UNIVERSITY OF CALIFORNIA SAN DIEGO

Cell States Explain Calcium Signaling Heterogeneity in MCF10a Cells in Response to ATP Stimulation

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Robert Kevin Foreman

Committee in charge:

Professor Scott Rifkin, Chair
Professor Roy Wollman, Co-Chair
Professor Nathan Lewis
Professor Jin Zhang
Professor Kun Zhang

2019

The Dissertation of Robert Kevin Foreman is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

____

____

Co-Chair

____

Chair

University of California San Diego

2019

DEDICATION

I dedicate this dissertation to my family and girlfriend.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Chapter 1 is a full reprint of the material as it appears in Mammalian gene expression is explained by cell state in bioRxiv 2019. Foreman, Robert and Wollman, Roy. The dissertation author was the primary investigator and author of this material. Anna Pilko is not an author, but constructed the GCaMP5 MCF10A cell line for a previous work.

Chapter 2 is being prepared for submission for the publication of the material. The authors are Robert Foreman, Evan Maltz, and Roy Wollman. The dissertation author was the primary investigator and author of this material. Evan Maltz performed the variational autoencoder models.

Jeffrey Moffitt provided many helpful emails while getting the MERFISH method to work.

VITA

2012	Bachelor of Science, University of North Carolina at Charlotte
2019	Doctor of Philosophy, University of California San Diego

FIELDS OF STUDY

Major Field: Bioinformatics and Systems Biology

Studies of Cell Signaling Heterogeneity
Professor Roy Wollman

ABSTRACT OF THE DISSERTATION

Cell States Explain Calcium Signaling Heterogeneity in MCF10a Cells in Response to ATP Stimulation

by

Robert Kevin Foreman

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2019

Professor Scott Rifkin, Chair

Professor Roy Wollman, Co-Chair

Regulated differences between cells fundamentally endows multicellular organisms with the huge diversity of complex functions and behaviors we observe in nature. Recent advances in single-cell technology such as droplet-based RNA-Seq are revealing at an unprecedented rate different cellular states and their dynamics. However, not all differences at the gene expression level are necessarily related to cell states. Differences between cells can arise from noisy effects such as transcriptional bursting. Similarly, not all phenotypic variability necessarily arises from systematic differences in gene expression. For example, variability can be explained

by post transcriptional regulation and/or intrinsic fluctuations in components of the calcium signaling network. In order to clarify the sources of variability in calcium signaling, we sought to use in situ sequential hybridization smFISH in order to obtain highly accurate single-cell expression counts paired to measurement of a complex phenotype, calcium signaling dynamics, which is an emergent property of gene expression plus post-transcriptional regulation. In this work, we identify an upper bound for how much gene expression variability could arise from allele specific transcriptional bursting, and then investigate how much of calcium signaling variability is explained by gene expression differences related to different cell states.

INTRODUCTION - Variability in Gene Expression and Calcium Signaling Response

Abstract

Cell states are regulated differences between individual cells and occur on many different scales. At one extreme there are cell types such as muscle cells vs neuronal cells with cell state differences so significant they manifest morphologically. At the other end of the spectrum there are immune cells that can appear quite similar, but may differ in their expression of a small number of genes affecting the cells behavior. It is currently unclear what the lower limit to the resolution of cellular identity is, are cells stratified up to this limit, and whether mammalian systems tend to distribute as clusters or continuum in expression space. These studies will address how intrinsic noise acts as a resolution limit to cell states, and show how much of calcium signaling heterogeneity is actually explained by systematic gene expression differences between cells.

Introduction

Despite significant research seeking to understand the genotype to phenotype relationship. It is still an open question how much of the variability in certain complex phenotypes, such as cellular signaling, is explained by systematic differences in gene expression between cells. Some previous studies found that non genetic differences explain a significant proportion of phenotypic variability(Cheong et al., 2011; Spencer et al., 2009). While other studies indicated that cellular heterogeneity seemed to originate in long-time scale extrinsic cell-to-cell differences(Selimkhanov et al., 2014; Toettcher et al., 2013; Yao et al., 2016). If cells are fundamentally very noisy, and signaling is corrupted by this noise then cells must be structured to function through robust low entropy self-organized processes. On the other hand, if cells are actually very accurate at responding to complex environmental cues then multicellular organisms could be functioning in a fundamentally different manner. To bridge this knowledge gap, we sought to measure both the complex emergent phenotype of signaling dynamics in response to ligand stimulation, and the gene expression state of a large number of single-cells.

In order to address subtle differences in cell state we needed to use a highly sensitive and accurate gene expression measurement. While there has been an explosion in scRNA-Seq technology over the last 5 years with the introduction of fluidic capture methods such as C1 Fluidigm and droplet based approaches(Gong et al., 2018; Macosko et al., 2015; Xin et al., 2016; Zilionis et al., 2017), these techniques can be limited in three key ways: low RNA sensitivity, low cell capture efficiency, and loss of spatial information about cells(Zhang et al., 2019). Low capture efficiency and RNA sensitivity affect the ability to detect rare cell populations and cell states maintained by subtle gene expression differences(Andrews and Hemberg, 2018). We want to preserve spatial information about cells in order to map the gene expression state of a cell to its corresponding calcium signaling phenotype. Furthermore, cellular state could be maintained by local environments that would be perturbed by tissue dissociation required for droplet based scRNA-Seq(Haque et al., 2017). Fortunately, over the last 5 years spatial transcriptomics methods have also developed very rapidly. These sequential hybridization techniques are very sensitive and the *in situ* nature of the technique allows the alignment of cells between live-cell calcium dynamical measurement and the single-cell gene expression measurement(Foreman and Wollman, n.d.).

While there are subtle differences in the implementation of spatial transcriptomics every method relies on performing a series of sequential hybridization, imaging, and quenching. Some methods simply used standard single-gene per hybridization/color, and the multiplexing scaled linearly in the number of dye colors times the number of hybridizations performed(Codeluppi et al., 2018). Others used spectral barcoding(Cai, 2013), simple combinatorial barcoding(Lubeck and Cai, 2012), and recently groups have established error robust barcoding methods which scale to 100s or 1000s of genes(Eng et al., 2019; Moffitt et al., 2016). Error robust barcoding allows for very sensitive, 99% true positive, and specific, 0.1% false positive rate, quantification of transcripts per cell(Moffitt et al., 2016). These sequential hybridization smFISH techniques are mature and have been used in cell culture and tissues of many different kinds. This era of spatial transcriptomics is in the early stages, and there are exciting applications in a variety of different areas from mapping cell states in whole tissues to understanding heterogeneity and spatial structure in tissues and cancers.

Beyond the *in situ* benefit of preserving spatial information of cells and spatial distribution of RNA within cells, one can also use the spatial information to map data from any image based assay onto the gene expression state of the cell. In this work we leverage this ability to make joint measurement of a live cell phenotype measured by a genetically encoded biosensor of cytoplasmic calcium with the gene expression measurement of many genes from the calcium signaling network. Thereby we are able to address questions related to the genetic basis of heterogeneity in calcium signaling.

The calcium signaling pathway is known to exhibit a range of qualitatively diverse dynamical patterns including: single peak(Yao et al., 2016), oscillations(Smedler and Uhlén, 2014), excitable pulses(Zhang et al., 2015), and various combinations(Giorgi et al., 2018; Taylor and Francis, 2014). These different response types were shown to lead to differential outcomes for cells that responded one way or another. For example, in endothelial cells responding to VEGF, cells that oscillate tend to proliferate while cells with a high intensity single peak tend to migrate(Noren et al., 2016). Single-cells could therefore be encoding a stratified set of responses to the same environmental cues, and effectively utilizing cell state differences to achieve a broader repertoire of response than would be possible if all cells responded identically. This thesis seeks to elucidate how much of the variability in signaling is due to cell state differences, and to evaluate how much cell state maintenance is limited by transcriptional bursting.

Chapter 1 seeks to identify a minimal unit of cell state. This limit is effectively set by the scale of intrinsic noise, allele specific transcriptional bursting, that induces short timescale variability in the gene expression state of a cell. Therefore, gene expression differences that are smaller than this intrinsic noise limit could not be effectively maintained over time. We find that in contrast to previous reports that intrinsic noise is a super-Poissonian limiting factor(Dar et al., 2015; Hansen et al., 2018a), after controlling for extrinsic differences between individual cells, gene expression is actually distributed near Poisson. This means that in terms of gene expression variability transcriptional bursting in an allele-specific manner is playing a smaller role than extrinsic variability from systematic differences between individual cells.

Chapter 2 leverages that there is a correlation between the gene expression state of the cell, and the differences observed in calcium signaling response to ATP. We use the variational autoencoders(Doersch, 2016) to learn the information capacity of calcium response encodings, and how much of the encoding differences between cells is explained by gene expression. We also identify cell states related to differences in calcium signaling dynamics, and validate that the cell states are meaningful by relating to the emergent phenotype.

CHAPTER 1 - Transcriptional Bursting does not Limit Cellular State Resolution

Abstract

Gene expression variability in mammalian systems plays an important role in physiological and pathophysiological conditions. This variability can come from differential regulation related to cell state (extrinsic) and allele-specific transcriptional bursting (intrinsic). Yet, the relative contribution of these two distinct sources is unknown. Here we exploit the qualitative difference in the patterns of covariance between these two sources to quantify their relative contributions to expression variance in mammalian cells. Using multiplexed error robust RNA fluorescent in situ hybridization (MERFISH) we measured the multivariate gene expression distribution of 150 genes related to Ca^{2+} signaling coupled with the dynamic Ca^{2+} response of live cells to ATP. We show that after controlling for cellular phenotypic states such as size, cell cycle stage, and Ca^{2+} response to ATP, the remaining variability is effectively at the Poisson limit for most genes. These findings demonstrate that the majority of expression variability results from cell state differences and that the contribution of transcriptional bursting is relatively minimal.

Introduction

Gene expression variability is ubiquitous in all biological systems. In multicellular organisms heterogeneity between different cell types and states confers specialized function giving rise to complexity in whole-system behavior(Eldar and Elowitz, 2010; Raj and van Oudenaarden, 2008; Suo et al., 2018; Symmons and Raj, 2016; Tabula Muris Consortium et al., 2018). Similarly, single-cell organisms and viruses were shown to utilize heterogeneity at the population level to create diverse phenotypes, such as bet-hedging strategies in changing environments(Rouzine et al., 2015; Veening et al., 2008; Vega and Gore, 2014). While variability can provide useful functional heterogeneity in a multicellular organism or cell population, it is not necessarily always beneficial(Raj and van Oudenaarden, 2008; Symmons and Raj, 2016). Unregulated stochastic events, i.e. noise, can limit cells ability to respond accurately to changing environments and can introduce phenotypic variability that can have a negative contribution to overall fitness. Indeed, many biological mechanisms including buffering(Stoeger et al., 2016) and feedback loops(Jangi and Sharp, 2014; Schmiedel et al., 2015) have been suggested to limit the detrimental effect of gene expression variability. Quantification of the different contributions of mechanisms that cause gene expression variability is an important step toward determining to what degree the variability represents uncontrolled “noise” or cellular stratification and function.

Two key contributors of gene expression variability are allele specific sources and global factors related to underlying cell state. Analysis of expression covariance between genes is a powerful approach to decompose gene expression variability into these two classes. Landmark works used this approach to investigate expression variability in bacterial cells, which laid a foundation for decomposing variability into allele-specific (intrinsic) sources and variability that originate from sources that affect multiple alleles and relate to the underlying cell state (extrinsic)(Elowitz, 2002; Paulsson, 2005). This work was later extended to yeast(Raser and O’Shea, 2004) and mammalian systems(Raj et al., 2006; Sigal et al., 2006; Singh et al., 2012). The decomposition into allele-specific and cell state components is not always simple. Allele-specific noise in an upstream component can propagate into downstream genes(Sigal et al., 2006) whereas temporal fluctuations in the shared components can have nontrivial consequences on expression distributions(Paulsson, 2004; Pedraza and van Oudenaarden, 2005; Shahrezaei et al., 2008). Finally, use of the terms “intrinsic” and “extrinsic” is sometimes ill-defined and some models include a “coupled intrinsic” mode as well which is a form of shared variability and hence “extrinsic”(Rodriguez et al., 2019). Despite the sometimes confusing nomenclature, the use of expression covariance to distinguish between allele-specific and shared factors is a powerful decomposition approach.

In addition to covariance based approaches, the relationship between gene expression distribution variance and mean provides a useful quantitative framework to gain insights into sources of expression variability(Munsky et al., 2012). The comparison of expression variability between genes is not straightforward as expression variance scales with its mean. Three statistical tools are commonly used to describe mean normalized variance: the coefficient of variation (CV), coefficient of variation squared (CV^2), and Fano factor. CV and CV^2 are both unitless measures where the CV is defined as the standard deviation divided by the mean and the CV^2 is simply the CV squared, or the variance divided by the mean squared. The CV and CV^2 are useful to compare the scale of variance between different genes because of their unitless nature. The third measure, the Fano factor, is the variance divided by the mean and therefore not unitless, but it has a special property of being equal to one in the case of a Poisson process. Many biological processes have a variance to mean ratio that is at least Poisson so the Fano factor can define a 'standard dispersion', as a result, distributions with Fano factor smaller/bigger than one are considered under/over-dispersed, respectively. Therefore a simple quantification of the distribution variance scaled by its mean can provide key insights into the underlying mechanism generating the observed distribution(Choubey et al., 2015; Hansen et al., 2018a).

Multiple studies across bacteria, yeast, and mammalian cells measured over-dispersed gene expression distributions. This observation can have two main interpretations. One interpretation is that the observed over-dispersion is simply a result of the superposition of an allele-specific Poisson variability and cell state variability(Battich et al., 2015). The other interpretation is that the allele-specific variability itself is not a simple Poisson process(Corrigan et al., 2016; Dar et al., 2015; Suter et al., 2011; Tantale et al., 2016). The latter interpretation was popularized by the introduction of a simple phenomenological model named the two-state or random telegraph model that represented genes as existing in either “on” or “off” states(Friedman et al., 2006; Fukaya et al., 2016; Kaern et al., 2005; Kepler and Elston, 2001; Lenstra et al., 2016; Molina et al., 2013; Paulsson, 2004; Peccoud and Ycart, 1995; Raj et al., 2006; Sanchez and Golding, 2013; Shahrezaei and Swain, 2008; Suter et al., 2011; Thattai and van Oudenaarden, 2004). More complex models with multiple states were also considered,(Corrigan et al., 2016; Nicolas et al., 2018; Suter et al., 2011; Tantale et al., 2016; Zoller et al., 2015) but the addition of multiple states does not change the model in a qualitative way. These models suggest that transcription should occur in distinct bursts with multiple transcripts generated when the gene is “on”. These two-state models can be described by two overall key parameters: the burst size and frequency that control the resulting gene expression distributions with lower burst frequency and larger burst size contributing to the overdispersion of the underlying distribution. Overall both interpretations, bursting and cell state, can explain the observed over-dispersion and it is currently unclear which one is correct.

The relative scales and sources of variability are very important to understand in the modern world of single-cell highly multiplexed measurements. These new technologies are revealing the complex structure of ‘cell space’ with cells occupying a large array of types(Han et al., 2018; Rosenberg et al., 2018; Tabula Muris Consortium et al., 2018), states(Cheng et al., 2019; Trapnell, 2015), and fronts(Shoval et al., 2012) that reflect functional stratification. Despite our knowledge that cell types and states manifest as gene expression heterogeneity, sometimes total gene expression variability is interpreted as arising from two-state transcriptional bursting alone(Larsson et al., 2019). The gap in our understanding of the relative contribution of cell state and allele-specific factors is hindering progress in assigning functional roles to observed variability(Dueck et al., 2016).

