0.05) and rand is a uniform random number between 0 and 1. After perturbation, the updated parameter set is ran through SSA, generating a protein histogram that can be assessed by our cost function. If the current *Cost* value is lower than the prior value, then the updated parameters are accepted for the next SA step. Otherwise, they are subjected to the Metropolis test for judgement as to whether they should be accepted or rejected (Metropolis et al., 1953).

Internal control labelling

Lentiviral vectors were created with a reporter plasmid for cell labelling. The transcription element contains the cytomegalovirus promoter driving iRFP670 expression. Reporter cell lines were transduced with an MOI of 1.39 in order to label approximately 25% of the population.

Cells were incubated for 72 hrs to dissipate stress prior to sorting. All cells were sorted for presence of iRFP670 (as measured by APC-A). APC-A(+) cells were used as control cells while APC-A(-) cells were processed into desired quintiles downstream. APC-A(+) cells were then spiked into quintiles. Percentiles of quintile populations were then judged by their relative signal to the APC-A(+) control population in order to remove well-to-well variability.

Fluorescence activated cell sorting and flow cytometry

Sorting was accomplished with a BD FACSAria cell sorter. Tracking of expression after sorting was accomplished with a BD LSRII. Cells were initially filtered using forward scatter (FSC-A) and side scatter (SSC-A). EGFP signal was measured with FITC-A or GFP-A. TagBFP signal was measured with DAPI-A or Pac-Blue-A. TdTomato signal was measured with PI-A or RFP-A. iRFP670 signal was measured with APC-A.

FACShacking

Prior to rewriting sorting gates, reporter cell line was run through cytometer (using BD FACSDiva software), gathering population data. Gate was constructed around desired population for use as “parent” gate. Flow cytometry standard (.fcs) files were gathered from parent gate population as well as the configuration file for sorting session (.xml). Both of these files were imported to MATLAB for manipulation. The population of the parent gate was binned by the covarying reporter. In each bin, the bottom, middle, and top quintiles of our allele-specific signal were assessed, determining the minimum and maximum values of each quintile within the bin. These values were then stitched together across all bins, forming three polygons. The “parent” gate within the configuration file was used as a template to form new gates. Copies of the “parent” gate were nested one level below, then altered to fit the polygons determined from

the cell population data. This configuration file was re-introduced into the cell sorter, resetting the session with conditioned quintile gates.

Photobleach control

For the photobleach control, cells were seeded in 384 well plates to allow for full saturation by light exposure. Photobleaching was achieved through 900 repeated exposures to 474 nm light at an interval of 4 s ON 8 s OFF. After photobleaching, cells were isolated by the same gating as the bottom 5% population.

ChIP-qPCR

All ChIP experiments were preceded by the isolation of the bottom, middle and top deciles of ubiquitin-regulated EGFP signal (as conditioned on forward scatter). After isolation, cells were split into two pools, Day 0 measurement (500K cells) and Day 11 measurement (50K cells). The Day 0 pool was immediately processed by ChIP-qPCR, while the Day 11 measurement was incubated for ten days, then processed. The LowCell# ChIP protein A kit from Diagenode was used for all replicates. Fixation was performed as per kit protocol. Sonication was performed on a S220 focused-ultrasonicator from Covaris, using M220 microTUBE, snap-cap tubes. Fourteen cycles of 1 min ON/1 min OFF were performed on samples to reach a target fragmentation size of 100-200 bp.

Three antibodies were used for immunoprecipitation, Rabbit IgG polyclonal antibodies IgG (ab37415), anti-histone H3 (tri methyl K4) ChIP grade antibodies (ab8580), and anti-histone H3 (di methyl K79) ChIP grade antibodies (ab3594). DNA was isolated and purified using the IPure v2 kit from Diagenode (C03010015). Isolated DNA was then assessed by qPCR using primers that amplify -929 to -810 relative to the TSS of our reporter gene. Sequences are as follows,

Forward: acagcagagatccagtttggtta

Reverse: agtctgcttcccgcgtcc

Positive and negative control primers were bought from Active Motif, human positive control primer set ACTB-1 (71003) and human negative control primer set 1 (71001). qPCR results were analyzed relative to ChIP input percentage. Analysis of variance (ANOVA) was used to determine significance.

qPCR thermocycler protocol

1. 95C for 3:00
2. 95C for 0:05
3. Annealing T for 0:30
4. Take picture, GOTO #2 x 45
5. 65C -> 95C, 0.5C and take picture per 0:05

References

1. Adler, Miri, Yael Korem Kohanim, Avichai Tendler, Avi Mayo, and Uri Alon. 2019. “Continuum of Gene-Expression Profiles Provides Spatial Division of Labor within a Differentiated Cell Type.” *Cell Systems* 8 (1): 43–52.e5.

2. Altschuler, Steven J., and Lani F. Wu. 2010. “Cellular Heterogeneity: Do Differences Make a Difference?” *Cell* 141 (4): 559–63.

3. Bahar, Rumana, Claudia H. Hartmann, Karl A. Rodriguez, Ashley D. Denny, Rita A. Busuttil, Martijn E. T. Dollé, R. Brent Calder, et al. 2006. “Increased Cell-to-Cell Variation in Gene Expression in Ageing Mouse Heart.” *Nature* 441 (7096): 1011–14.

4. Becskei, Attila, Benjamin B. Kaufmann, and Alexander van Oudenaarden. 2005. “Contributions of Low Molecule Number and Chromosomal Positioning to Stochastic Gene Expression.” *Nature Genetics* 37 (9): 937–44.

5. Benayoun, Bérénice A., Elizabeth A. Pollina, Duygu Ucar, Salah Mahmoudi, Kalpana Karra, Edith D. Wong, Keerthana Devarajan, et al. 2014. “H3K4me3 Breadth Is Linked to Cell Identity and Transcriptional Consistency.” *Cell* 158 (3): 673–88.

6. Brenner, Naama, Erez Braun, Anna Yoney, Lee Susman, James Rotella, and Hanna Salman. 2015. “Single-Cell Protein Dynamics Reproduce Universal Fluctuations in Cell Populations.” *The European Physical Journal. E, Soft Matter* 38 (9): 102.

7. Cembrowski, Mark S., and Vilas Menon. 2018. “Continuous Variation within Cell Types of the Nervous System.” *Trends in Neurosciences* 41 (6): 337–48.

8. Chang, Hannah H., Martin Hemberg, Mauricio Barahona, Donald E. Ingber, and Sui Huang. 2008. “Transcriptome-Wide Noise Controls Lineage Choice in Mammalian Progenitor Cells.” *Nature* 453 (7194): 544–47.

9. Cheong, Raymond, Alex Rhee, Chiaochun Joanne Wang, Ilya Nemenman, and Andre Levchenko. 2011. “Information Transduction Capacity of Noisy Biochemical Signaling Networks.” *Science* 334 (6054): 354–58.

10. Corrigan, Adam M., Edward Tunnacliffe, Danielle Cannon, and Jonathan R. Chubb. 2016. “A Continuum Model of Transcriptional Bursting.” *eLife* 5 (February). https://doi.org/10.7554/eLife.13051.