To address this knowledge gap, we utilized the two key properties of expression variability: covariance and dispersion. We measured gene covariance and dispersion using joint measurements of individual cells; where for each cell multiple cell state features were measured as well as a highly multiplexed measurement of gene expression. We used sequential hybridization smFISH (MERFISH implementation)(Moffitt et al., 2016) that allowed us to accurately measure the expression of 150 genes in ~5000 single-cells. Since expression covariance between genes from the same pathway is higher compared to genes that have distinct functions(Sigal et al., 2006; Stewart-Ornstein et al., 2012), we focused on a single signaling network and biological function, Ca^{2+} response to ATP in epithelial cells, a biological response important to wound healing(Funaki et al., 2011; Handly et al., 2015; Handly and Wollman, 2017).The key advantage of Ca^{2+} response is that the overall signaling response can be measured in less than fifteen minutes, a fast timescale that precludes any ATP induced changes in transcription. Using the combined dataset we were able to separate the correlated and uncorrelated components using a simple multiple linear regression model guided by the changes in the covariance matrix. We found that after removing all shared components, the remaining allele-specific variability shows very little over-dispersion for most genes measured. Overall these results indicate that transcriptional bursting is only a minor contributor to the overall observed expression variability.

Results

To assess the relative contribution of the overall expression variability that stems from allele-specific sources versus underlying cell state variability, we took advantage of the fact that these two sources have different expression covariance signatures. Figure 1.1 shows simulated data to illustrate how covariance signatures can be utilized to decompose sources of variability. By definition, allele-specific variability is uncorrelated to any other gene whereas variability that is due to heterogeneity in the underlying cell state will likely be shared between genes with similar function (Figure 1.1A). When transcriptional bursting dominates (Figure 1.1B top) the shared regulatory factors will have a small contribution, there will be little correlation between genes and the expression variance will remain largely unchanged after conditioning expression level on any cell state factors (Figure 1.1B top right). The residual intrinsic variance will have a Fano factor greater than one. On the other hand, when cell state variability dominates (Figure 1.1B bottom), expression between genes will be highly correlated and conditioning the expression on cell state factors will reduce both the variance and correlation between genes. At the limit, when all shared factors are accounted for, the correlation between genes will approach zero and the Fano factor of the residuals will approach one, the Poisson limit (Figure 1.1B bottom right). When the contribution of bursting and cell state is comparable (Figure 1.1B middle) conditioning on cell state factors will have some effect but the final Fano factor will be higher than one even when the correlation is zero (Figure 1.1B middle right). Conditioning on cell state factors has a dual effect on correlation and Fano factor and therefore it is possible to assess whether the conditioning removed all the obvious extrinsic variability. When all the extrinsic variability is conditioned out, one can confidently interpret whether the residual intrinsic variability is under or overdispersed.

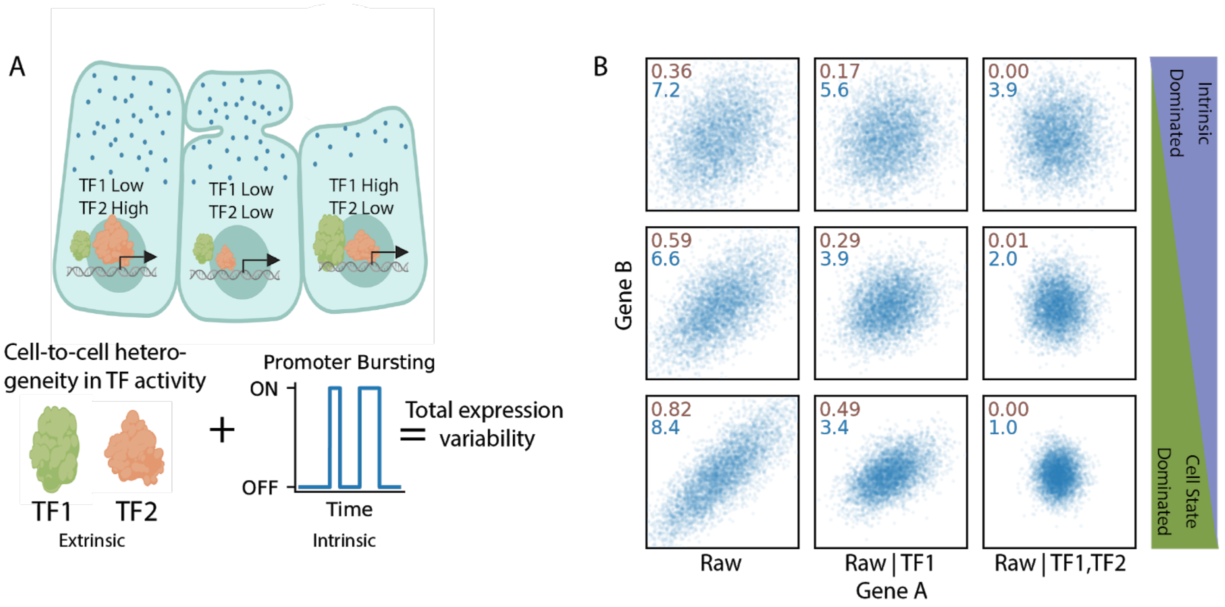
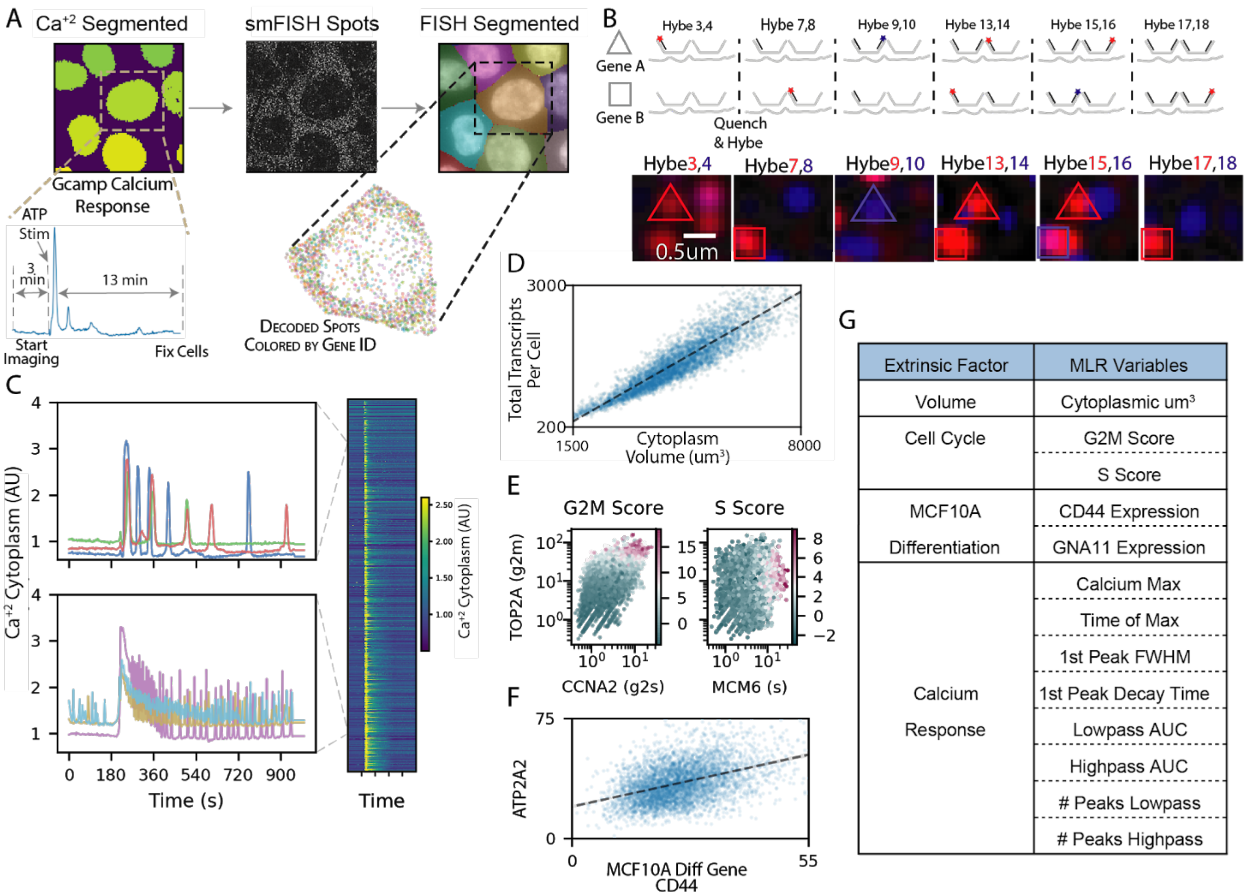


Figure 1.1 Transcriptional bursting and trans-acting factors are two distinct causes of cell-to-cell heterogeneity. (a) Cartoon depicting that different cells can have different activities of trans-factor (TF) regulatory molecules in addition to the effects of transcriptional bursting. (b) Simulated data showing that variability from shared regulatory factors results in correlation between two genes with three example cases: intrinsic dominated noise (top three panels), mixture of cell-state and allele specific sources (middle three), and cell-state dominated (bottom three). This correlation is diminished when the expression levels are conditioned on the levels of these shared regulatory factors (middle and right). After conditioning on all trans-acting regulatory factors the remaining variability due to transcriptional bursting alone is potentially significantly smaller (right). Inset text is the pearson correlation coefficient between Gene A and Gene B (brown) and the fano factor of

To distinguish between the possible situations described above requires accurate highly multiplexed single-cell measurements of gene expression and a sufficient number of cellular features that correlate with the underlying cell state factors controlling gene expression. To achieve this we developed an experimental protocol that combines MERFISH, multiplexed and error robust protocol of counting RNA transcripts using fluorescent in situ hybridization,(Chen et al., 2015; Moffitt et al., 2016) with rich profiling of the underlying cell state (Figure 1.2). We used the MCF10A mammary epithelial cell line, which is often used in studies of cellular variability due to their non-transformed nature and their accessibility to imaging(Qu et al., 2015; Selimkhanov et al., 2014). We focused on genes that share biological function: involvement in the Ca^{2+} signaling network, a key pathway important to the cellular response to tissue wounding(Justet et al., 2019; Minns and Trinkaus-Randall, 2016). The two advantages of Ca^{2+}

signaling are that 1. we expect that genes that share a function will show a high degree of correlation in their expression levels(Stewart-Ornstein et al., 2012). 2. Ca^{2+} signaling is fast and we can measure the overall emergent phenotype of the network in less than 15 minutes (Figure 1.2A). In our protocol cells were rapidly fixed after live cell imaging (10-15 min from ATP stimulation to fixation) and therefore the gene expression measured in the same cell is unlikely to have changed as a result of the agonist.

MERFISH is a multiplexing scheme of smFISH where transcript identity is barcode-based, and the barcodes are imaged over several rounds of hybridization. During each hybridization round, dye-labeled oligos are hybridized to a subset of RNA species being measured, the sample is imaged and RNA appear as diffraction limited spots, then the dye molecules are quenched, and the process is repeated until all barcode ‘bits’ are imaged. By linking diffraction limited spots across imaging rounds, we can decode the RNA barcodes by identifying the subset of images where a bright diffraction limited spot appears at the same XYZ coordinate (Figure 1.2B). The use of combinatorial labeling allows exponential scaling of the number of genes images with the number of imaging rounds. The scaling is mostly limited by the built-in error correction(Chen et al., 2015). In this experiment, we used 24 imaging rounds (8 hybs x 3 colors) where each RNA molecule was labeled in 4 imaging rounds. An example of the MERFISH data is shown in Figure 1.2B. Overall we measured the expression of 150 genes including 131 genes annotated as involved in Ca^{2+} signaling network(Bandara et al., 2013; Kanehisa et al., 2019; Kanehisa and Goto, 2000), 17 genes to mark stages of the cell cycle(Whitfield et al., 2002), and two genes that correlate with the sub-differentiated state of MCF10A cells(Qu et al., 2015).



Extrinsic Factor	MLR Variables
Volume	Cytoplasmic um^3
Cell Cycle	G2M Score
	S Score
MCF10A Differentiation	CD44 Expression
	GNA11 Expression
Calcium Response	Calcium Max
	Time of Max
	1st Peak FWHM
	1st Peak Decay Time
	Lowpass AUC
	Highpass AUC
	# Peaks Lowpass
	# Peaks Highpass

Figure 1.2 Paired single-cell MERFISH and live-cell calcium imaging. (a) Experimental overview - live cells are imaged for their calcium response to ATP before being fixed and imaged to measure gene expression of 150 genes. (b) smFISH spots are imaged over several rounds of hybridization and aligned such that individual genes are encoded as specific series of dark and bright spots throughout all rounds of hybridization. (c) Left, representative calcium trajectories demonstrating the heterogeneous response to ATP stimulation, top vs bottom left. The Right panel is an image plot of all 5000+ successfully paired to smFISH cells. (d) Cellular volume is measured and the correlation between total transcripts per cell and the cellular volume is shown. (e) Left, shows marker gene expression for cell cycle related genes used to derive a g2m score (coloring). Right, is the same as the left panel with a representative gene used to derive the S score for each cell. (f) Correlation of a representative gene (ATP2A2) with a gene that marks the differentiation status of MCF10A cells (CD44). (g) table of the cell state features categories and the complete list of the 13 factors used in the multiple linear regression (MLR) statistical model.

Our decomposition into allele-specific and cell state components is based on conditioning on multiple cell state factors. While, it would be ideal to directly measure the regulatory factors that causatively control gene expression variability, more accessible measurements, e.g. cell size or cell cycle stage, that are correlated with these causative regulatory factors are sufficient for the conditioning process. Given that the genes we probe are related to Ca^{2+} signaling we first extracted key features from time-series of cytoplasmic Ca^{2+} response measured with a calibrated GCaMP5 biosensor (S1.2 A). The live cell imaging of cytoplasmic Ca^{2+} levels (Figure 1.2C) showed a highly heterogeneous response, qualitatively and quantitatively similar to previous work on Ca^{2+} signaling in MCF10A cells where we observed a mixed population response with a wide range of response phenotypes(Handly and Wollman, 2017; Yao et al., 2016). We used a feature-based representation of Ca^{2+} response to represent cellular factors that we anticipate correlate with underlying cell state (Figure 1.2G and S1.2). In addition to Ca^{2+} features that are specific to Ca^{2+} signaling, we also measured a few global features of the cell that are likely to be correlated with expression changes of most genes. Specifically, we measured cell volume, cell cycle stage, and two markers of MCF10A differentiation status (Figure 1.2 DEF). As was shown in the past, cell volume strongly correlated with the total number of transcripts per cell (Figure 1.2D) indicating that at least for some genes cell state factors must be important contributors to their expression variability(Hansen et al., 2018a; Padovan-Merhar et al., 2015). However, not all genes show the same strength correlation with volume, and some cell cycle genes are more complexly related to volume (S1.4). Similarly, the cell cycle stage and MCF10A differentiation status were correlated with specific genes (Figure 1.2EF). Overall we measured 13 different cellular features that will be used to decompose variance in all 131 Ca^{2+} related genes we measured. By focusing on a smaller number of specific features that relate to the Ca^{2+} response augmented by established global cell state features like cell size and cell cycle state we expected to be able to capture most of the expression variability that comes from underlying cell state heterogeneity.

To decompose the observed expression into multiple components we used standard multiple linear regression (MLR)(Battich et al., 2015; Hansen et al., 2018a). Figure 1.3A shows the scatter plots of expression of two representative genes (ATP2A2 and RRM1) plotted against cell volume, cell cycle, differentiation markers, and Ca^{2+} feature. The scatter plots show that 1. there is indeed a correlation between expression and some of these cell state features 2. The amount of variance that is explained by each cell state feature can change between genes. Overall the simple MLR model with 13 independent measurements was able to explain between ~15-85% of the observed variance with a median of 0.62 (Figure 1.3B). To assess the relative contribution of each cell state feature we looked into the relative fraction of explanatory power for each feature category (Figure 1.3C). Overall, cell volume has the most explanatory power, but for some genes, cell cycle and Ca^{2+} features contribute meaningfully to the explained variance. While some of the features had a small effect in terms of the overall variance explained by the feature, in most cases, the effects were very unlikely to be a result of pure random sampling, permutation-based statistical testing showed that most genes measured here are statistically correlated with at least one calcium feature (Figure 1.3D).