11. Dar, Roy D., Brandon S. Razooky, Leor S. Weinberger, Chris D. Cox, and Michael L. Simpson. 2015. “The Low Noise Limit in Gene Expression.” *PloS One* 10 (10): e0140969.

12. Dattani, Justine, and Mauricio Barahona. 2017. “Stochastic Models of Gene Transcription with Upstream Drives: Exact Solution and Sample Path Characterization.” *Journal of the Royal Society, Interface / the Royal Society* 14 (126). https://doi.org/10.1098/rsif.2016.0833.

13. Dey, Siddharth S., Jonathan E. Foley, Prajit Limsirichai, David V. Schaffer, and Adam P. Arkin. 2015. “Orthogonal Control of Expression Mean and Variance by Epigenetic Features at Different Genomic Loci.” *Molecular Systems Biology* 11 (5): 806.

14. Dueck, Hannah, Mugdha Khaladkar, Tae Kyung Kim, Jennifer M. Spaethling, Chantal Francis, Sangita Suresh, Stephen A. Fisher, et al. 2015. “Deep Sequencing Reveals Cell-Type-Specific Patterns of Single-Cell Transcriptome Variation.” *Genome Biology* 16 (June): 122.

15. Eldar, Avigdor, and Michael B. Elowitz. 2010. “Functional Roles for Noise in Genetic Circuits.” *Nature* 467 (7312): 167–73.

16. Elowitz, Michael B., Arnold J. Levine, Eric D. Siggia, and Peter S. Swain. 2002. “Stochastic Gene Expression in a Single Cell.” *Science* 297 (5584): 1183–86.

17. Furusawa, Chikara, and Kunihiko Kaneko. 2012. “A Dynamical-Systems View of Stem Cell Biology.” *Science* 338 (6104): 215–17.

18. Gillespie, Daniel T. 1976. “A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions.” *Journal of Computational Physics* 22 (4): 403–34.

19. Gong, Lanqi, Qian Yan, Yu Zhang, Xiaona Fang, Beilei Liu, and Xinyuan Guan. 2019. “Cancer Cell Reprogramming: A Promising Therapy Converting Malignancy to Benignity.” *Cancer Communications* 39 (1): 48.

20. Haber, Adam L., Moshe Biton, Noga Rogel, Rebecca H. Herbst, Karthik Shekhar, Christopher Smillie, Grace Burgin, et al. 2017. “A Single-Cell Survey of the Small

Intestinal Epithelium." *Nature* 551 (7680): 333–39.

21. Iyer-Biswas, Srividya, F. Hayot, and C. Jayaprakash. 2009. "Stochasticity of Gene Products from Transcriptional Pulsing." *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 79 (3 Pt 1): 031911.

22. Ji, Ni, Teije C. Middelkoop, Remco A. Mentink, Marco C. Betist, Satto Tonegawa, Dylan Mooijman, Hendrik C. Korswagen, and Alexander van Oudenaarden. 2013. "Feedback Control of Gene Expression Variability in the Caenorhabditis Elegans Wnt Pathway." *Cell* 155 (4): 869–80.

23. Ko, M. S. 1991. "A Stochastic Model for Gene Induction." *Journal of Theoretical Biology* 153 (2): 181–94.

24. Ko, M. S., H. Nakauchi, and N. Takahashi. 1990. "The Dose Dependence of Glucocorticoid-Inducible Gene Expression Results from Changes in the Number of Transcriptionally Active Templates." *The EMBO Journal* 9 (9): 2835–42.

25. Larsson, Anton J. M., Per Johnsson, Michael Hagemann-Jensen, Leonard Hartmanis, Omid R. Faridani, Björn Reinius, Åsa Segerstolpe, Chloe M. Rivera, Bing Ren, and Rickard Sandberg. 2019. "Genomic Encoding of Transcriptional Burst Kinetics." *Nature* 565 (7738): 251–54.

26. Losick, Richard, and Claude Desplan. 2008. "Stochasticity and Cell Fate." *Science* 320 (5872): 65–68.

27. Molina, Nacho, David M. Suter, Rosamaria Cannavo, Benjamin Zoller, Ivana Gotic, and Félix Naef. 2013. "Stimulus-Induced Modulation of Transcriptional Bursting in a Single Mammalian Gene." *Proceedings of the National Academy of Sciences of the United States of America* 110 (51): 20563–68.

28. Muramoto, Tetsuya, Danielle Cannon, Marek Gierlinski, Adam Corrigan, Geoffrey J. Barton, and Jonathan R. Chubb. 2012. "Live Imaging of Nascent RNA Dynamics Reveals Distinct Types of Transcriptional Pulse Regulation." *Proceedings of the National Academy of Sciences of the United States of America* 109 (19): 7350–55.

29. Phillips, Nicholas E., Aleksandra Mandic, Saeed Omidi, Felix Naef, and David M. Suter. 2019. "Memory and Relatedness of Transcriptional Activity in Mammalian Cell Lineages." *Nature Communications* 10 (1): 1208.

30. Raj, Arjun, Charles S. Peskin, Daniel Tranchina, Diana Y. Vargas, and Sanjay Tyagi. 2006. "Stochastic mRNA Synthesis in Mammalian Cells." *PLoS Biology* 4 (10): e309.

31. Schübeler, Dirk, David M. MacAlpine, David Scalzo, Christiane Wirbelauer, Charles Kooperberg, Fred van Leeuwen, Daniel E. Gottschling, et al. 2004. "The Histone Modification Pattern of Active Genes Revealed through Genome-Wide Chromatin Analysis of a Higher Eukaryote." *Genes & Development* 18 (11): 1263–71.

32. Selimkhanov, Jangir, Brooks Taylor, Jason Yao, Anna Pilko, John Albeck, Alexander Hoffmann, Lev Tsimring, and Roy Wollman. 2014. “Systems Biology. Accurate Information Transmission through Dynamic Biochemical Signaling Networks.” *Science* 346 (6215): 1370–73.

33. Shahrezaei, Vahid, Julien F. Ollivier, and Peter S. Swain. 2008. “Colored Extrinsic Fluctuations and Stochastic Gene Expression.” *Molecular Systems Biology* 4 (May): 196.

34. Sherman, Marc S., Kim Lorenz, M. Hunter Lanier, and Barak A. Cohen. 2015. “Cell-to-Cell Variability in the Propensity to Transcribe Explains Correlated Fluctuations in Gene Expression.” *Cell Systems* 1 (5): 315–25.

35. Sigal, Alex, Ron Milo, Ariel Cohen, Naama Geva-Zatorsky, Yael Klein, Yuvalal Liron, Nitzan Rosenfeld, Tamar Danon, Natalie Perzov, and Uri Alon. 2006. “Variability and Memory of Protein Levels in Human Cells.” *Nature* 444 (7119): 643–46.

36. Suter, David M., Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. 2011. “Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics.” *Science* 332 (6028): 472–74.

37. Suvà, Mario L., Nicolo Riggi, and Bradley E. Bernstein. 2013. “Epigenetic Reprogramming in Cancer.” *Science* 339 (6127): 1567–70.

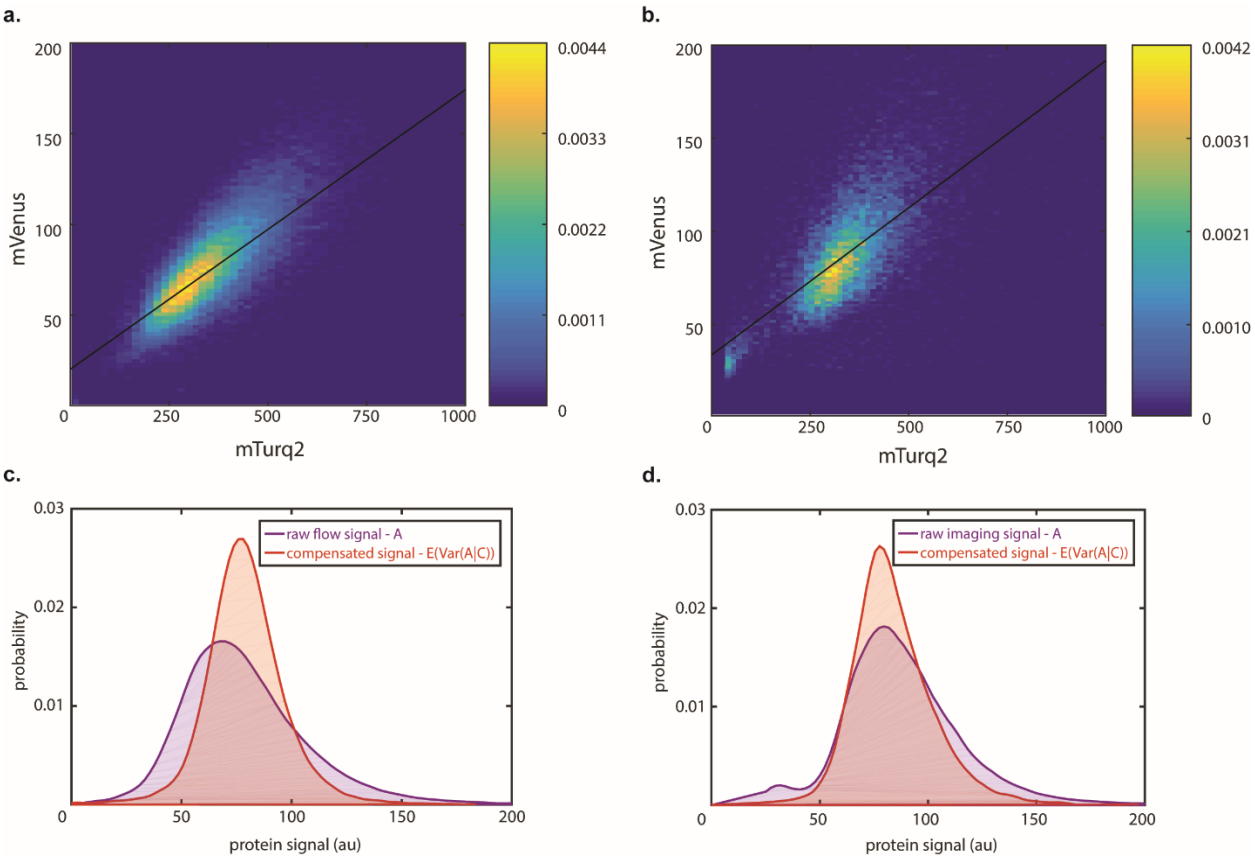
38. Symmons, Orsolya, and Arjun Raj. 2016. “What’s Luck Got to Do with It: Single Cells, Multiple Fates, and Biological Nondeterminism.” *Molecular Cell* 62 (5): 788–802.

39. Thomas, Philipp. 2017. “Making Sense of Snapshot Data: Ergodic Principle for Clonal Cell Populations.” *Journal of the Royal Society, Interface / the Royal Society* 14 (136). https://doi.org/10.1098/rsif.2017.0467.

40. Villani, Alexandra-Chloé, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, et al. 2017. “Single-Cell RNA-Seq Reveals New Types of Human Blood Dendritic Cells, Monocytes, and Progenitors.” *Science* 356 (6335). https://doi.org/10.1126/science.aah4573.

41. Yu, Ji, Jie Xiao, Xiaojia Ren, Kaiqin Lao, and X. Sunney Xie. 2006. “Probing Gene Expression in Live Cells, One Protein Molecule at a Time.” *Science* 311 (5767): 1600–1603.

Supplemental Figures



Supplementary Figure 1.1: Signal compensation by co-varying reporters for flow and imaging

(a) Heatmap of reporter system run on flow cytometry. The system used contains a singly-integrated mVenus reporter and multiply-integrated mTurquoise2 reporters. The black line represents line of best fit across population which is equivalent to the variance of the expectation value according to the double expectation theorem. **(b)** Heatmap of reporter system imaged by fluorescence microscopy. The system used contains a singly-integrated mVenus reporter and multiply-integrated mTurquoise2 reporters. The black line represents line of best fit across population which is equivalent to the variance of the expectation value according to the double expectation theorem. **(c)** Original distribution of mVenus signal measured by flow as compared to the compensated distribution through removal of co-varying noise. **(d)** Original distribution of mVenus signal measured by imaging as compared to the compensated distribution through removal of co-varying noise.

Supplemental Tables

Supplementary Table 1.1: Reactions and parameter values for *in silico* two-state model

	Reaction	Parameter value	Description
1.	$\mathrm{G}^{off} \rightarrow \mathrm{G}^{on}$	2.64 h^{-1}	Activation of Gene
2.	$\mathrm{G}^{on} \rightarrow \mathrm{G}^{off}$	0.212 h^{-1}	Deactivation of Gene
3.	$\mathrm{G}^{on} \rightarrow \mathrm{G}^{on} + mRNA$	0.907 h^{-1}	mRNA synthesis from active state of Gene
4.	$\mathrm{G}^{off} \rightarrow \mathrm{G}^{off} + mRNA$	0.0686 h^{-1}	mRNA synthesis from inactive state of Gene
5.	$\mathrm{mRNA} \rightarrow \emptyset$	0.1 h^{-1}	mRNA decay
6.	$\mathrm{mRNA} \rightarrow \mathrm{mRNA} + \mathrm{P}$	0.43 h^{-1}	Protein synthesis
7.	$\mathrm{P} \rightarrow \emptyset$	0.03 h^{-1}	Degradation of Protein

Supplementary Table 1.2: Reactions and parameter values for *in silico* slow fluctuation model