A key uniqueness of our approach is that gene expression is measured in a multiplexed fashion allowing the estimation of the correlation between genes. Figure 1.4A shows the correlation matrix of the raw counts, and the counts conditioned on cell state features. As expected, as we increase the number of cell state features included in the MLR, the overall gene to gene correlation goes down. Interestingly, the full MLR model, that only includes 13 identical terms for all genes is able to reduce the overall correlation between gene significantly. To quantify the bulk correlation we measured the amount of variance that is explained by the first two components of a Principal Component Analysis (PCA) (Figure 1.4B). Without conditioning on any cellular feature, the first two components explain >40% of the variance. This is reduced substantially to <10% of the overall variance, in the full MLR. The substantial reduction in the gene to gene correlation demonstrates that we were able to condition away most of the shared components. Still, the remaining correlation was not completely removed and therefore we added another term to the model that is based on the first two principal components of a PCA analysis after taking all other features into account. These two components most likely represent some cell state features that were not sufficiently captured by our 13 cellular features. With the addition of the last “hidden” feature, the overall variance that is shared is very close to values from shuffled data. Overall the analysis of expression covariance demonstrates that our simple MLR sufficiently captures most of the information related to cell state that is required for conditioning expression distribution.

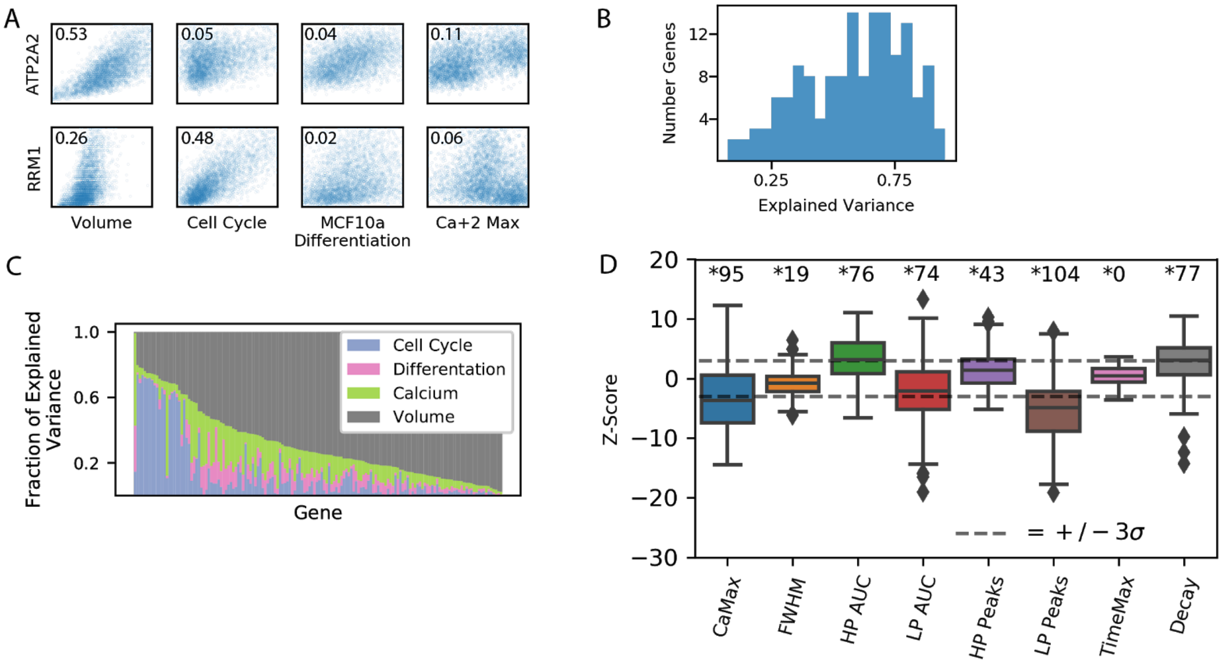


Figure 1.3 Decomposition of gene expression variability using multiple linear regression. (a) Representative scatter plots of correlation between two individual genes (rows) and different cell state factors (columns). The percent of variance explained by each factor in the MLR model for each gene is annotated in the corner. (b) A histogram of the overall explained variance for each gene. (c) Stacked bar plot showing which cell state features categories contribute to the explained variance of the MLR. (d) The significance of calcium features for each gene were estimated by Z-scoring the slope of the feature in a null distribution of bootstrapped shuffled data slopes. The number of statistically significant genes for each feature is shown above (adjusted P-value (Bonferroni) < 0.05).

Finally, we wanted to determine the overall dispersion remaining in the allele-specific gene expression distribution. The allele-specific variability is estimated as the residual variability in the raw gene expression counts after conditioning on cell state factors. As we increase the number of cell state features we conditioned on, we saw a substantial reduction in the distributions of Fano factor magnitudes (Figure 1.4C). When all 13 cell state features and the two hidden features estimated based on PCA are included, the Fano factor is very close to one for most of the genes. Note that we do not perform any correction for technical noise so the limit of one is only theoretical. Similarly, analysis of the coefficient of variation square (CV^2) vs the expression means on a log-log plot shows that all genes are very close to the Poisson limit (Figure 1.4D). The proximity to the Poisson limit is similar across all expression levels. Therefore, these data indicate that super-Poissonian transcriptional bursting plays a very minor role in allele-specific variability. It is unclear if the few genes that do show over-dispersion whether they have significant levels of transcriptional bursting or whether our conditioning procedure failed to sufficiently remove cell state effect.

Discussion

Here we analyzed the relative contribution of gene specific variability that arises from transcriptional bursting, i.e. episodic synthesis of multiple transcripts from a gene, and variability that is shared among multiple genes. Our approach is enabled by very rich single cell measurement that include live cell Ca^{2+} response to ATP, global cell state factors such as size and cell cycle stage, and the expression level of 150 genes all in the same single cells. Using this data, we were able to decompose gene expression variability into gene-specific and cell state components. We show that after removing covariability from gene expression distributions, the remaining variability follows a simple Poisson model. The residual allele specific variability is not over-dispersed and therefore not consistent with models of transcriptional bursting where a gene is actively transcribed only during a small fraction of time.

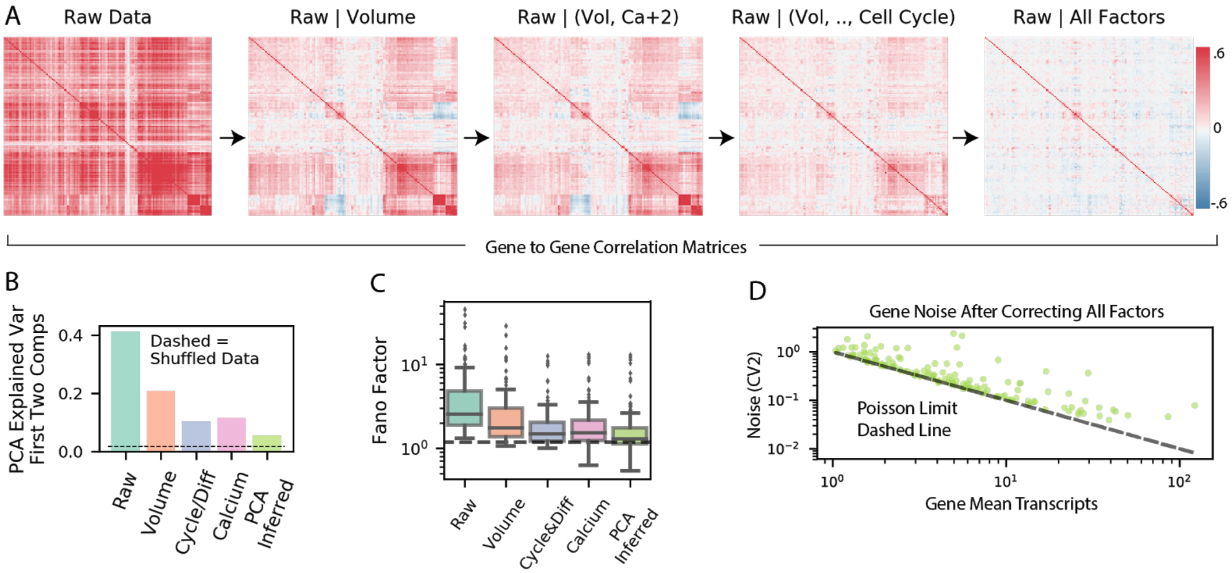


Figure 1.4 Residual variability from MLR models contains significantly less covariation between genes and close to poisson variability within individual genes. (a) Gene-gene correlation matrices showing the reduction of covariance after conditioning on cell state features. (b) Explained variance of first 2 components of PCA for each stage of MLR models showing reduction in shared variability with increasing number of cell state factors. (c) Fano factor distributions at different levels of cell state conditioning are shown as boxplots. Dashed line is the Poisson expectation. (d) Scatter plot of residual gene expression coefficient variation squared for each gene after decomposition of all cell state features. Poisson expectation is shown as dashed line.

The popularity of the transcriptional bursting model is evident by the large number of papers that fit the entire RNA and protein distributions to the two state model without considering other sources of variability(Dey et al., 2015; Molina et al., 2013; Skupsky et al., 2010; Suter et al., 2011). In other cases, cell state was considered using dual reporters(Sigal et al., 2006; Strebinger et al., 2018), assuming timescale separation(Dar et al., 2012), or conditioning on forward scatter(Sherman et al., 2015). However, without multiplexed expression measurements it is difficult to determine whether conditioning on cell state was done to completion. The high goodness of fit of the two-state model to uncorrected or partially corrected distributions that shows substantial bursting could simply be a case of over interpretation of model fit. RNA binding systems, such as MS2, allow direct live-cell observation of transcription bursting, and many groups have observed burst-like punctuated transcription(Corrigan et al., 2016; Ferguson and Larson, 2013; Fritzsch et al., 2018; Muramoto et al., 2010). While direct visualization is compelling, it is unclear if punctuated transcriptional events are due to stochastic transition of promoter state, as suggested by two state model, or due to stochasticity in the activity of an upstream regulatory element. Furthermore, difficulty in quantifying the number of mRNAs synthesized in each such event make it difficult to distinguish between a two-state model and a one state model with a low rate of transcription that will generate a Poisson distribution. In fact, our results are consistent with recent measurements that showed that TTF1 mRNA is generated in "bursts" of 1-2 mRNA(Rodriguez et al., 2019). Furthermore, the two alleles of TTF1 showed coordination between these bursts suggesting that the observed transcriptional events are coupled through trans-regulatory factors. Finally, temporal changes in global rates of transcriptions(Shah et al., 2018; Skinner et al., 2016) can also make the interpretation of a single allele temporal reporter challenging. It is important to note that our work focuses on genes that encode for calcium signaling activity and might not represent all genes, such as reporters controlled by viral promoters(Dar et al., 2012; Singh et al., 2010) and genes that are key to cellular differentiation(Hansen and van Oudenaarden, 2013; Ochiai et al., 2014). Overall it is advisable to use more caution when interpreting gene expression variability as evidence of transcriptional bursting.

Our measurements are based on cytoplasmic RNA and it is possible that mechanisms related to RNA processing reduce the dispersion of RNA distribution in the cytoplasm after it was generated in an over-dispersed manner through bursting(Battich et al., 2015). Cells include a large number of RNA binding proteins many with unknown function and it is possible that some function as part of post-transcriptional noise reduction mechanisms(Hansen et al., 2018b). However, some of the proposed mechanisms such as nuclear export of RNA were shown to act as amplifiers of observed dispersion(Hansen et al., 2018a). Therefore the degree by which post-transcriptional mechanism can be used to reduced expression noise is an important open question. Until additional data will help clarify the ubiquity of such mechanisms, the most parsimonious interpretation is simply that RNA synthesis does not happen in large allele-specific bursts.

Recent technological advances in the ability to measure single cell gene expression with scRNAseq and sequential smFISH approaches are providing an unparalleled view into the underlying “cell state space”. The distribution of cells in “cell state space” and the definition of cell types and states within this space are key open research areas that will likely to further grow in importance with further improvements in single cell measurement technologies(Eng et al., 2019; Wagner et al., 2016). Our work has two important implications on our understanding of this “cell state space”, at least with regards to the heterogeneity of a single cell type: 1. All the shared variability was reduced using only a simple representation of cell state as 13 linear coefficients. Furthermore, most of these 13 features had only a very small contribution to the overall explanatory power suggesting that cell state distribution can be represented by few latent dimensions. An observation that emboldens efforts to learn the cell state manifold(Moon et al., 2018). 2. Expression noise, i.e. unregulated variability in gene expression that is a result of stochastic biochemical interactions in effect defines a “resolution limit” of the cell state space. Our results indicate that the highly heterogeneous distribution of cells within cell state space is likely not due to the inability of cells to control their expression levels rather our work indicates functional stratification of cells within this space. Collectively these contributions pave the way to a more rigorous definition of cell state that is based on concepts of signal to noise where the signal is represented by regulated differences between cells and noise is due to unregulated stochastic events. Such definitions will help identify the functional role of cellular heterogeneity.

Acknowledgements

Chapter 1 is a full reprint of the material as it appears in Mammalian gene expression is explained by cell state in bioRxiv 2019. Foreman, Robert and Wollman, Roy. The

dissertation author was the primary investigator and author of this material. Anna Pilko is not an author, but constructed the GCaMP5 MCF10A cell line for a previous work.

Materials and Methods

Cell Culture

MCF10a cells were grown in complete media (above) and passaged at 70-90% confluency. Cells were seeded onto coated 40mm #1.5 coverslips (Bioptech) and grown to confluence in 5mm diameter PDMS wells 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 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 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.

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).

Coverslip Modification

40mm coverslips (Bioptech) were allyl-silane functionalized according to (Moffitt et al., 2016) which briefly consists of washing coverslips in 50% methanol and 50% 12M HCl, and then incubating at room temperature in 0.1% (vol/vol) triethylamine (Millipore), 0.2% (vol/vol) allyltrichlorosilane (Sigma) in chloroform for 30 minutes. Washing with chloroform then 100% ethanol and air drying with nitrogen gas. These were stored in a desiccator for less than a month until use.

Calcium Imaging

Cells were stained with 0.1 ug/mL Hoescht for 20 minutes then rinsed with imaging media. Each well was imaged and stimulated consecutively as follows: image 3 minutes of Gcamp before stimulating with 6uM ATP in imaging media then imaged for another 13 minutes. Gcamp was imaged every 2-3 seconds and Hoechst was imaged every 4 minutes for segmentation. Immediately following imaging of a well, that well was fixed with 4% formaldhyde in PBS. The next well was imaged, and then the previously imaged/fixed well was washed 3X with PBS.

Sequential FISH Staining

PDMS wells were removed and cells were briefly fixed for 2 minutes, washed 3X with PBS, and then permeabilized with 0.5% Triton X-100 in PBS for 15 minutes. Coverslips were washed 3X with 50mM Tris and 300mM NaCl (TBS), and then immersed in 30% formamide in TBS (MW) for 5 minutes to equilibrate, all the liquid was aspirated from the petri dishes, and 30uL of 75uM encoding probes and 1uM locked poly-T oligos were added on top of the coverslip and a piece of parafilm was place on top of the coverslip to evenly spread the small volume over the surface and prevent evaporation. The entire petridish was also sealed with parafilm and incubated at 37C for 36-48 hours. The parafilm was removed and the coverslip was washed 2X with MW buffer with 30 minute incubation at 47C for both washes. A 4% polyacrylamide hydrogel was then cast to embed the cells before clearing with 2% SDS, 0.5%

Triton X-100, and 8U/mL proteinase k (NEB P8107S), according to previously published methods. Coverslips were incubated in clearing buffer for 24 hours then washed 3X in TBS for 15 minutes each at room temperature. (Moffitt et al., 2016)

Sequential FISH Imaging

smFISH staining was imaged on a custom modified Zeiss Axiobserver Z1 body with Andor Zyla 4.2 sCMOS camera and 1.4NA 63 Plan-Apo oil immersion objective. Illumination light was provided by luxeon rebel LEDs (Deep Red, Lime, Blue, and Royal Blue) to excite Cy5, Atto565, Alexa 488, Hoechst, and 200nm Deep Blue fiducial markers. The microscope was controlled by micro-manager (Ausubel et al., 2001) and custom MATLAB software. Automated washing during sequential rounds of hybridization was accomplished by using a previous published setup (Moffitt et al., 2016; Moffitt and Zhuang, 2016). Briefly, FC2 bioptech flow chambers were attached to a gilson minipuls peristaltic pump pulling liquid from reservoirs attached to hamilton MVP valves. The pump and valves were controlled with arduino, and serial commands with Python https://github.com/ZhuangLab/storm-control/tree/master/storm_control/fluidics. This setup was used to automatically wash cells with TBS, then 2mL of TCEP (Sigma) in TBS incubated for 15 minutes, then rinse with TBS, then flow in 2mL of wash buffer (10% ethylene carbonate in TBS with 2mM Vanadyl Ribonucleoside Complex (NEB)), followed by 3mL 3nM readout probes in wash buffer incubated for 15 minutes, then rinsed with 2mL wash buffer, then 1mL of TBS, and finally 3mL of imaging buffer. Imaging buffer is 0.15U/mL rPCO (OYCO), 2mM PCA (Sigma), 2mM Trolox (Sigma), 50mM pH 8.0 Tris-HCl, 300mM NaCl, and 40U/mL murine rnase inhibitor (NEB).