	Reaction	Parameter value	Description
1.	$\emptyset \rightarrow md$	0.2 h^{-1}	Upstream factor synthesis/accumulation
2.	$md \rightarrow \emptyset$	0.01 h^{-1}	Upstream factor decay/dissipation
3.	$\mathrm{G}^{off} \rightarrow \mathrm{G}^{on}$	2.64 h^{-1}	Activation of Gene
4.	$\mathrm{G}^{on} \rightarrow \mathrm{G}^{off}$	0.212 h^{-1}	Deactivation of Gene
5.	$\mathrm{G}^{on} \xrightarrow{md/\langle md \rangle} \mathrm{G}^{on} + mRNA$	1.56 h^{-1}	mRNA synthesis from active state of Gene driven by upstream factor
6.	$\mathrm{G}^{off} \rightarrow \mathrm{G}^{off} + mRNA$	0.0686 h^{-1}	mRNA synthesis from inactive state of Gene
7.	$\mathrm{mRNA} \rightarrow \emptyset$	0.1 h^{-1}	mRNA decay
8.	$\mathrm{mRNA} \rightarrow \mathrm{mRNA} + \mathrm{P}$	0.25 h^{-1}	Protein synthesis
9.	$\mathrm{P} \rightarrow \emptyset$	0.03 h^{-1}	Degradation of Protein

CHAPTER 2

Persistence of the intracellular calcium response to extracellular ATP

Abstract

In mammalian cells, extracellular ATP acts as a damage-associated molecular pattern (DAMP), creating a downstream release of intracellular calcium that triggers the innate immune response. This calcium spike encodes spatial information on a population level, with both the number of cells responding and average response level varying depending on the local ATP concentration. However, beyond this dose-dependent variability is further response variability, degrading total information capacity. The non-exclusive mechanisms underpinning this variability are intrinsic and pre-existing factors. To interrogate these mechanisms and identify the presence of distinct response archetypes within the population, we developed a ratiometric reporter system that allowed for the isolation of response variability. We then used this system to measure the persistence of response identity, intuiting a ceiling of intrinsic variability and determining the relative decay of pre-existing factors comprising calcium response identity up to 18 hours.

Introduction

When responding to stimuli, isogenic populations of mammalian cells exhibit significant cell-to-cell variability [5,10,22]. The underlying causes for this variability are unclear, but possible causes can be broken down into two groups, intrinsic fluctuations and variability in pre-existing factors [8,15]. Intrinsic fluctuations are defined as the stochastic chemical processes directly involved in the response, while pre-existing factors can include any deterministic factors,

such as expression variability of upstream network components, and global factors such as cell cycle and volume [9,16]. These factors are non-exclusive, and could all contribute to variability, allowing for accumulation of their fluctuations, but it has been estimated that the majority of variability is generated by pre-existing factors [9,23].

Given that we could view pre-existing factors as deterministic mappings producing defined outputs [21], is this the same as saying that heterogeneity in cellular responses amounts to distinct, functional cell states? Not necessarily, deterministic heterogeneity doesn't mean distinct clusters of behavior. Spectrums of phenotypic heterogeneity behavior occupying a Pareto front would be functional but not clustered [17,18]. To discern the difference, we could use clustering methods to decompose heterogeneous cell responses into more homogenous populations, seeking clear response archetypes [25]. But on its own, this approach wouldn't uncover the mechanisms underlying response variability, making the nature of these distinct cell states unclear. As different physiological mechanisms have different timescales of turnover, a potential approach is the measurement of the persistence of response identity as an indication of mechanism.

To study response variability, we focused on a single response network, the intracellular calcium response to ATP. ATP acts as a damage-associated molecular pattern (DAMP), which are endogenous signalling ligands released from the cell upon physical damage [6]. DAMPs act as the first indication of cellular damage, telling the surrounding healthy cells to activate transcriptional programs that trigger innate immune responses [4,24]. Depending on the nature of the propagation of DAMP signals, different kinds of information can be carried. Spatial information, that is, how far a cell is from the damage, could be determined if the DAMP signal acted as a gradient across distance [19]. If the DAMP is propagated through positive regulation,

then spatial information will likely be sacrificed in order to allow for long-distance signal transduction. This context has relevance to response variability as it informs how much variability could be allowed for while still transmitting functional information.

ATP displays release and diffusion behavior as a propagation strategy [12], allowing for a graded response to the signal, with high sensitivity. This is because extracellular ATP levels are quite low (~1 nM), approximately six orders of magnitude less than cytoplasmic levels [7], 2010). This means that any loss of membrane integrity leads to an immediate increase in extracellular ATP levels, allowing for this highly sensitive response to cellular damage. For a given cell, the response is mediated by the P2Y receptor, which is activated through phosphorylation by the binding of ATP. Phosphorylated P2Y then triggers phospholipase C (PLC) to degrade phosphatidylinositol 4,5-bisphosphate (PIP2), one of the products of which is inositol triphosphate (IP3) [2,3]. IP3 is then able to bind the IP3 receptor on the endoplasmic reticulum to release its store of calcium into the cytoplasm, creating an intracellular calcium spike. Calcium is then pumped back into the ER through the sarco/endoplasmic reticulum calcium ATPase (SERCA) channel, charging the system for another response [13]. This calcium release can be measured intracellularly through the use of a calcium indicator, GCaMP [14].

To interrogate the mechanisms of cellular heterogeneity, we measured the persistence of calcium response identity to extracellular ATP stimulation. We constructed a reporter system incorporating a ratiometric GCaMP-mCherry system into the MCF10A cell line. We performed a dual-stimulation experiment on this system, tracking cells for a variable period of time between simulations. For each time point and cell, the maximum GCaMP and average mCherry signal was taken, then compared to the other time point. These signals were then correlated across timepoints to understand the relationship between time and calcium response autocorrelation.

Results

For our reporter system, we used the calcium activity biosensor GCaMP5 [1] fused to the fluorescent reporter mCherry [20]. On its own, the GCaMP signal would have noise introduced from its own gene expression. By normalizing the GCaMP signal with mCherry signal, we can remove this noise, leaving a ratiometric calcium response signal that reflects response variability while not perturbing the underlying physiology its measuring. This fusion protein was multiply integrated into MCF10A, a mammary epithelial cell line using transposase technology. This system was imaged by fluorescence microscopy using MATLAB image processing to track cells and obtain single-cell traces. Cells were segmented with distinct fluorophore labelling and fluorescent channels. Hoechst staining was used to identify nuclei (Fig. 1a), and mCherry fluorescence was used to identify cytoplasm, as our GCaMP-mCherry fusion protein was excluded from the nucleus due to size (~78.8 kDa) (Fig. 1b).

Imaging was performed during dual-stimulation of our reporter system. The system was perturbed with 10 uM ATP, creating an intracellular calcium spike, then cells were tracked using their nuclear signal for a variable period of relaxation prior to a second perturbation (Fig. 2). To capture the calcium spike, the maximum GCaMP signal per stimulation was measured. This signal was normalized by the mean mCherry signal during each stimulation to generate a normalized, ratiometric, signal. Calcium spikes were compared and correlated, generating a mean correlation coefficient for a given period of relaxation. This allowed for us to measure the persistence of the calcium identity response through determining the decay in calcium response identity over time (Fig. 3). By measuring this persistence, we can infer which mechanisms are responsible for calcium response variability by comparing the timescale of these factors' fluctuations.