FISH Oligo Pool Design Amplification

Oligopools were ordered from CustomArray. The oligos were designed using previously published software (Moffitt and Zhuang, 2016). Briefly, design involves selecting 30bp regions with 40-60% GC for each target gene that maximizes specificity of the oligo by finding shared 15-mer substrings against all other transcripts in the human genome. These regions are concatenated with sequences for 3 readout probe binding sequences and flanking 20bp Primers. Probes were amplified according to the another previously published work (Wang et al., 2018). Briefly, limited cycle qPCR with a T7 promoter on the reverse primer. The PCR was terminated 1 cycle after saturation during the extension phase. PCR product was column purified, then in vitro transcription further amplified the oligos (NEB Quick High Yield Kit), t7 reactions were purified with desalting columns, and converted to ssDNA with Maxima RT H- (Thermo).

Gcamp Image Processing

Cell nuclei were segmented using custom Python 3.6 scripts. Cell nuclei were segmented using the Hoechts staining. Nuclear images were low pass filters with gaussian of sigma 5 pixels. Then regional maxima were found with corner_peaks from scikit-image these peaks were used as seeds in a watershed of the negative intensity of the images, and thresholded with otsu of the smoothed nuclear images. This was repeated for each time point and the centroid of each nuclear mask was tracked across time using linear assignment. Segmented nuclei were used as masks to calculate the mean intensity within each cell mask in the Gcamp channel and also the channel for mCherry-fusion expression marker for Gcamp. Finally Gcamp values were divided by the mCherry values to give expression normalized calcium trajectories.

Calcium Trajectory Feature Extraction

Calcium trajectories were processed with wavelets to find lowpass, smoothed, and highpass trajectories by thresholding coefficients of different scale wavelets. Peaks were detected in the lowpass and highpass with scipy’s find_peaks and prominence thresholds of 0.1 and 0.15

respectively. Decay time of the first major peak after ATP stim was calculated, FWHM of the first peak after ATP was calculated, the AUC of highpass and lowpass was calculated with numpy's trapz, the maximum of each calcium was calculated, and the time of maximum was also calculated from smooth trajectories.

Alignment to Live Cell Images

EM microgrids (G400F1-Cu EMS) were glued (23005 biotium) to 40mm Bioptech coverslips. These grids were imaged in brightfield to determine the stage coordinate of fiduciary marks on the microgrids. A rotation and translation transformation was fitted between the live cell and smFISH coordinates of microgrid fiduciary marks. This ensured that we imaged the same FOVs, but additional alignment was performed after imaging. smFISH images were downscaled until they had a pixel size matching the live cell imaging (63x vs 10x with same Andor Zyla Camera so 6.3X downscaling). Cross-correlation template matching with live cell templates and smFISH candidate images was performed iteratively with range of rotational angles (-5 to +5 degrees) in order a second set of 'image' translations and rotations that maximize the cross correlation scores. A threshold was then applied and downsampled images were stitched together and overlaid to confirm successful alignment.

smFISH Image Alignment

All rounds of hybridization contained 200nm blue beads (F8805 ThermoFisher) that were imaged in addition to smFISH oligos. First the coordinates of putative beads were determined with subpixel accuracy by upsampling images by a factor of 5 (~20.5 nm pixel size) and finding peak coordinates of normalized cross-correlation between a gaussian 'bead template' and bead images in 3D. Next a translational transformation was estimated from these putative beads with a custom algorithm designed to be robust to false detection of beads. Briefly, neighborhoods of beads with a radius of maximum shift (100 pixels), were found and the differences each of these pairs was calculated. Next, the bead coordinate differences were density clustered and bead pairs from the largest cluster were used in a least square error optimization of translation vector that minimizes residual of all bead pairs after translation. This fit was performed in 3D and any FOVs with a residual >0.5 pixels XY or 1.2um (3 frames) in Z were discarded.

Chromatic Aberration Correction

Tetraspeck (4-color) 100nm beads were imaged in all channels used for smFISH imaging. The subpixel centers of these beads were found as described above, and the misalignment of channels was calculated as a function of the XY image coordinate. Images were then interpolated in 2D to correct for systematic differences between channels. (Mostly only necessary at the edges of the images due to large camera sensor size).

Gene Calling

Spots were called with a reimplemented algorithm deeply inspired by (Moffitt et al., 2016), and the code is available at https://github.com/wollmanlab/PySpots. Images were taken every 0.4um in Z, but groups of 3 images 1 above and below the current Z slice being processed were maximum projected to form a pseudo Z slice to be further processed. Then two Z slice were skipped before form another pseudo Z slice. This local max projections help gene calling perhaps do to making the imaging more robust to misaligned images, or uncorrected planarity issues in the objective. Second, fiduciary 200nm beads were used to fit XYZ translation transformations described in the image alignment section, and all psuedo Z slices were warped to correct for chromatic aberration and translations from stage reproducibility error. Registered and chromatic aberration fixed images were then high pass filtered by subtracting a gaussian convolution with sigma 2.2 pixels from the original images. These high pass filtered images were then deconvolved for 20 iterations of lucy richardson deconvolution using the flowdec package.(Czech et al., 2018) Finally after deconvolution the images were blurred by gaussian convolution with a sigma of 0.9 pixels. The output at this step for each site imaged is a matrix of (2048, 2048, 24, #Z) elements. Where 2048 is the image width and height, 24 is the number of codebits used to encode gene identity (3 colors X 8 rounds sequential hybridization) and #Z is the number of pseudo Z slices. Next, each Z slice was processed separately on a per pixel basis to assign each pixel as its gene identity or as background. This process was done by dividing each of the 24 images by the 95th percentile of that image to make the intensities for different codebits more similar, and L-2 normalizing each pixel. Then for each pixel the Euclidean distance to L-2 normalized codebit vectors was calculated, and if that distance was less than the volume of a nonoverlapping hypersphere for all codewords (0.5176) then the pixel was classified as that closest codeword. This approach is essentially testing whether the intensities from all 24 codebits point in the direction of a particular codeword in 24-dimensional space. Finally, these classified images (2048, 2048, #Z) were segmented to collect groups of connected components with that same gene label. Finally genes calls were thresholded on the number of pixels for each group of connected components, and the average intensity of the set of connected components.

Calculation of Cell Volume

A 3-D histogram of gene calls for each cell was calculated and smoothed with a gaussian filter of 10 pixels. The number of voxels (1um, 1um, 1um) with at least 0.5 RNA was calculated and used as the volume for each cell.

Simulation of Gene Variance Decomposition

For each of the three combinations of cell state and allele-specific noise simulations there were three transcription factors and two genes simulated. Transcription factors were poisson distributed, and genes were simulated as gamma distributions with shapes dependent on additive combinations of transcriptions 1, 2, and 3. The scale of the gamma distributions were varied to control the amount of ‘allele specific variability’, and the amount of gene correlation was controlled by the fraction of shape shared between genes. For each of the 3 combinations of different noises there were 4 linear models fitted using python statsmodels ols package. For each gene a model was fitted for gene ~ tf1 and gene ~ tf1+tf2. Then the residuals from the fit were adjusted by adding back the mean of expression for that gene, and these mean adjusted residuals are the distribution of the gene conditioned on tf1 or (tf1, tf2).

Cell Cycle Features

Cell cycle features were calculated using the scanpy package (Wolf et al., 2018).

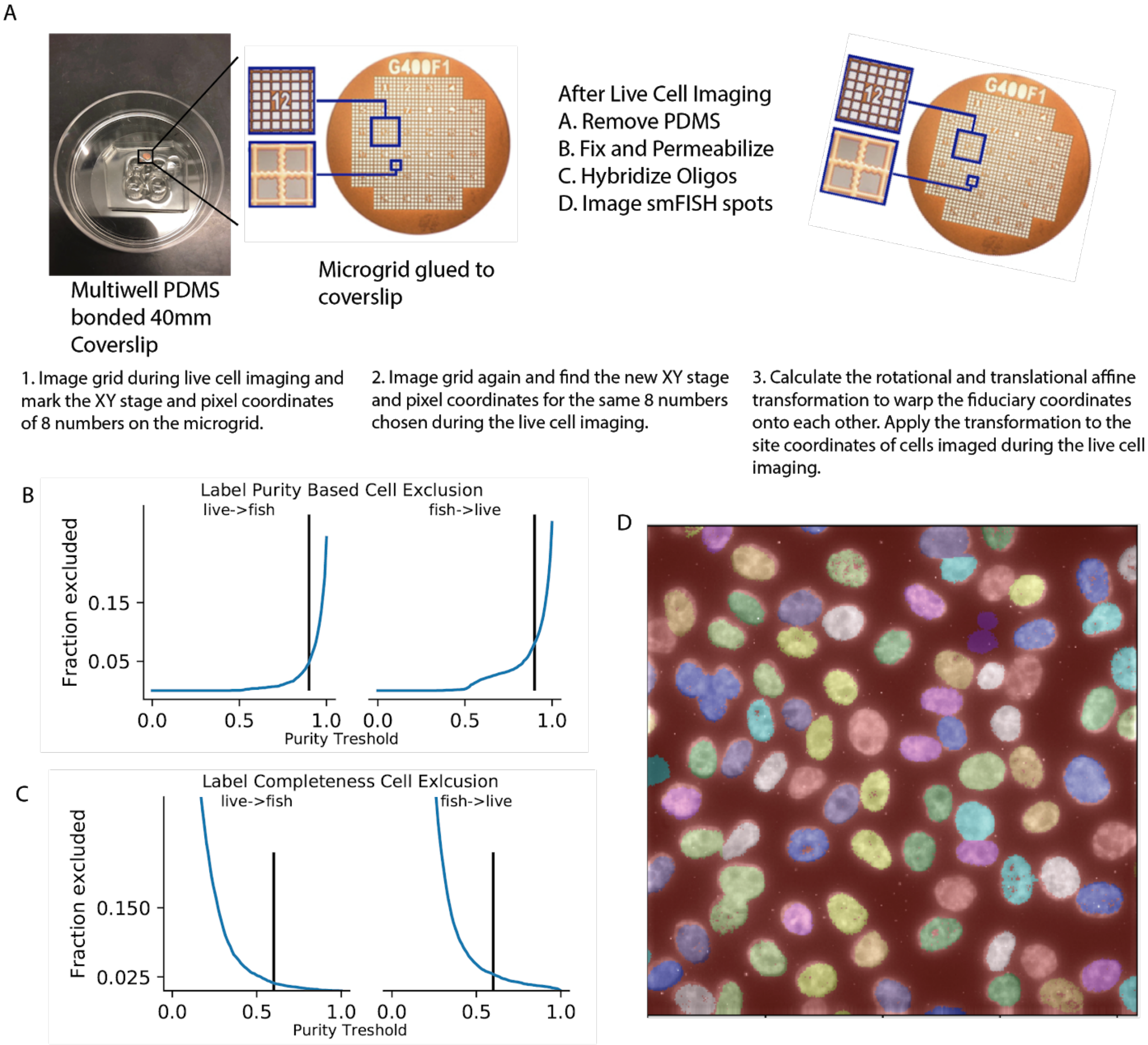
Gene Variance Decomposition

The same method (linear model residuals) as in the simulation was used to decompose variance for gene expression. In order to investigate residual correlations between genes with different sets of conditioning variables, the decomposition was repeated from different combinations of feature combinations. The first stage involved only gene ~ volume, and then gene ~ volume + s_phase + g2m_phase...finally for the inferred features we used PCA components #1 and #2 as features: gene ~ pca_comp1 + pca_comp2.

Statistical Test of Calcium Feature Significance

Volume adjusted gene expression counts were fitted with a linear model based on calcium features. For every gene separate and every calcium feature separately a shuffled linear model was also calculated. That is, for each calcium feature and gene many bootstrap models were estimated where a single calcium feature was shuffled and the model was fitted. The slopes of these fitted models on shuffled data formed a null distribution, and then the p-value of the feature for that gene was considered (100-Qtile(unshuffled slope in shuffled bootstraps)) where 0 is 1/#Bootstraps.

Supplemental Figures

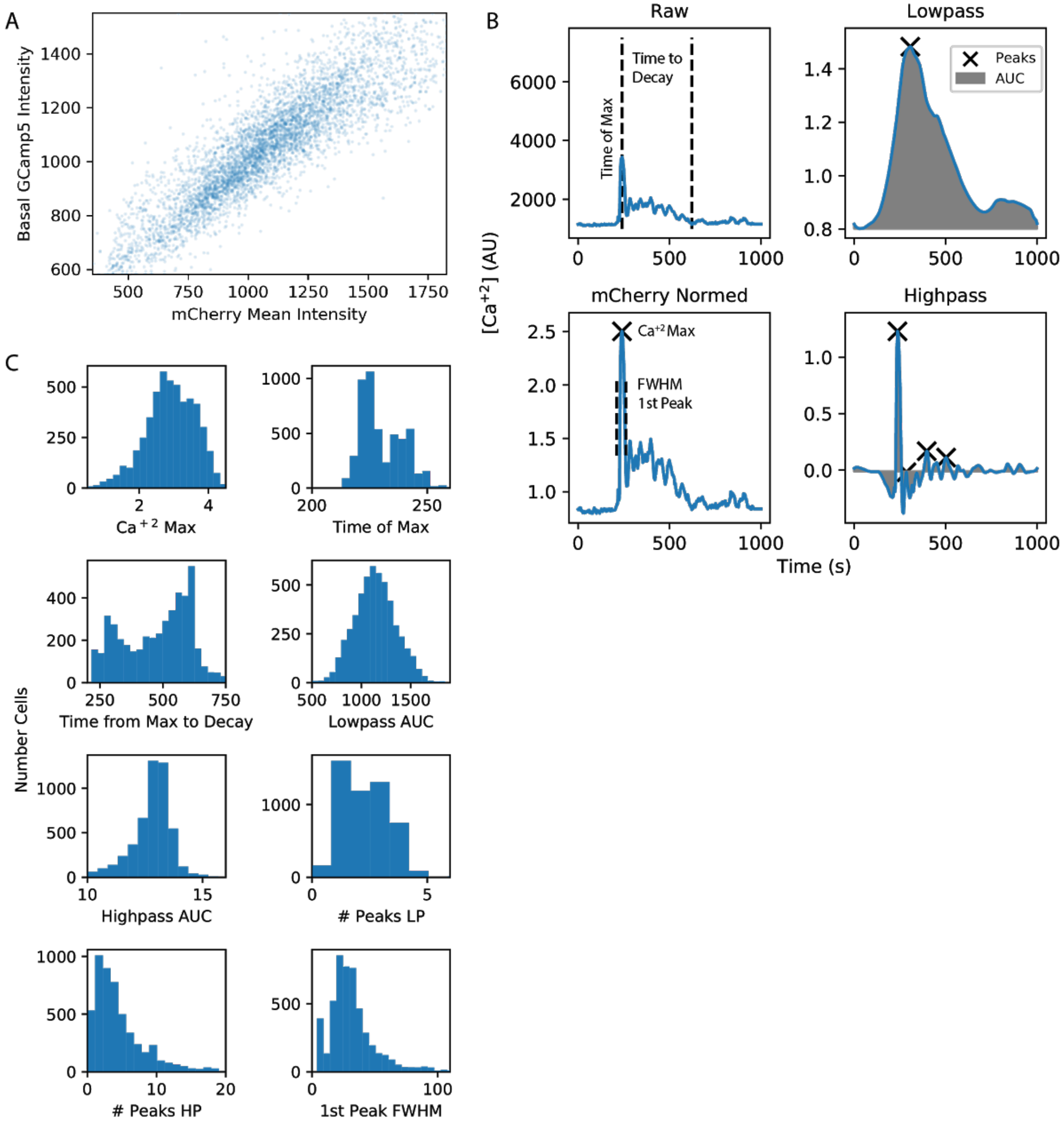


1. Image grid during live cell imaging and mark the XY stage and pixel coordinates of 8 numbers on the microgrid.

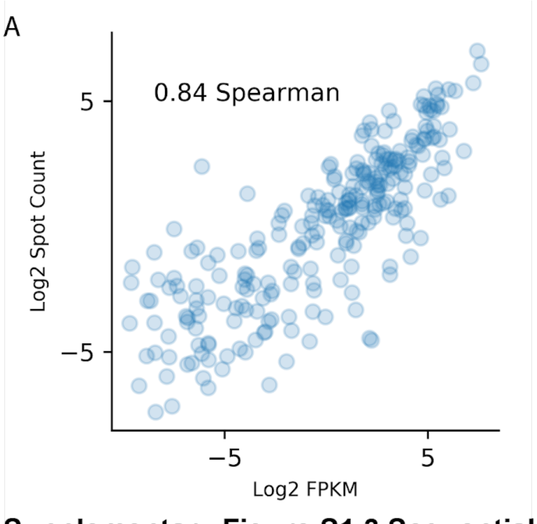
2. Image grid again and find the new XY stage and pixel coordinates for the same 8 numbers chosen during the live cell imaging.

3. Calculate the rotational and translational affine transformation to warp the fiduciary coordinates onto each other. Apply the transformation to the site coordinates of cells imaged during the live cell imaging.

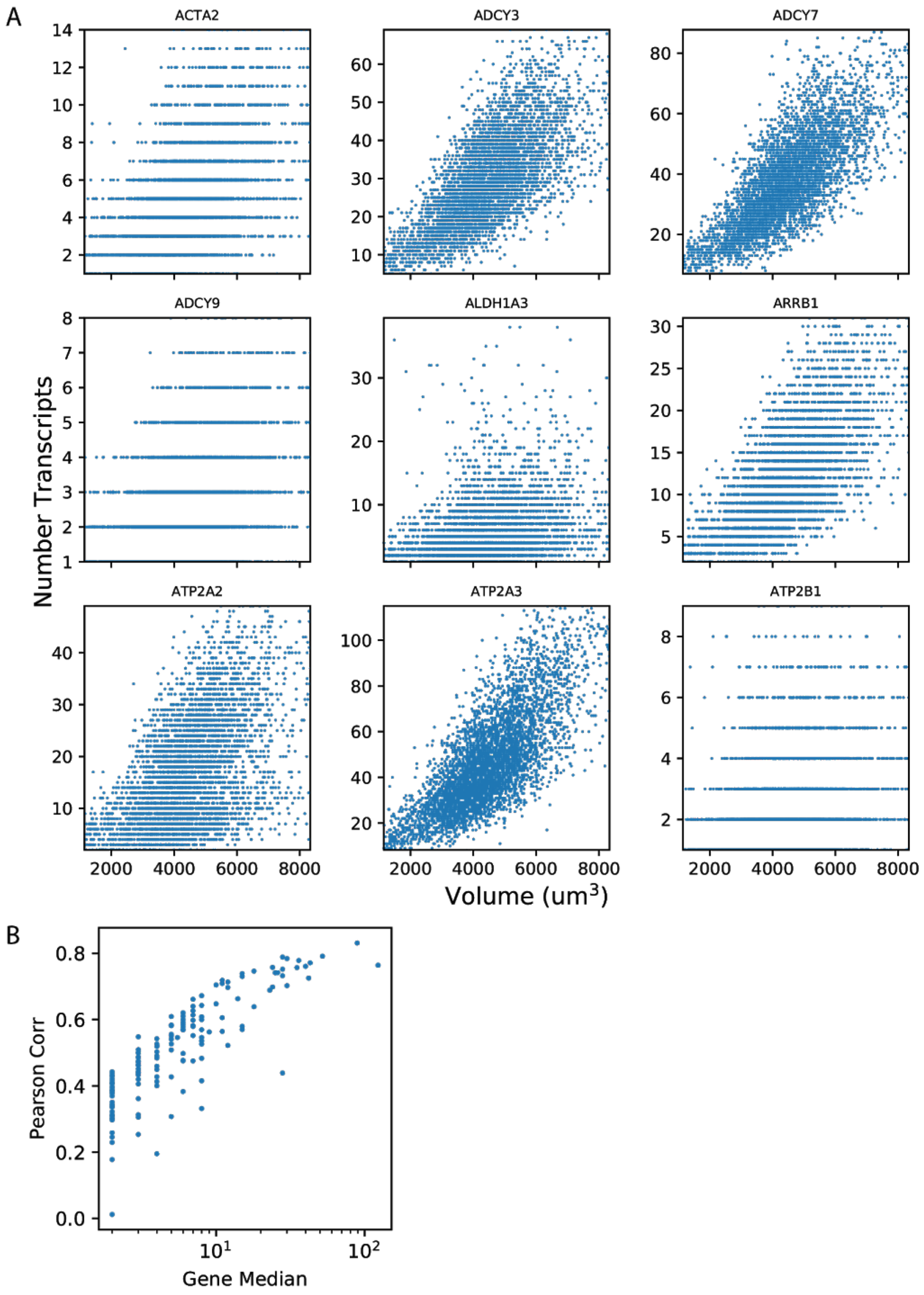
Supplementary Figure S1.1 - Alignment of calcium images cells with smFISH imaging cells and quality control filtering of cells. (a) PDMS wells and the fiduciary grid attached to a 40mm coverslip (left). Example of orientation of grid during the live cell imaging (middle). Right panel shows that during the smFISH imaging the grid can be rotated, but coordinates of the fiduciary numbers are identifiable. (Below) description of the steps taken to align the fiduciary grids. (b) Purity of label is the fraction of pixels in the calcium nucleus labels that are the same as the nucleus labels from the FISH imaging (left), and vice-versa (right). Vertical lines represent thresholds used to discard cells which were not uniquely mapped between live calcium and fixed FISH imaging. (c) Label completeness is the fraction pixels that were non-zero in the paired segmentation and vice-versa (left, right). Vertical lines were the thresholds of completeness used in quality filtering of alignments.



Supplementary Figure S1.2 Feature based representation of calcium trajectories. (a) Ca+2 is measured with GCaMP sensor fused with mCherry. The high correlation between basal GCaMP intensity and mCherry intensity indicates that mCherry intensity can be used to normalize sensor expression variability. (b) Top left shows raw GCaMP Ca+2 trajectories with the Time to Decay of 1st peak calcium feature, and the Time of Max feature for the example trajectory (dotted lines). Bottom left shows the mCherry normalized trajectory with the FWHM of 1st peak and the intensity value of Ca+2 Max feature for an example trajectory. Top right shows the lowpass filtered trajectory and the AUC (grey) as well as where the peaks in the lowpass are (x). Bottom right shows the same AUC (grey) and number of peaks (x) but for the high pass filtered trajectory. (c) Histograms of each calcium feature for all ~5000 cells are shown in each subpanel.



Supplementary Figure S1.3 Sequential hybridization smFISH is accurate measure of gene expression. (a) The scatter plot and correlation of RNA-Seq FPKM vs spot counts from the sequential smFISH measurements.



Supplementary Figure S1.4 The relationship between volume and gene expression counts for different genes. (a) Nine randomly selected genes are shown as a scatter plot of volume vs spot counts per cell. (b) The Pearson correlation of volume and gene expression as a function of the gene's median expression.

CHAPTER 2 - Relating Gene Expression Variability to Differences in Calcium Signaling Responses

Abstract

Signaling networks allow cells to respond to changes in their environment, both internally and externally. Despite the importance of accurate signaling, individual cells transducing signals often exhibit a large amount of cell-to-cell variability, and it is currently unclear how much of this heterogeneity is simply noise or actually related to phenotype functionality. In this work, use measurements of live-cell calcium signaling response to ATP stimulation in MCF10a mammalian epithelial cells, and also measure 336 gene expression levels using smFISH in the same cells as the calcium measurement. We find that 55% of the heterogeneity in the calcium signaling phenotypes is explained by systematic variation in the underlying gene expression state of cells. This finding suggests that cells could have high fidelity to respond to their environment, and do so in a stratified manner that is related to cellular state.

Introduction

Modern single-cell expression techniques such as scRNA-Seq and sequential hybridization smFISH are uncovering gene expression variability among populations of cells(Dixit et al., 2016; Wang et al., 2018). The new level of throughput of these techniques enables discovering and characterizing cell types/states at an unprecedented rate(Han et al., 2018; Suo et al., 2018; Tabula Muris Consortium et al., 2018). However, from simple gene expression studies alone it can be unclear whether every bit of gene expression variability is actually a manifestation of cell populations being stratified into functionally/phenotypically relevant states(Andrews and Hemberg, 2018; Nguyen et al., 2018; Shoval et al., 2012). Gene networks have been shown to exhibit a large degree of robustness, and redundancy in their function(Blanchini and Franco, 2011). Differences in gene expression between cells could still have the same phenotypic mapping due to robustness and redundancy(Tanaka et al., 2015; Whitacre, 2012).

In this work, we propose a new approach for identifying cell states, and simultaneously mapping the relationship between gene expression and emergent complex phenotypic behaviors such as signaling dynamics. We use live-cell calcium biosensor GCaMP5(Akerboom et al., 2012) to measure cytoplasmic free calcium in MCF10a breast epithelial cells in response to ATP stimulation. ATP is a wound associated ligand and induces heterogeneous calcium signaling responses in single-cell. The live cell calcium measurement is paired to fixed single-cell gene expression measurement of 336 genes using sequential hybridization smFISH (MERFISH)(Moffitt et al., 2016) by aligning cells from both rounds of imaging using fiduciary grids. The paired measurement leverages the power of joint measurements of gene expression state and an emergent complex phenotype to allow validation of putative cellular states. Simultaneously through the measurement of a large number of cells the covariance between gene expression and different dynamical features of calcium signaling can be naturally observed without perturbation.

It is not obvious that heterogeneity observed in calcium signaling signaling would be correlated with differences in gene expression. In many examples of signaling responses to ligand treatment, the heterogeneity in response is dominated by differences in the ligand's receptor expression between different cells(Cheong et al., 2011). The variability in receptor expression can often be explained by intrinsic post-transcriptional noise in expression(Hansen et al., 2018a) rather than cell state regulated differences between cells(Spencer et al., 2009). It is also very possible that signaling variability between cells can be explained by non-noisy systematic differences between cell states(Yao et al., 2016). In the ERK signaling pathway, multiple groups have observed that cell-to-cell variability is dominated by systematic differences between cells(Selimkhanov et al., 2014; Toettcher et al., 2013). However, in these ERK examples it is unclear whether the systematic differences result from post-transcriptional regulation of the signaling networks, or longer time gene expression based cellular states.

In this work, we find that in MCF10a cells the variability in cytoplasmic calcium signaling response to ATP is partially explained by cell state differences in gene expression. These cell states are correlated with differences in cell cycle and an orthogonal effect of MCF10a sub differentiation.

Results

Four overlapping questions will be addressed: 1. What gene expression states exist? 2. Do expression states correlate with calcium ATP response? 3. How much of the calcium heterogeneity is explained by gene expression? 4. How do different genes affect calcium signaling dynamics? We we will describe how we measure the calcium response, then how we make the gene expression measurements, and then we will address the four questions above.

MCF10a mammalian breast epithelial cells respond heterogeneously to stimulation with ATP, a ligand that is passively released from wounded cells(Handly and Wollman, 2017). Previous studies used modeling to suggest that the variability between cells has an extrinsic variation component rather than simple intrinsic noise in the signaling network itself(Yao et al., 2016). We sought to resolve how much of the heterogeneity results from cell state differences by simultaneously measuring single-cell response to ATP and the gene expression state of the calcium signaling network in the same cells. If the calcium response variability arises from post-transcriptional effects such as intrinsic fluctuations of calcium network components then we expected calcium response type will not be correlated with gene expression. Conversely if gene expression clusters explain the variability then cell state regulated differences manifest as differences in calcium signaling network that propagate to qualitatively different response dynamics to ATP stimulation.

This work measures calcium signaling with GCaMP5 and fluorescent live cell imaging(Akerboom et al., 2012). The result is a 1001 dimensional response across time for each single-cell, but due to the curse of dimensionality simple comparison of cell-to-cell similarity using euclidean distance fails. We embedded cells in a space where euclidean distance is a better measure of cell-to-cell similarity using a variational autoencoder. VAEs are generative models that compress a high dimensional signal or image into a lower dimensional 'latent vector' by optimizing a convolutional neural network that can successfully reconstruct the original high dimensional signal or image from the embedded 'latent vector'. This reconstruction optimization ensures that information is not lost in the embedding process. The 12 dimensional representation of calcium can further be reduced by techniques such as UMAP or tSNE in order to visualize cells in the new space, but all analysis except for during plotting was performed with the 12 dimension VAE representation of calcium response in cells (Figure 2.1CDEF).

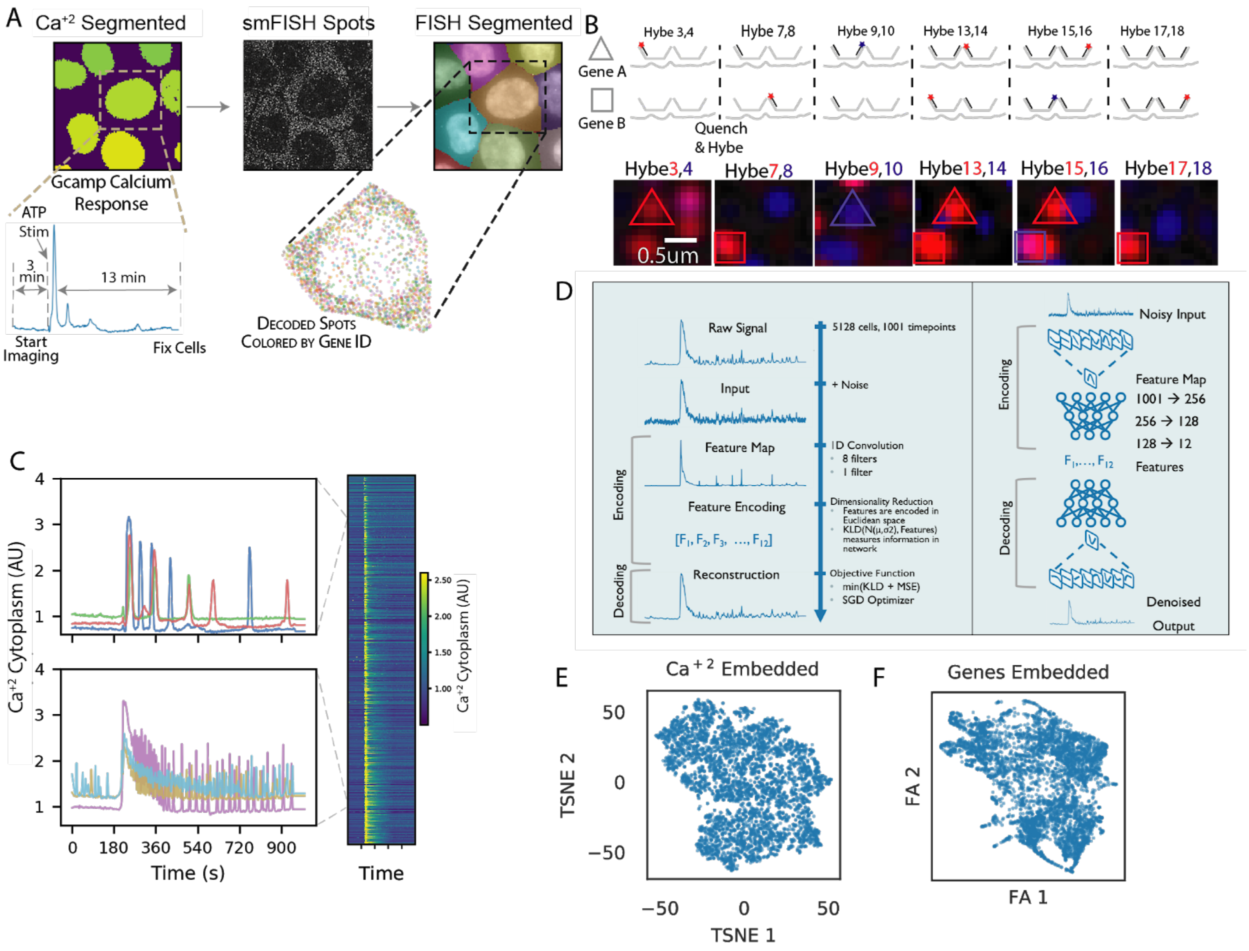


Figure 2.1 Calcium response to ATP in MCF10A cells is heterogenous. (a) Experimental overview that single cells are measured for their live cell response to ATP stimulation, fixed, and then imaged for single-cell smFISH measurement of gene expression. (b) The cells from calcium imaging are paired to gene expression measurements by using a fiduciary grid to align the images from both datasets. (c) Single-cell calcium trajectories showing that different cells respond to ATP with qualitatively distinct amplitudes and dynamics. (d) Variational autoencoder structure for representing calcium trajectories where similar cells are are less distant than dissimilar cells in the encoded latent vector. (e) TSNE embedding of 12 dimensional VAE to visualize calcium cells in 2D. (f) Force atlas projection of gene expression neighborhood graph (SCANPY).