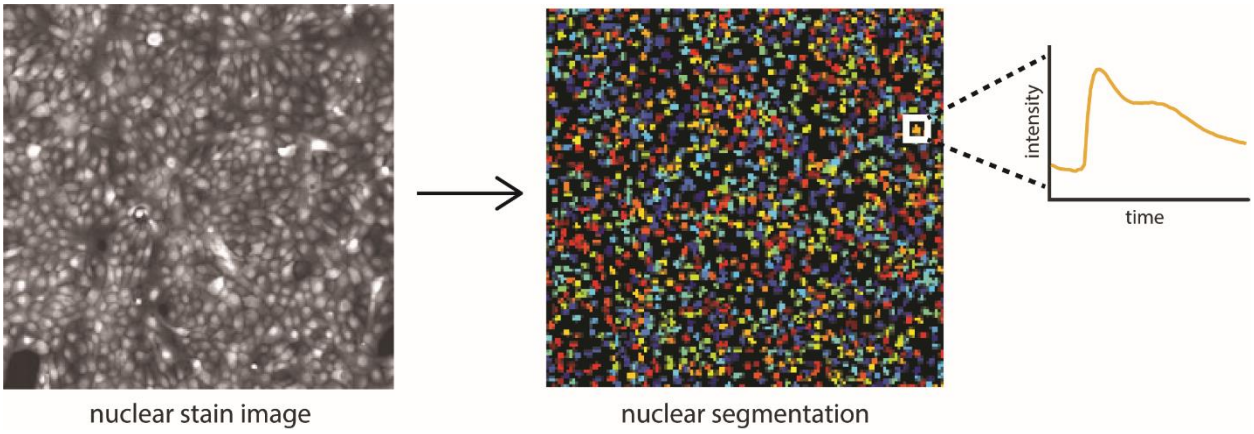


Figure 2.1: Overview of nuclear segmentation and single-cell

Processing of nuclear stain image (Hoechst) to segment and isolate nuclei, allowing for single-cell definition and cell tracking.

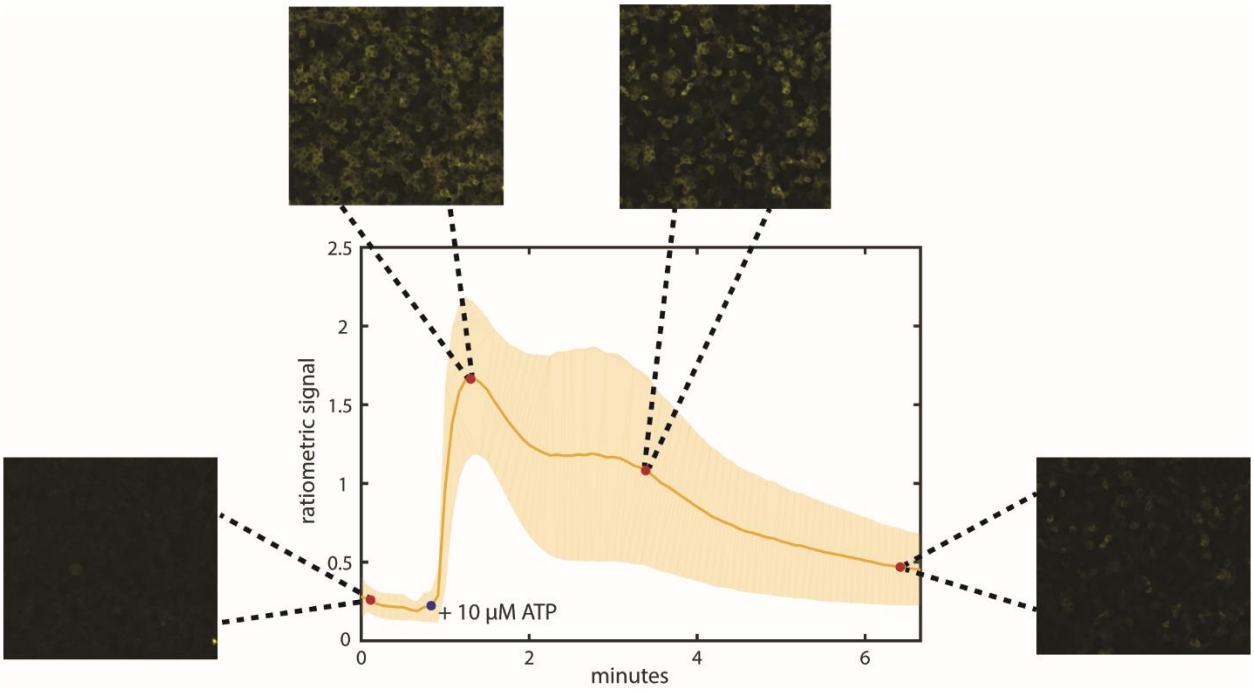


Figure 2.2: Features of the intracellular calcium response to extracellular ATP

Averaged intracellular calcium response (as measured by GCaMP-mCherry construct) to ATP. Major features of response labeled, including basal signal, peak calcium and decaying oscillations after perturbation.

Measuring the correlations of mCherry and maximum GCaMP signal separately (Fig. 3a), we can determine the relative persistence of reporter gene expression to response variability. GCaMP's signal is contingent on fluctuations in both response variability and reporter expression variability, but mCherry's signal is only contingent on the latter. The elevated correlation coefficients of mCherry over GCaMP (Fig. 3a) then imply that the response variability component of the GCaMP signal is fluctuating at a faster timescale than reporter expression variability. This expression variability is more reflective of global fluctuations than allele-specific effects, as our reporter was multiply integrated, diluting these effects.

The decay of the ratiometric calcium response reflects the removal of the shared reporter expression variability, leaving the faster fluctuating components behind. Consequently, the coefficients of the ratiometric signal are depressed relative to the original signals. However, the decay starts to level off around 0.5, suggesting a possible multi-component mechanism. The intrinsic variability of the calcium response would be removed merely by performing a second stimulation at any time point, as this would reflect stochastic chemical fluctuations that are independent across time points [16]. This means decay below the highest correlation coefficient measured (~0.82) is due to pre-existing factors. By comparing correlation decay with fluctuations of key pre-existing factors, such as the protein expression of upstream components like the P2Y receptor, we can determine the mechanisms responsible for generating variability in this system.