Gene expression is measured using MERFISH(Moffitt et al., 2016); a technique that uses multiple rounds of hybridization and a barcoding strategy to measure 100s of genes' expression levels *in situ.* MERFISH works by encoding gene identities as a specific sequence of bright and dark spots at the same pixel coordinates across multiple rounds of hybridization, imaging, and quenching (Figure 2.1B). These combinatorial codes could scale to 10,000s genes, but are typically limited to 100s due to error-correcting codes and necessity of sparse spots in images(Wang et al., 2018). We measured 336 genes 300+ genes annotated as involved in calcium signaling, 16 genes related to cell cycle, and 2 genes associated with sub differentiation in MCF10a cells(Bandara et al., 2013; Kanehisa and Goto, 2000; Whitfield et al., 2002).

Genes were filtered by their expression level to remove any genes that were expressed at less than 2.65 transcripts/cell in the 95th percentile of cells. The counts of transcripts per cell were then normalized to cellular volume, measured through imaging, and log transformed with a pseudocount of one. These normalized gene counts were used to cluster cells based on their gene expression without knowledge of the calcium response using phenograph clustering based on Louvain modularity with k=13(DiGiuseppe et al., 2018). We observed that, despite being identified independently, the clusters in gene expression space demonstrated a distinct patterning when overlayed in the embedded calcium space (Figure 2.2A). Furthermore, when plotting the calcium trajectories grouped by cluster there are striking differences in the qualitative dynamics of the calcium response (Figure 2.2B). Additionally the dot plot of gene expression in the different clusters shows that there are significant differences across groups of genes. Together these data indicate that cell state maintained by gene expression programs are correlated with the qualitatively different patterns of calcium signaling dynamics in response to ATP.

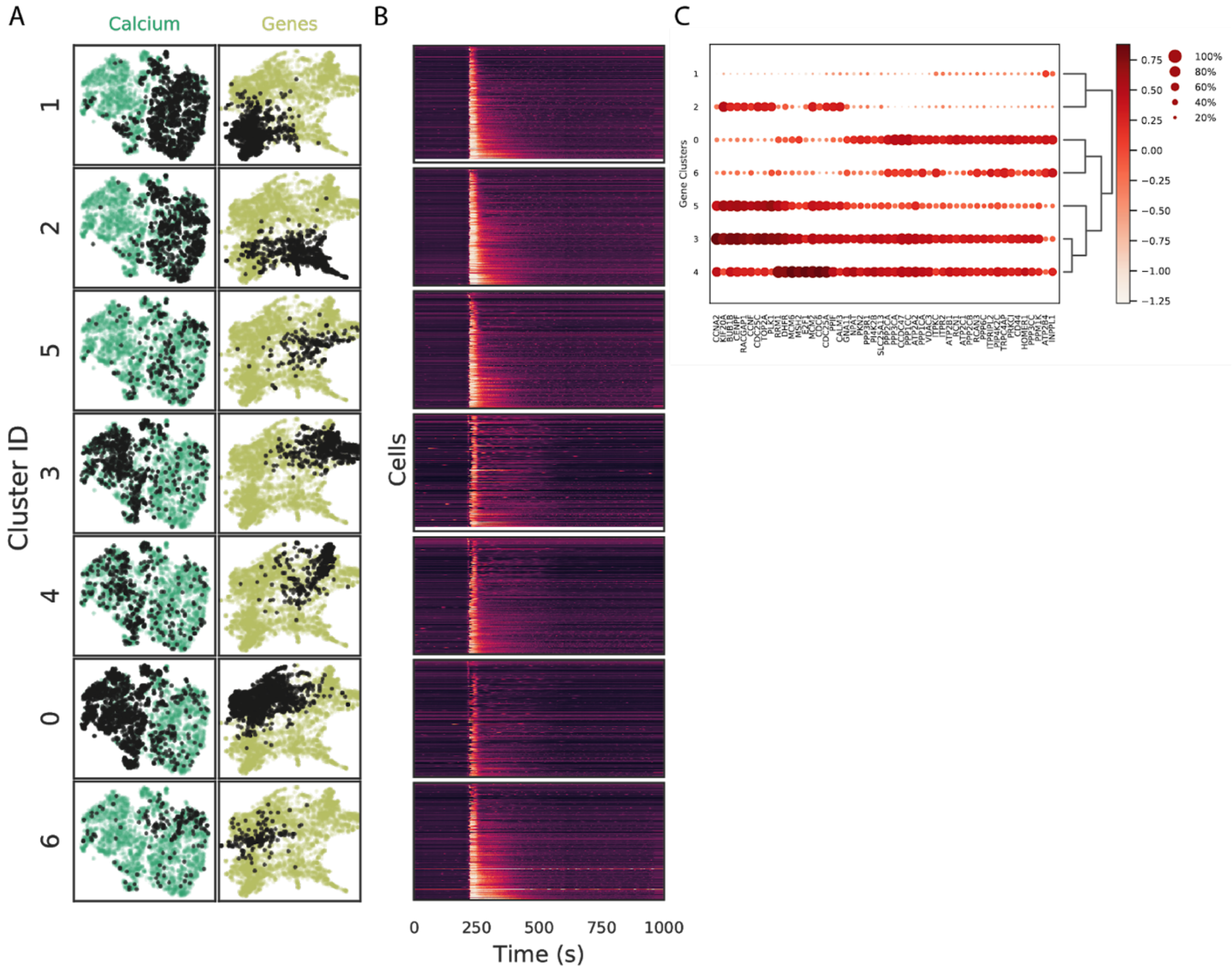


Figure 2.2 Gene expression differences are correlated with differences in calcium signaling dynamics. (a) Each cluster from gene expression is shown overlayed in black on calcium (left) and gene (right) TSNE and force atlas embeddings, respectively. (b) Calcium trajectory images for each corresponding cluster. Brighter colors indicate higher cytoplasmic free calcium. (c) Dot plot of top 40 differentially expressed genes in the clusters. Dot size indicates the percentage of cells and dot color indicates expression level.

Figure 2.3A shows expression of marker genes for cell cycle and 2.3B shows the expression of a sub differentiation marker, CD44. Differences in these two cell state components explain the two major axes of differences in gene expression. Interestingly, the two cell state terms are orthogonal. Further, specific calcium pumps measured vary independently with either cell cycle, or the sub differentiation state of the cells(Figure 2.3C).

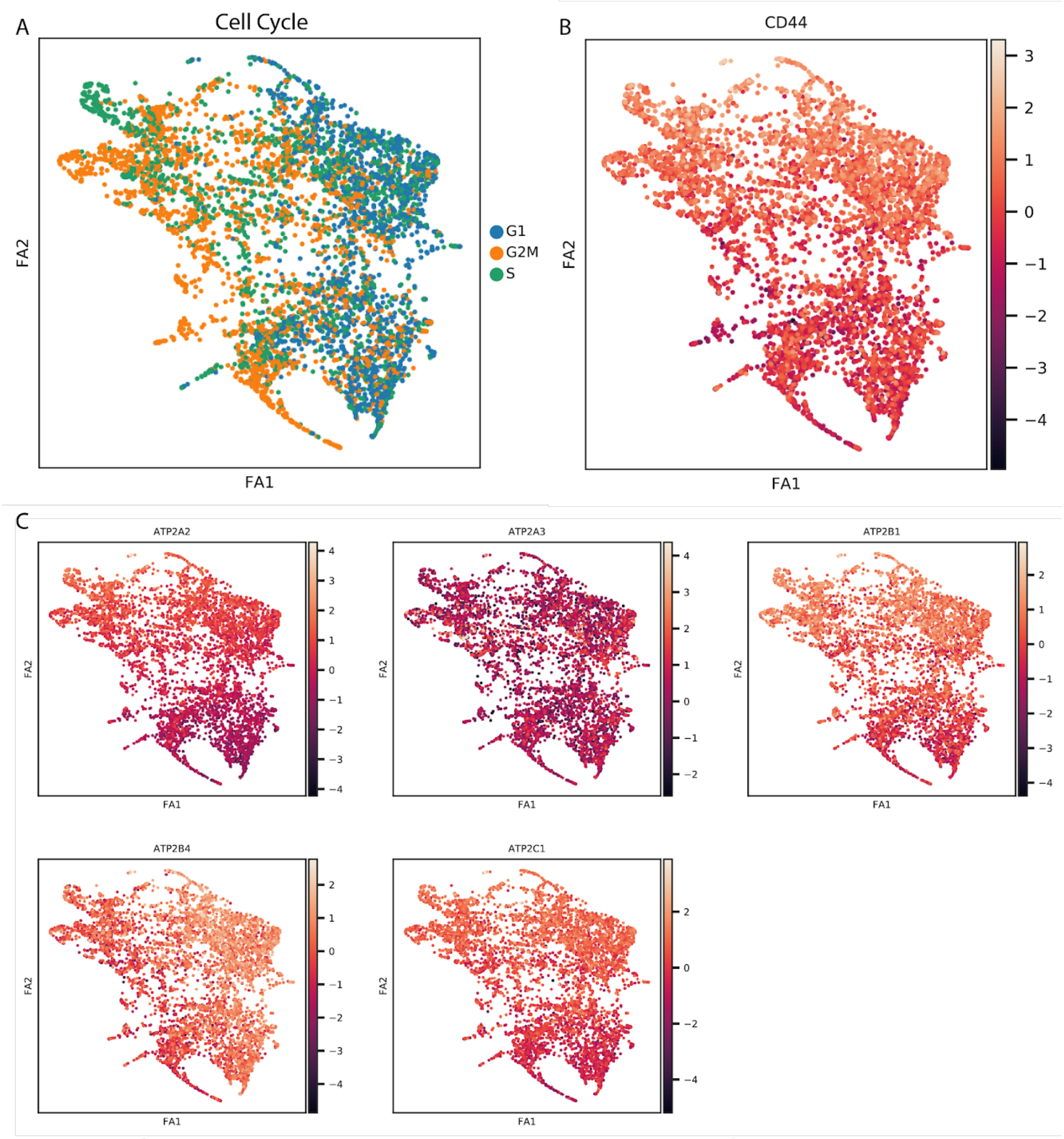


Figure 2.3 Cell state related differences in gene expression. (a) Cell cycle prediction from marker genes overlayed on force atlas projection of gene expression. (b) CD44 expression, a sub differentiation marker in MCF10a cells. (c) SERCA, PMCA, and SPCA calcium pump expression showing that differences in pump expression potentially explain much of the differences in calcium dynamics.

Next we sought to determine if differences in pump expression were consistent with predictions of their effect on calcium response. Clusters 3, 4, and 0 (Figure 2.2) have lower calcium maximum response and smaller full width at half max, FWHM. If differences in calcium pump expression are causally explaining these differences then we expect the expression of pumps to be the highest in cells with smaller CaMax and smaller FWHM. Figure 2.4A shows that on a per cluster basis there is a trend between cytoplasmic Ca^{+2} exporters and the calcium dynamical features. This indicates that differences in calcium responses likely arise from calcium signaling network differences correlated with cellular states, such as cell cycle and sub differentiation, rather than simple fluctuations in the ATP ligand

receptor.

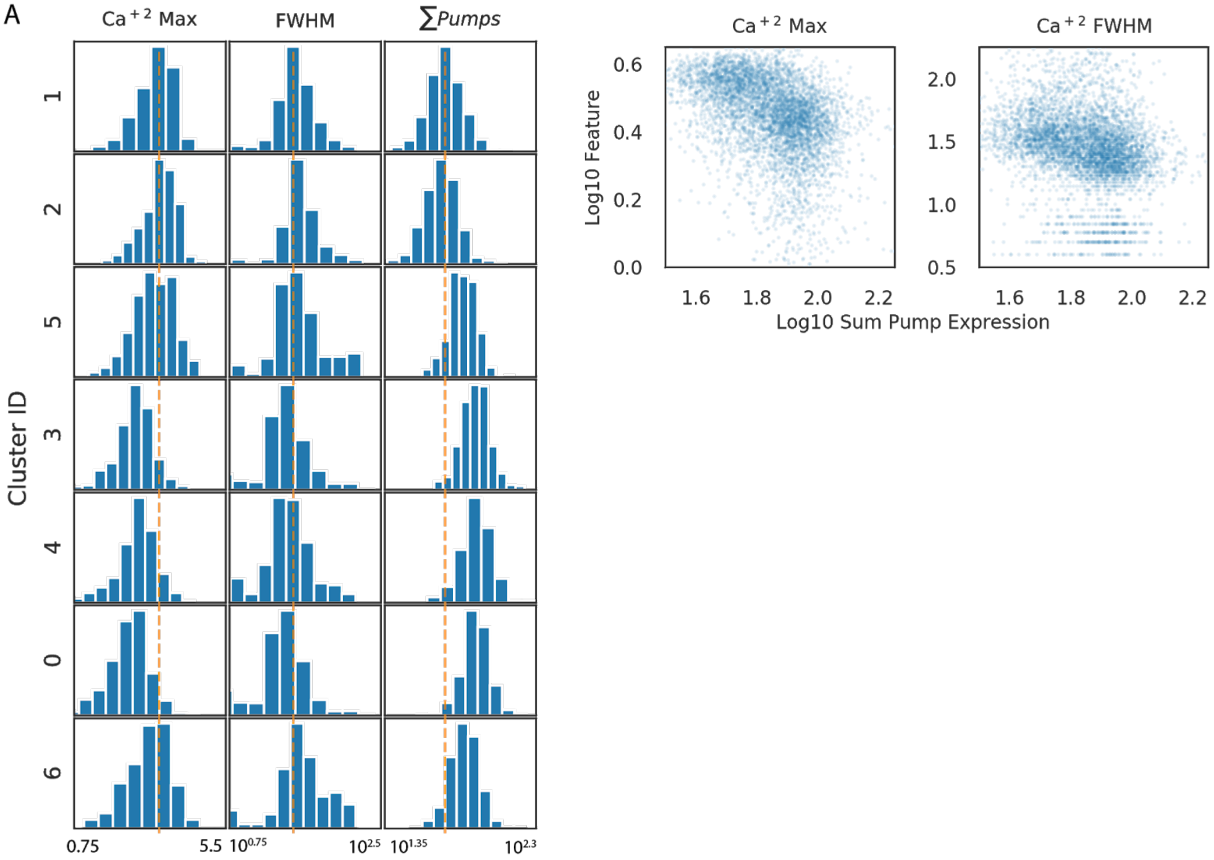


Figure 2.4 Calcium pump expression covaries with maximum free calcium and FWHM of first peak in calcium response. (left) Calcium maximum response histograms separated by cluster id. (middle) FWHM of first peak in calcium response histograms also separated by cluster. (right) the sum expression of all PMCA, SERCA, and SPCA pumps measured. These genes export calcium from the cytoplasm. (b) Left panel shows the negative correlation between the maximum calcium response and the expression of cytoplasmic calciume exporters. Right panel shows the same negative correlation for FWHM.