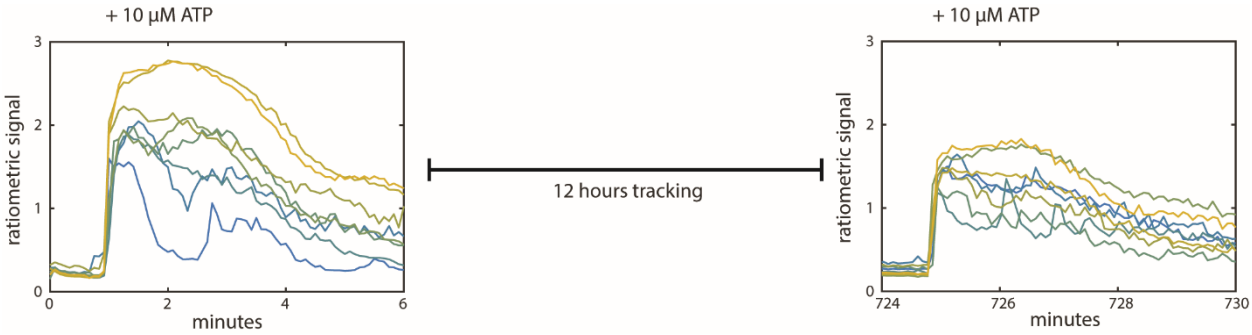


Figure 2.3: Overview of dual ATP stimulation experiment

Breakdown of aspects of dual stimulation experiment. Plots represent stimulations of 10 µM ATP and concurrent imaging of calcium signal for six minutes. Inter-stimulation period consists of intermittent imaging to capture Hoechst and mCherry signal.

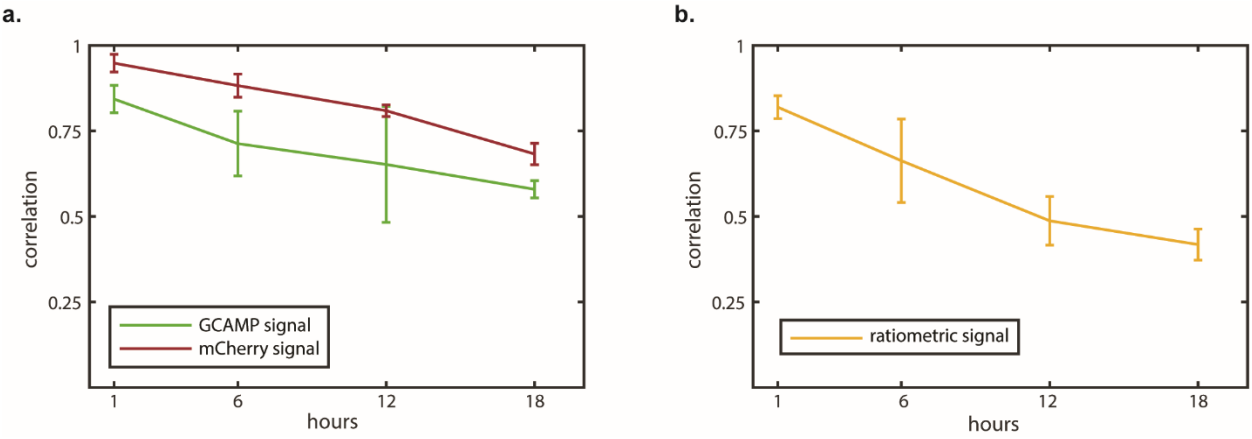


Figure 2.4: Correlation of peak calcium response over time

(a) Correlations of fluorescent signals given different time periods. Correlations broken down by GCaMP and mCherry signal prior to normalization. GCaMP correlation is calculated by comparing maximum values (peak feature), while mCherry correlation is calculated by average signal over stimulation period. mCherry signal is best indication of tracking error rate.
(b) Correlations of calcium peaks given different time periods. Ratiometric signal provided by normalizing GCaMP signal per cell by mCherry signal per cell. Peak calcium is then calculated and correlated to other stimulation, creating correlation versus time.

Discussion

In this work we developed a fusion reporter system that eliminated reporter expression noise while accurately measuring calcium response. We then utilized that system in an ATP dual stimulation experiment that allowed for the isolation of response variability in intracellular calcium release. Comparing signals within our reporter we confirmed that this response variability fluctuates on a faster timescale than the expression variability of our reporter. Utilizing our isolated signal, we demonstrated a definitive ceiling of intrinsic variability within the calcium response, and consequently intuited that the remainder of correlation decay is due to the fluctuations of key pre-existing factors.

To best capture functional calcium response heterogeneity, an understanding of the informational content of the calcium response is necessary. That is, not every feature of the calcium response may contain encoded functional information and therefore shouldn't be used as a feature to define heterogeneity. The calcium spike feature of the intracellular calcium response is a scalar quantity, which represents less information capacity than the full dynamic signal of the calcium response [16]. However, this feature represents the most information dense time interval in the full response [16], and can be explicative of gene expression variability in related networks [9]. Furthermore, the calcium peak encodes spatial information, allowing for an estimation of a cell's distance from cellular damage [12]. This provides a new context, as variability in calcium response levels would degrade this spatial information, acting as "true noise". A buffering mechanism for this noise may be paracrine signalling creating response fidelity through population averaging [11].

By defining a maximum value for intrinsic variability and observing a persistent calcium response identity, we have clarified that this process isn't memoryless. The network inducing

calcium release and recharge is deterministic, replicating a similar response upon a second stimulation. Any stochasticity in upstream factors becomes encoded as as an input for downstream factors. To determine which components of this network are responsible for the remaining identity loss, a co-fluctuation experiment could be carried out to link the turnover of a network component to the decay in calcium spike correlation. Suggested targets would be the P2Y or IP3 receptors, which both transduce a graded, and not saturated, response to 10 μM extracellular ATP.

The persistence of calcium response identity has implications as to whether intracellular calcium response heterogeneity contains distinct cellular states. The ergodic nature of this variability can be tested through determining if there is a “floor” of calcium response autocorrelation. Performing more dual-stimulation experiments at longer timescales could determine whether the correlation decay levels off or completely decays. A multi-day persistent correlation would imply distinct clusters within the cell state space for the intracellular calcium response, confirming recent reports [25].

Acknowledgements

Thank you to Anna Pilko for her help with cell line construction, Jason Yao for his mentoring at the inception of this project, and Roy Wollman for assistance with project conception and direction.

Lead contact and materials availability

Further information and requests for resources and reagents should be directed to, and will be fulfilled by the corresponding author Roy Wollman (rwollman@ucla.edu).

Materials and methods

Cell Culture

MCF10a cells were grown in complete media (above) and passaged at 70-90% confluency. Cells were seeded onto coated well of 96 well plate and grown to confluence before changing media to complete media without EGF and 1% horse serum, instead of normal 5%, 6-8 hours before imaging. Coating solution consists of sterile filtered 10ug/mL fibronectin, 10ug/mL bovine serum albumin, and 30ug/mL type I collagen in DMEM.

mCherry GCaMP5 Fusion MCF10A Cell Line Creation

To generate stable cell lines constitutively expressing cGamp5fusion-mcherry, MCF10A cells grown in the standard conditions and co-transfected using Neon transfection system (Invitrogen cat#MPK1025) and transposase expression vector pCMV-hyPBase (Sanger institute) in the 4:1 ratio with 0.625 ug of transposase and 2ug of transposon plasmid per well in 6 well dish. Electroporation parameters: Pulse voltage (v) 1,100 2003 Pulse width (ms) 20 Pulse number 2 Cell density (cells/ml) 2 x 10^5 Transfection efficiency 45% Viability 65% Tip type 10 µ Stable, polyclonal cell populations were established after blasticidin selection (10 µg/mL).