Interestingly, cluster 6 has high expression of cytoplasmic exporters, but the calcium max response and FWHM are more consistent with lower pump expression. We therefore examined the differences in pump expression per cluster for each pump separately (Figure 2.5A). Previous figures aggregate the pump expression across all exporters, but not all pumps necessarily have equal contribution to calcium export. Cluster 6 can be explained if ATP2A2 and ATP2A3 have a smaller effect size than ATP2C1 and ATP2B1 because cluster 6 has significant contribution of it's 'pump score' attributed to ATP2A3.

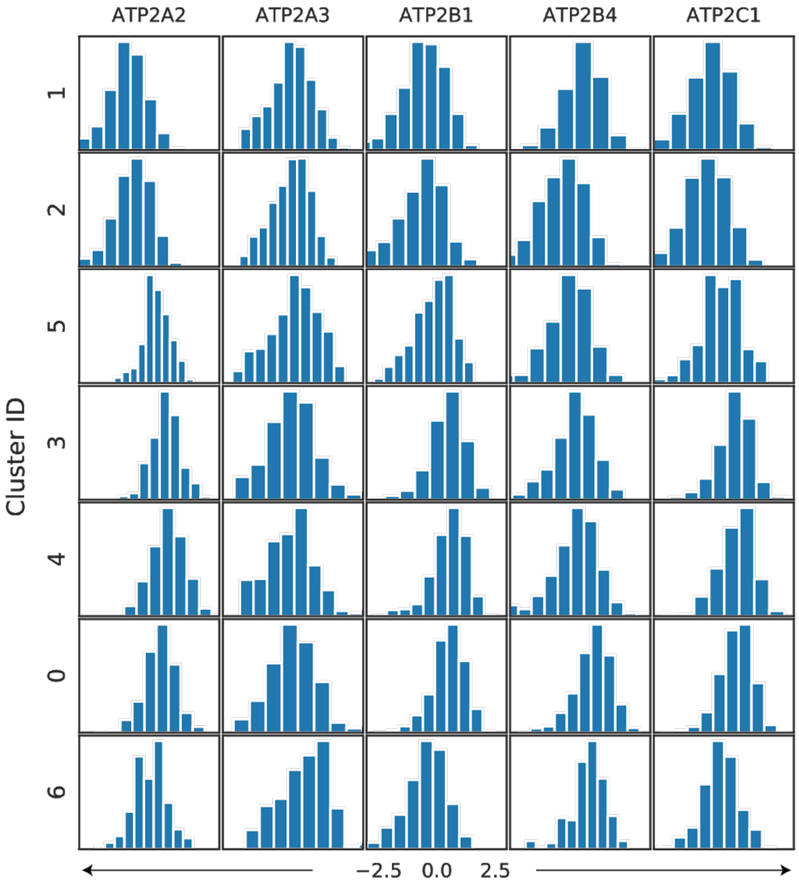


Figure 2.5 Per gene pump expression levels by cluster. Each column is the expression of an individual pump gene separated by cluster.

We are able to identify clusters in gene expression which correlate with heterogeneity observed in cellular signaling, and differences in calcium pump expression are consistent with predictions about how calcium dynamics should be different. Yet we sought to further describe how much of the variance in calcium signaling is actually explained by gene expression. In order to do this we built two different predictive models, one is very a very simple k-nearest neighbor regression prediction of calcium signaling response from gene expression, and the second is a convolutional neural network that predicts calcium signaling from gene expression. We rationalize that the simple model may sacrifice some performance in order to reduce model complexity, and the full neural network model allows us a glimpse of an upper bound on the variance explained given our data.

In the KNN regression model, we predict the latent space of calcium VAE encoding from the full gene expression vector. The output of this model is decoded by the previously trained VAE to output a full length predicted calcium trajectory. We repeat this procedure for ~1000 cells in a withheld test set, and then assess the explained variance. We measure explained variance by calculating the MSE, mean squared error, of the predicted trajectory with the measured trajectory for each of the ~1000 samples. The mean of all MSEs is normalized by the average MSE of all ~1000 samples and the average calcium trajectory of all ~1000 cells. This is the proportion of total variance not explained by the prediction and the explained variance is one minus this proportion. {actualy data from KNN model}

We also fitted a heteroencoder convolutional neural network to predict calcium trajectories from gene expression data (Figure 2.6A). This model demonstrates that gene expression predicts the intensity and dynamics of the first peak after ATP stimulation well, but not the high frequency pulses some cells exhibit (Figure 2.6BC). Using the same normalized MSE method to calculate explained variance, the heteroencoder explains 55% of the variability in calcium signaling response to ATP in MCF10A cells. We conclude that most of the variability in signaling in our MCF10a model is explained by gene expression and underlying correlated cellular states of cell cycle and MCF10a sub differentiation state.

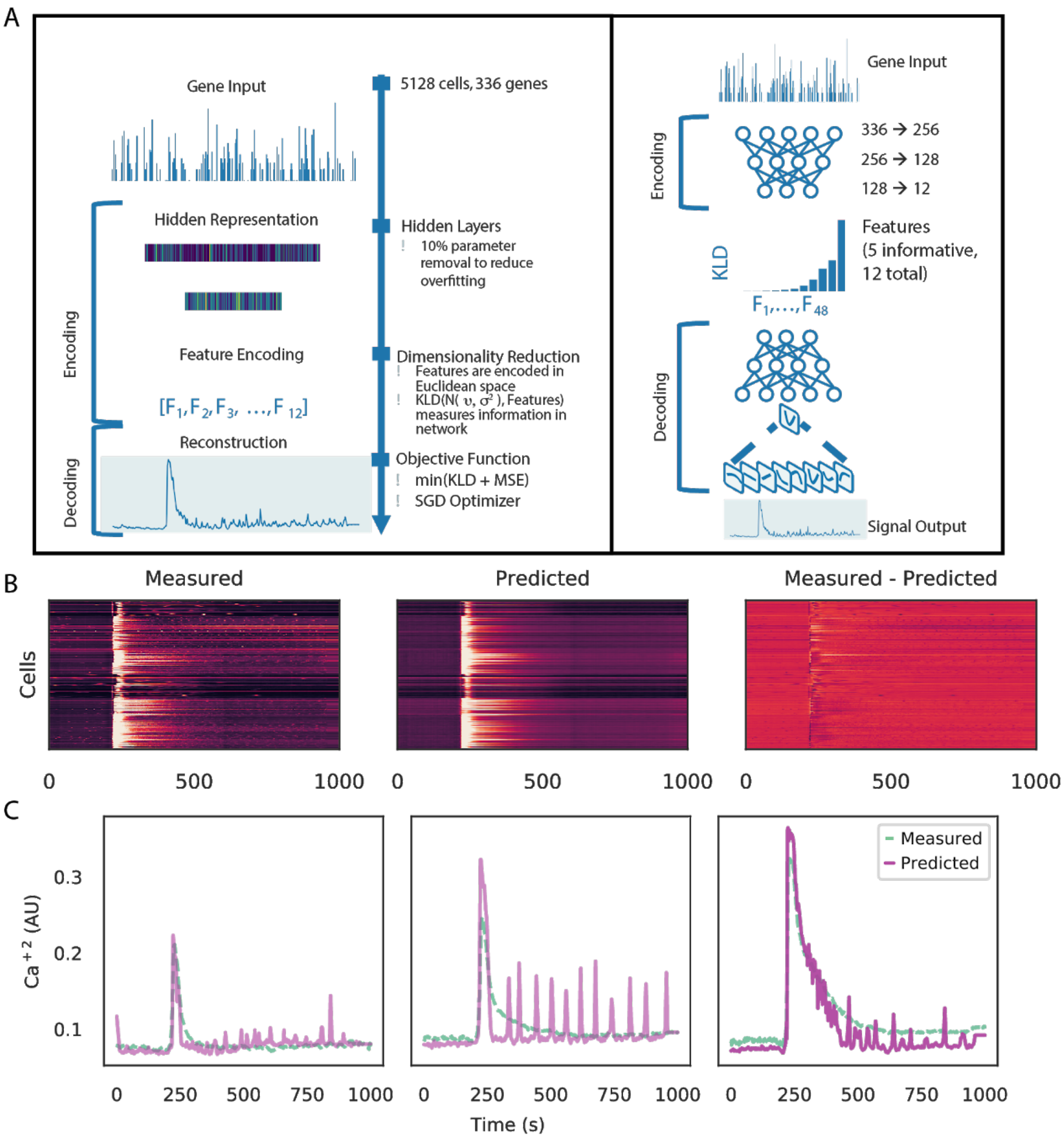


Figure 2.6 Gene expression predicts calcium response. (a) Model structure for predicting calcium trajectories from gene expression data. (b) Left shows the measured trajectories as an image plot. Middle is the reconstruction from the gene expression data of a with held test set. Right is the difference/residual of the predicted. (c) Three representative traces and their reconstructions from the test set. The model predicts the first peak well, but high frequency pulsing is not captured well by the model.

Discussion

While there is clearly a relationship between gene expression and calcium signaling heterogeneity, we were only able to explain roughly half of the overall signaling heterogeneity. There are two particular aspects of the trajectory that were not predicted well by gene expression: 1. Pulses in calcium and 2. Prestimulation basal calcium levels in some cells. It is not immediately clear whether the failure to explain these features is due to insufficient data, or their variability is rooted in post transcriptional regulation/noise.

We measured 150 expressed calcium related genes, but this is by no means an exhaustive measurement of all elements of the calcium signaling network. We could not sufficiently target some calcium related genes because the transcript length was either too short, too homologous to off-target genes, or the expression level was too high. Therefore it is reasonable that more comprehensive assays of the gene expression state could reveal that even more of the cellular heterogeneity is explained by gene expression. At the same time, it would be surprising if all of the variability was explained by gene expression. We know that fluctuations in receptor concentrations can have large impacts on signaling response, and that many of these fluctuations come from protein expression noise rather than variance in the number of transcripts(Brock and Jovin, 2001; Cheong et al., 2011).

The failure to capture pulses could be a limitation of the models explored. In our model, if we cannot predict the timing of pulses with high accuracy then the MSE score is severely affected by an out of phase pulse prediction. Therefore, if gene expression is predictive of the propensity of cells to pulse at different frequencies, but the timing is somewhat stochastic then our model will generally favor not predicting any pulses. A future direction could be predicting a feature based representation of the pulses, or reconsideration of an error function that is more robust to stochastic timed pulses.

This work can be also be extended by recent advances in genome manipulation using CRISPRi and CRISPRa(Dixit et al., 2016) in order to more comprehensively investigate both how much of the variability is related to gene expression, and how different genes specifically impact calcium signaling. Spatial transcriptomics combined with these pooled genome manipulation techniques allows investigation of the genotype to phenotype mapping in signaling pathways on an entirely new scale, and it will be very interesting to see how their future use develops.

Acknowledgements

Chapter 2 is being prepared for submission for the publication of the material. The authors are Robert Foreman, Evan Maltz, and Roy Wollman. The dissertation author was the primary investigator and author of this material. Evan Maltz performed the variational autoencoder models.

Methods

Cell Culture and Experiments

Experiments were performed as described in Chapter 1 methods.

Gene Expression Analysis

Raw transcript counts were volume normalized as described in Chapter 1. Genes were filtered to exclude any genes that were expressed at less than 2.65 transcripts/cell in the 95th percentile of cells. This gene expression matrix was log1p transformed and used for clustering and embedding. Clusters were generated using the Phenograph package(DiGiuseppe et al., 2018) with k=13 and with z-scaled gene expression counts. Gene embeddings were performed with the SCANPY package by creating a diffusion map from the neighborhood graph of cells (default parameters), and lower dimensional viewings were generated using force atlas projection.

Variational Autoencoder Calcium Embedding

The purpose of the variational autoencoder is to generate accurate reconstructions of calcium signaling, measure the available information content, and produce a continuous, minimal encoding. Autoencoders are generally structured as a reciprocal pair of networks: the encoder and the decoder. These networks cooperate to learn minimal representations (i.e. latent vectors) of each example in the dataset during the training phase. Training is performed by showing the encoder an example from the training set, which it attempts to encode in a significantly reduced space, then the decoder is given the encoding and attempts to reconstruct the original example. To optimize learning, the raw signaling data was preprocessed by normalizing each cell by its GCaMP abundance, then scaling all values between 0 and 1 while preserving each point's relative value. To prevent and measure overfitting, 10% of the total dataset was withheld from training and used only to evaluate model performance. Additionally, 10% of the weights in the network are randomly removed during each training iteration to further decrease overfitting. Architectures were trained and evaluated based on the sum of squared errors and KL divergence between the latent vectors and normal distributions. The inverse sum of the terms of the objective function defines the evidence lower bound (ELBO), or the marginal likelihood of the data, which is maximized during training to approximate the reconstruction distribution. Training was performed using stochastic gradient descent with a learning rate schedule based on step decay to gradually refine parameters during training. KL annealing, the process of slowly increasing the weight of the KL divergence, was also tested but did not significantly improve end results. Because the mean KL divergence of each latent vector is used, varying the number of latent vectors did not affect learning so long as there were at least the minimal number of informative vectors. During each round of training, a small amount of normally distributed noise is added to each example with mean 0 and standard deviation of 0.002 to act as a high pass filter for technical noise.

Using this objective function, several architectures were tested and evaluated by qualitatively comparing reconstructions and quantitatively plotting the error. The networks tested ranged from a simple feed-forward network with minimal units to fully connected convolutional models with adversarial networks. The chosen model is depicted in Figure 2.1D, representing the simplest model tested that was able to accurately reconstruct the signals and produce minimal error.

Variational Heteroencoder (Gene to Calcium Model)

Like the variational autoencoder, the variational heteroencoder was designed to accurately reproduce calcium signaling patterns using a continuous encoding; however, the input data is derived from gene expression. Thus, the goal of the heteroencoder is to map gene expression to signaling using the same learning method as described above (e.g. learning rate schedule, etc.). The decoder used here has the same architecture as the autoencoder above, though using a different encoder. The encoder was copied from another variational autoencoder trained explicitly on single-cell gene expression data. Preprocessing included taking the log pseudocount, then scaling each gene across all cells to between 0 and 1. Several architectures were tested for this autoencoder, the encoder network architecture from the best model was used as the encoder network for the heteroencoder shown in Figure 2.7A. The gene encoder and signaling decoder ensemble was the only architecture tested. Again, KL annealing was tested and did not improve visual reconstruction accuracy or produce a higher ELBO. Three variations were tested, learning new weights end to end, using pre-trained weights for the encoder, and pre-trained decoder. Pre-training failed to improve accuracy, so the final model used was the end to end trained model.

Explained Variance of Calcium Prediction

Calcium trajectories were predicted using a convolutional neural network described above. We used mean squared error as a goodness of fit metric to determine how much of the variance in calcium trajectories was unexplained by gene expression. This value was normalized by the MSE of all trajectories with the population average trajectory to yield a proportion of variance unexplained by the prediction model. Explained variance was then considered 1-UnexplainedVariance.