mCherry GCamp5 Fusion Construct Creation

For pPB - mCherry vector construction a PCR product encoding GCaMP5 sensor incorporating the CaMP3 mutation T302L R303P D380Y and no stop codon (Addgene plasmid #31788) was directionally ligated into pENTR/D-TOPO vector (Invitrogen K243520) resulting in a pEntry_GCaMP5G construct.
(For:caccATGGGTTCTCATCATCATCATCATCATGGTATGGCTAGCATGAC, REV: TTACTTCGCTGTCATCATTTGTACAAACTCTTCGTAG) pEntry_GCaMP5G was linearized with PCR reaction using standard Phusion® Hot Start Flex 2X Master Mix (NEB Cat# M0536L)

protocol (FOR: cgcgccgacccag , REV: ctcgagggatccggatcctcccttcgctgtcatcatttgtacaaac). PCR product was then subjected to DpnI digestion (NEB cat# R0176S) and gel purification with Zymoclean Gel DNA Recovery Kit (ZYMO cat#D4001). A sequence encoding mCherry and a5' linker was PCR amplified (FOR : gaggatccggatccctcgagAccatggtgagcaagggc REV :aagaaagctgggtcggcgcgcttgtacagctcgtccatg). mCherry2-C1 was a gift from Michael Davidson (Addgene plasmid # 54563). GeneArt Seamless Cloning and Assembly Enzyme Mix (Invitrogen cat# A14606) was used to assemble a construct encoding for GCaMP5 sensor fused with a short linker to mCherry called pENTRY-GCaMP5fusedmCherry. LR recombination between this entry clone and a custom gateway PiggyBack transposon vector with 1 µl LR Clonase II enzyme (Invitrogen: cat #11791020) resulted in the final construct of pPB_CAG_GCaMP5fusedmCherry_blast.

Longitudinal dual-stimulation experiment

The system used for this experiment was MCF10A with the exogenous calcium activity reporter GCaMP5 [1] fused to mCherry [20]. For analysis purposes, GCaMP5 signal was normalized by the average mCherry signal over the local time period to create a ratiometric reporter signal. Wells in a 96 well plate were seeded with this system prior to the experiment as detailed in ‘Cell Culture’. Prior to stimulation, cells were first incubated in 1:2,000 dilution of 10 mg/ml Hoechst. Dual-stimulation experiments were performed by stimulation with 10 µM ATP. Imaging began prior to the calcium response and was then imaged for six minutes. Tracking was then performed with minimal imaging for Hoechst and mCherry signal for the time period of interest prior to a second stimulation, which

Measurements of single-cell calcium response to an increase in extracellular ATP perturbation

Image acquisition of the 96-well plate was conducted on a Nikon Ti microscope using a 10× objective (Selimkhanov et al, 2014). The microscope was automated using micro-manager (www.micro-ma nager.org) through its MATLAB scripting interface. The images were segmented to locate the Hoechst signals as the positions of the nuclei. The calcium signals were mapped to the Hoechst images taken at different time points to establish single-cell trajectories. The procedure is identical to that described in Selimkhanov et al (2014).

References

1. Akerboom, Jasper, Tsai-Wen Chen, Trevor J. Wardill, Lin Tian, Jonathan S. Marvin, Sevinç Mutlu, Nicole Carreras Calderón, et al. 2012. "Optimization of a GCaMP Calcium Indicator for Neural Activity Imaging." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 32 (40): 13819–40.

2. Berridge, M. J., P. Lipp, and M. D. Bootman. 2000. "The Versatility and Universality of Calcium Signalling." *Nature Reviews. Molecular Cell Biology* 1 (1): 11–21.

3. Bootman, Martin D. 2012. "Calcium Signaling." *Cold Spring Harbor Perspectives in Biology* 4 (7): a011171.

4. Chen, Grace Y., and Gabriel Nuñez. 2010. "Sterile Inflammation: Sensing and Reacting to Damage." *Nature Reviews. Immunology* 10 (12): 826–37.

5. Cohen-Saidon, Cellina, Ariel A. Cohen, Alex Sigal, Yuvalal Liron, and Uri Alon. 2009. "Dynamics and Variability of ERK2 Response to EGF in Individual Living Cells." *Molecular Cell* 36 (5): 885–93.

6. Cordeiro, João V., and António Jacinto. 2013. "The Role of Transcription-Independent Damage Signals in the Initiation of Epithelial Wound Healing." *Nature Reviews. Molecular Cell Biology* 14 (4): 249–62.

7. Corriden, Ross, and Paul A. Insel. 2010. "Basal Release of ATP: An Autocrine-Paracrine Mechanism for Cell Regulation." *Science Signaling* 3 (104): re1.

8. Elowitz, Michael B., Arnold J. Levine, Eric D. Siggia, and Peter S. Swain. 2002. "Stochastic Gene Expression in a Single Cell." *Science* 297 (5584): 1183–86.

9. Foreman, Robert, and Roy Wollman. 2019. “Mammalian Gene Expression Variability Is Explained by Underlying Cell State.” *bioRxiv*. https://doi.org/10.1101/626424.

10. Geva-Zatorsky, Naama, Nitzan Rosenfeld, Shalev Itzkovitz, Ron Milo, Alex Sigal, Erez Dekel, Talia Yarnitzky, et al. 2006. “Oscillations and Variability in the p53 System.” *Molecular Systems Biology* 2 (June): 2006.0033.

11. Handly, L. Naomi, Anna Pilko, and Roy Wollman. 2015. “Paracrine Communication Maximizes Cellular Response Fidelity in Wound Signaling.” *eLife* 4 (October): e09652.

12. Handly, L. Naomi, and Roy Wollman. 2017. “Wound-Induced Ca2+ Wave Propagates through a Simple Release and Diffusion Mechanism.” *Molecular Biology of the Cell* 28 (11): 1457–66.

13. Lemon, G., W. G. Gibson, and M. R. Bennett. 2003. “Metabotropic Receptor Activation, Desensitization and Sequestration-I: Modelling Calcium and Inositol 1,4,5-Trisphosphate Dynamics Following Receptor Activation.” *Journal of Theoretical Biology* 223 (1): 93–111.

14. Nakai, J., M. Ohkura, and K. Imoto. 2001. “A High Signal-to-Noise Ca(2+) Probe Composed of a Single Green Fluorescent Protein.” *Nature Biotechnology* 19 (2): 137–41.

15. Rhee, David Y., Dong-Yeon Cho, Bo Zhai, Matthew Slattery, Lijia Ma, Julian Mintseris, Christina Y. Wong, et al. 2014. “Transcription Factor Networks in Drosophila Melanogaster.” *Cell Reports* 8 (6): 2031–43.

16. Selimkhanov, Jangir, Brooks Taylor, Jason Yao, Anna Pilko, John Albeck, Alexander Hoffmann, Lev Tsimring, and Roy Wollman. 2014. “Systems Biology. Accurate Information Transmission through Dynamic Biochemical Signaling Networks.” *Science* 346 (6215): 1370–73.

17. Sheftel, Hila, Oren Shoval, Avi Mayo, and Uri Alon. 2013. “The Geometry of the Pareto Front in Biological Phenotype Space.” *Ecology and Evolution* 3 (6): 1471–83.

18. Shoval, O., H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh, and U. Alon. 2012. “Evolutionary Trade-Offs, Pareto Optimality, and the Geometry of Phenotype Space.” *Science* 336 (6085): 1157–60.

19. Sonnemann, Kevin J., and William M. Bement. 2011. “Wound Repair: Toward Understanding and Integration of Single-Cell and Multicellular Wound Responses.” *Annual Review of Cell and Developmental Biology* 27 (June): 237–63.

20. Su, Steven, Siew Cheng Phua, Robert DeRose, Shuhei Chiba, Keishi Narita, Peter N. Kalugin, Toshiaki Katada, Kenji Kontani, Sen Takeda, and Takanari Inoue. 2013. “Genetically Encoded Calcium Indicator Illuminates Calcium Dynamics in Primary Cilia.” *Nature Methods* 10 (11): 1105–7.

21. Symmons, Orsolya, and Arjun Raj. 2016. “What’s Luck Got to Do with It: Single Cells, Multiple Fates, and Biological Nondeterminism.” *Molecular Cell* 62 (5): 788–802.

22. Tay, Savaş, Jacob J. Hughey, Timothy K. Lee, Tomasz Lipniacki, Stephen R. Quake, and Markus W. Covert. 2010. “Single-Cell NF-kappaB Dynamics Reveal Digital Activation and Analogue Information Processing.” *Nature* 466 (7303): 267–71.

23. Toettcher, Jared E., Orion D. Weiner, and Wendell A. Lim. 2013. “Using Optogenetics to Interrogate the Dynamic Control of Signal Transmission by the Ras/Erk Module.” *Cell* 155 (6): 1422–34.

24. Vénéreau, Emilie, Chiara Ceriotti, and Marco Emilio Bianchi. 2015. “DAMPs from Cell Death to New Life.” *Frontiers in Immunology* 6 (August): 422.

25. Yao, Jason, Anna Pilko, and Roy Wollman. 2016. “Distinct Cellular States Determine Calcium Signaling Response.” *Molecular Systems Biology* 12 (12): 894.

Conclusions and future directions

In this work we developed reporter systems designed to capture cellular heterogeneity in two different contexts, an exogenous, idealized, reporter system purely measuring gene expression, and a ratiometric system capturing an endogenous response to stimuli. These systems were both designed to isolate signal components and measure their persistence and behavior across time, with the aim of answering fundamental questions about cellular identity within isogenic populations. Through this work we have found a higher degree of cellular identity and stratification than imagined in previous paradigms, which focus on the ability of stochastic behavior to create variability in cellular processes.

Our exogenous system was designed to isolate allele-specific variability of a single reporter through the removal of shared noise. The behavior and fluctuations of this signal were tracked through imaging and compared against an *in silico* version of the two-state model, and found to have a more persistent identity. In order to account for this persistence, we developed a model incorporating a time-dependent upstream component that influenced the transcription rate. To test this model further we developed a fluctuation analysis technique that tracked our system's isolated allele-specific variability across ten days. Comparing this relaxation of distinct cell states against a control recapitulating the timescale of RNA bursting, we see a long term cis-identity. This technique also allowed for the isolation of batches of distinct cell populations, which could be subsequently analysed using immunoprecipitation methods. Our batches were analyzed using ChIP-qPCR, measuring the enrichment of histone modifications associated with transcriptional activity near the transcription start site of our reporter. High expression was linked to enrichment of H3K4me3, demonstrating co-fluctuations of chromatin states and gene expression within populations of isogenic cells.

These results confirm our cis-fluctuating model, providing direct evidence of chromatin regulation influencing gene expression. Future studies, ideally using single-cell ChIP methods, will map out more of the epigenetic landscape, implicating other histone modifications in cellular heterogeneity. One question along this lineage of thought are the degree to which cells enriched by these modifications represent a permanent identity, or whether certain stresses can dislodge intractable subpopulations; a question that is relevant to both our understanding of cell fate as well as approaches towards cancer cell reprogramming. Another question is the influence of chromatin regulation across the genome, balancing reliability and diversity to generate optimal functional heterogeneity, and the consequences of this regulation going awry in ageing cell populations.

Our ratiometric reporter system directly measured the intracellular calcium response to extracellular ATP by removing noise introduced by the gene expression of our reporter. We performed dual-stimulation experiments on this system, tracking the identity of individual cell responses between these two stimulations. By correlating these identities across a population, we gain an understanding of the mechanisms driving this response. Through looking at this correlation over time, we can immediately identify a ceiling for intrinsic fluctuations, and can surmise that the remainder of the correlation is due to deterministic, pre-existing factors. This correlation further decays but persists after a day, showing a semi-stable phenotypic identity.

These results serve as an excellent starting point for interrogating the mechanisms of the intracellular calcium response. A possible approach could measure related network components, determining whether co-fluctuations exist, thus cementing a mechanistic link. Further time point measurements could confirm the ergodicity of this response, allowing us to discern whether distinct calcium response archetypes exist within isogenic populations of

MCF10A. The presence of these archetypes would raise questions as to how cells decode heterogeneous responses, including decoding the spatial information transmitted by the calcium response to ATP.

Early indications of these findings were single-cell transcriptomic studies demonstrating that populations of cells occupy a large and varied phenotypic space [1,2,3,4]. The clash of these results with theoretical frameworks treating cell processes as simple stochastic mechanisms seemed obvious but without hard proof of deterministic states producing cellular heterogeneity, this phenomenology was insufficient. However, the demonstration of histone modification diversity creating phenotypic diversity challenges the two-state model of gene expression and invites new theory to catch up to the wealth of single-cell data being gathered.

References

1. Buganim, Yosef, Dina A. Faddah, Albert W. Cheng, Elena Itskovich, Styliani Markoulaki, Kibibi Ganz, Sandy L. Klemm, Alexander van Oudenaarden, and Rudolf Jaenisch. 2012. “Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase.” *Cell* 150 (6): 1209–22.

2. Kumar, Roshan M., Patrick Cahan, Alex K. Shalek, Rahul Satija, Ajay DaleyKeyser, Hu Li, Jin Zhang, et al. 2014. “Deconstructing Transcriptional Heterogeneity in Pluripotent Stem Cells.” *Nature* 516 (7529): 56–61.

3. Moignard, Victoria, Iain C. Macaulay, Gemma Swiers, Florian Buettner, Judith Schütte, Fernando J. Calero-Nieto, Sarah Kinston, et al. 2013. “Characterization of Transcriptional Networks in Blood Stem and Progenitor Cells Using High-Throughput Single-Cell Gene Expression Analysis.” *Nature Cell Biology* 15 (4): 363–72.

4. Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. 2014. “The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells.” *Nature Biotechnology* 32 (4): 381–86.