REFERENCES

Akerboom, J., Chen, T.-W., Wardill, T.J., Tian, L., Marvin, J.S., Mutlu, S., Calderón, N.C., Esposti, F., Borghuis, B.G., Sun, X.R., Gordus, A., Orger, M.B., Portugues, R., Engert, F., Macklin, J.J., Filosa, A., Aggarwal, A., Kerr, R.A., Takagi, R., Kracun, S., Shigetomi, E., Khakh, B.S., Baier, H., Lagnado, L., Wang, S.S.-H., Bargmann, C.I., Kimmel, B.E., Jayaraman, V., Svoboda, K., Kim, D.S., Schreiter, E.R., Looger, L.L., 2012. Optimization of a GCaMP calcium indicator for neural activity imaging. J. Neurosci. 32, 13819–13840.

Andrews, T.S., Hemberg, M., 2018. Identifying cell populations with scRNASeq. Mol. Aspects Med. 59, 114–122.

Bandara, S., Malmersjö, S., Meyer, T., 2013. Regulators of calcium homeostasis identified by inference of kinetic model parameters from live single cells perturbed by siRNA. Sci. Signal. 6, ra56.

Battich, N., Stoeger, T., Pelkmans, L., 2015. Control of Transcript Variability in Single Mammalian Cells. Cell 163, 1596–1610.

Blanchini, F., Franco, E., 2011. Structurally robust biological networks. BMC Syst. Biol. 5, 74.

Brock, R., Jovin, T.M., 2001. Heterogeneity of signal transduction at the subcellular level: microsphere-based focal EGF receptor activation and stimulation of Shc translocation. J. Cell Sci. 114, 2437–2447.

Cai, L., 2013. Turning single cells into microarrays by super-resolution barcoding. Brief. Funct. Genomics 12, 75–80.

Cheng, S., Pei, Y., He, L., Peng, G., Reinius, B., Tam, P.P.L., Jing, N., Deng, Q., 2019. Single-Cell RNA-Seq Reveals Cellular Heterogeneity of Pluripotency Transition and X Chromosome Dynamics during Early Mouse Development. Cell Rep. 26, 2593–2607.e3.

Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., Zhuang, X., 2015. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 348, aaa6090.

Cheong, R., Rhee, A., Wang, C.J., Nemenman, I., Levchenko, A., 2011. Information Transduction Capacity of Noisy Biochemical Signaling Networks. Science.

Choubey, S., Kondev, J., Sanchez, A., 2015. Deciphering Transcriptional Dynamics In Vivo by Counting Nascent RNA Molecules. PLoS Comput. Biol. 11, e1004345.

Codeluppi, S., Borm, L.E., Zeisel, A., La Manno, G., van Lunteren, J.A., Svensson, C.I., Linnarsson, S., 2018. Spatial organization of the somatosensory cortex revealed by osmFISH. Nat. Methods 15, 932–935.

Corrigan, A.M., Tunnacliffe, E., Cannon, D., Chubb, J.R., 2016. A continuum model of transcriptional bursting. Elife 5.

Dar, R.D., Razooky, B.S., Singh, A., Trimeloni, T.V., McCollum, J.M., Cox, C.D., Simpson, M.L., Weinberger, L.S., 2012. Transcriptional burst frequency and burst size are equally modulated across the human genome. Proc. Natl. Acad. Sci. U. S. A. 109, 17454–17459.

Dar, R.D., Razooky, B.S., Weinberger, L.S., Cox, C.D., Simpson, M.L., 2015. The Low Noise Limit in Gene Expression. PLoS One 10, e0140969.

Dey, S.S., Foley, J.E., Limsirichai, P., Schaffer, D.V., Arkin, A.P., 2015. Orthogonal control of expression mean and variance by epigenetic features at different genomic loci. Mol. Syst. Biol. 11, 806.

DiGiuseppe, J.A., Cardinali, J.L., Rezuke, W.N., Pe'er, D., 2018. PhenoGraph and viSNE facilitate the identification of abnormal T-cell populations in routine clinical flow cytometric data. Cytometry Part B: Clinical Cytometry.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T.M., Lander, E.S., Weissman, J.S., Friedman, N., Regev, A., 2016. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell 167, 1853–1866.e17.

Doersch, C., 2016. Tutorial on Variational Autoencoders. arXiv [stat.ML].

Dueck, H., Eberwine, J., Kim, J., 2016. Variation is function: Are single cell differences functionally important?: Testing the hypothesis that single cell variation is required for aggregate function. Bioessays 38, 172–180.

Eldar, A., Elowitz, M.B., 2010. Functional roles for noise in genetic circuits. Nature 467, 167–173.

Elowitz, M.B., 2002. Stochastic Gene Expression in a Single Cell. Science.

Eng, C.-H.L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.-C., Cai, L., 2019. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. Nature.

Ferguson, M.L., Larson, D.R., 2013. Measuring transcription dynamics in living cells using fluctuation analysis. Methods Mol. Biol. 1042, 47–60.

Foreman, R., Wollman, R., n.d. Mammalian gene expression variability is explained by underlying cell state.

Friedman, N., Cai, L., Xie, X.S., 2006. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. Phys. Rev. Lett. 97, 168302.

Fritzsch, C., Baumgärtner, S., Kuban, M., Steinshorn, D., Reid, G., Legewie, S., 2018. Estrogen-dependent control and cell-to-cell variability of transcriptional bursting. Mol. Syst. Biol. 14, e7678.

Fukaya, T., Lim, B., Levine, M., 2016. Enhancer Control of Transcriptional Bursting. Cell 166, 358–368.

Funaki, T., Matsuda, A., Ebihara, N., Murakami, A., Chauhan, S., Jurkunas, U.V., Dana, R., 2011. Atp Contributes Corneal Endothelial Wound Healing Via P2x7 Receptor. Invest. Ophthalmol. Vis. Sci. 52, 6436–6436.

Giorgi, C., Danese, A., Missiroli, S., Patergnani, S., Pinton, P., 2018. Calcium Dynamics as a Machine for Decoding Signals. Trends Cell Biol. 28, 258–273.

Gong, H., Do, D., Ramakrishnan, R., 2018. Single-Cell mRNA-Seq Using the Fluidigm C1 System and Integrated Fluidics Circuits. Methods Mol. Biol. 1783, 193–207.

Handly, L.N., Pilko, A., Wollman, R., 2015. Paracrine communication maximizes cellular response fidelity in wound signaling. Elife 4, e09652.

Handly, L.N., Wollman, R., 2017. Wound-induced Ca2+ wave propagates through a simple release and diffusion mechanism. Mol. Biol. Cell 28, 1457–1466.

Hansen, C.H., van Oudenaarden, A., 2013. Allele-specific detection of single mRNA molecules in situ. Nat. Methods 10, 869–871.

Hansen, M.M.K., Desai, R.V., Simpson, M.L., Weinberger, L.S., 2018a. Cytoplasmic Amplification of Transcriptional Noise Generates Substantial Cell-to-Cell Variability. Cell Syst 7, 384–397.e6.

Hansen, M.M.K., Wen, W.Y., Ingerman, E., Razooky, B.S., Thompson, C.E., Dar, R.D., Chin, C.W., Simpson, M.L., Weinberger, L.S., 2018b. A Post-Transcriptional Feedback Mechanism for Noise Suppression and Fate Stabilization. Cell 173, 1609–1621.e15.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., Fang, Q., Wang, Y., Yue, R., Li, T., Huang, H., Orkin, S.H., Yuan, G.-C., Chen, M., Guo, G., 2018. Mapping the Mouse Cell Atlas by Microwell-Seq. Cell 173, 1307.

Haque, A., Engel, J., Teichmann, S.A., Lönnberg, T., 2017. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 9, 75.

Jangi, M., Sharp, P.A., 2014. Building robust transcriptomes with master splicing factors. Cell 159, 487–498.

Justet, C., Chifflet, S., Hernandez, J.A., 2019. Calcium Oscillatory Behavior and Its Possible Role during Wound Healing in Bovine Corneal Endothelial Cells in Culture. Biomed Res. Int. 2019, 8647121.

Kaern, M., Elston, T.C., Blake, W.J., Collins, J.J., 2005. Stochasticity in gene expression: from theories to phenotypes. Nat. Rev. Genet. 6, 451–464.

Kanehisa, M., Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., Tanabe, M., 2019. New approach for understanding genome variations in KEGG. Nucleic Acids Res. 47, D590–D595.

Kepler, T.B., Elston, T.C., 2001. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. Biophys. J. 81, 3116–3136.

Larsson, A.J.M., Johnsson, P., Hagemann-Jensen, M., Hartmanis, L., Faridani, O.R., Reinius, B., Segerstolpe, Å., Rivera, C.M., Ren, B., Sandberg, R., 2019. Genomic encoding of transcriptional burst kinetics. Nature 565, 251–254.

Lenstra, T.L., Rodriguez, J., Chen, H., Larson, D.R., 2016. Transcription Dynamics in Living Cells. Annu. Rev. Biophys. 45, 25–47.

Lubeck, E., Cai, L., 2012. Single-cell systems biology by super-resolution imaging and combinatorial labeling. Nat. Methods 9, 743–748.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., Trombetta, J.J., Weitz, D.A., Sanes, J.R., Shalek, A.K., Regev, A., McCarroll, S.A., 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214.

Minns, M.S., Trinkaus-Randall, V., 2016. Purinergic Signaling in Corneal Wound Healing: A Tale of 2 Receptors. J. Ocul. Pharmacol. Ther. 32, 498–503.

Moffitt, J.R., Hao, J., Bambah-Mukku, D., Lu, T., Dulac, C., Zhuang, X., 2016. High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. Proc. Natl. Acad. Sci. U. S. A. 113, 14456–14461.

Molina, N., Suter, D.M., Cannavo, R., Zoller, B., Gotic, I., Naef, F., 2013. Stimulus-induced modulation of transcriptional bursting in a single mammalian gene. Proc. Natl. Acad. Sci. U. S. A. 110, 20563–20568.

Moon, K.R., Stanley, J.S., Burkhardt, D., van Dijk, D., Wolf, G., Krishnaswamy, S., 2018. Manifold learning-based methods for analyzing single-cell RNA-sequencing data. Current Opinion in Systems Biology 7, 36–46.

Munsky, B., Neuert, G., van Oudenaarden, A., 2012. Using gene expression noise to understand gene regulation. Science 336, 183–187.

Muramoto, T., Müller, I., Thomas, G., Melvin, A., Chubb, J.R., 2010. Methylation of H3K4 Is required for inheritance of active transcriptional states. Curr. Biol. 20, 397–406.

Nguyen, A., Khoo, W.H., Moran, I., Croucher, P.I., Phan, T.G., 2018. Single Cell RNA Sequencing of Rare Immune Cell Populations. Front. Immunol. 9, 1553.

Nicolas, D., Zoller, B., Suter, D.M., Naef, F., 2018. Modulation of transcriptional burst frequency by histone acetylation. Proc. Natl. Acad. Sci. U. S. A. 115, 7153–7158.

Noren, D.P., Chou, W.H., Lee, S.H., Qutub, A.A., Warmflash, A., Wagner, D.S., Popel, A.S., Levchenko, A., 2016. Endothelial cells decode VEGF-mediated Ca2+ signaling patterns to produce distinct functional responses. Sci. Signal. 9, ra20.

Ochiai, H., Sugawara, T., Sakuma, T., Yamamoto, T., 2014. Stochastic promoter activation affects Nanog expression variability in mouse embryonic stem cells. Sci. Rep. 4, 7125.

Padovan-Merhar, O., Nair, G.P., Biaesch, A.G., Mayer, A., Scarfone, S., Foley, S.W., Wu, A.R., Churchman, L.S., Singh, A., Raj, A., 2015. Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. Mol. Cell 58, 339–352.

Paulsson, J., 2004. Summing up the noise in gene networks. Nature 427, 415–418.

Paulsson, J., 2005. Models of stochastic gene expression. Physics of Life Reviews.

Peccoud, J., Ycart, B., 1995. Markovian Modeling of Gene-Product Synthesis. Theor. Popul. Biol. 48, 222–234.

Pedraza, J.M., van Oudenaarden, A., 2005. Noise propagation in gene networks. Science 307, 1965–1969.

Qu, Y., Han, B., Yu, Y., Yao, W., Bose, S., Karlan, B.Y., Giuliano, A.E., Cui, X., 2015. Evaluation of MCF10A as a Reliable Model for Normal Human Mammary Epithelial Cells. PLoS One 10, e0131285.

Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., Tyagi, S., 2006. Stochastic mRNA synthesis in mammalian cells. PLoS Biol. 4, e309.

Raj, A., van Oudenaarden, A., 2008. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell 135, 216–226.

Raser, J.M., O'Shea, E.K., 2004. Control of stochasticity in eukaryotic gene expression. Science 304, 1811–1814.

Rodriguez, J., Ren, G., Day, C.R., Zhao, K., Chow, C.C., Larson, D.R., 2019. Intrinsic Dynamics of a Human Gene Reveal the Basis of Expression Heterogeneity. Cell 176, 213–226.e18.

Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S., Chen, W., Pun, S.H., Sellers, D.L., Tasic, B., Seelig, G., 2018. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science 360, 176–182.

Rouzine, I.M., Weinberger, A.D., Weinberger, L.S., 2015. An evolutionary role for HIV latency in enhancing viral transmission. Cell 160, 1002–1012.

Sanchez, A., Golding, I., 2013. Genetic determinants and cellular constraints in noisy gene expression. Science 342, 1188–1193.

Schmiedel, J.M., Klemm, S.L., Zheng, Y., Sahay, A., Blüthgen, N., Marks, D.S., van Oudenaarden, A., 2015. Gene expression. MicroRNA control of protein expression noise. Science 348, 128–132.

Selimkhanov, J., Taylor, B., Yao, J., Pilko, A., Albeck, J., Hoffmann, A., Tsimring, L., Wollman, R., 2014. Accurate information transmission through dynamic biochemical signaling networks. Science 346, 1370–1373.

Shahrezaei, V., Ollivier, J.F., Swain, P.S., 2008. Colored extrinsic fluctuations and stochastic gene expression. Mol. Syst. Biol. 4, 196.

Shahrezaei, V., Swain, P.S., 2008. Analytical distributions for stochastic gene expression. Proc. Natl. Acad. Sci. U. S. A. 105, 17256–17261.

Shah, S., Takei, Y., Zhou, W., Lubeck, E., Yun, J., Eng, C.-H.L., Koulena, N., Cronin, C., Karp, C., Liaw, E.J., Amin, M., Cai, L., 2018. Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH. Cell 174, 363–376.e16.

Sherman, M.S., Lorenz, K., Hunter Lanier, M., Cohen, B.A., 2015. Cell-to-Cell Variability in the Propensity to Transcribe Explains Correlated Fluctuations in Gene Expression. Cell Syst 1, 315–325.

Shoval, O., Sheftel, H., Shinar, G., Hart, Y., Ramote, O., Mayo, A., Dekel, E., Kavanagh, K., Alon, U., 2012. Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. Science 336, 1157–1160.

Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Liron, Y., Rosenfeld, N., Danon, T., Perzov, N., Alon, U., 2006. Variability and memory of protein levels in human cells. Nature 444, 643–646.

Singh, A., Razooky, B., Cox, C.D., Simpson, M.L., Weinberger, L.S., 2010. Transcriptional bursting from the HIV-1 promoter is a significant source of stochastic noise in HIV-1 gene expression. Biophys. J. 98, L32–4.

Singh, A., Razooky, B.S., Dar, R.D., Weinberger, L.S., 2012. Dynamics of protein noise can distinguish between alternate sources of gene-expression variability. Mol. Syst. Biol. 8, 607.

Skinner, S.O., Xu, H., Nagarkar-Jaiswal, S., Freire, P.R., Zwaka, T.P., Golding, I., 2016. Single-cell analysis of transcription kinetics across the cell cycle. Elife 5, e12175.

Skupsky, R., Burnett, J.C., Foley, J.E., Schaffer, D.V., Arkin, A.P., 2010. HIV Promoter Integration Site Primarily Modulates Transcriptional Burst Size Rather Than Frequency. PLoS Comput. Biol. 6, e1000952.

Smedler, E., Uhlén, P., 2014. Frequency decoding of calcium oscillations. Biochim. Biophys. Acta 1840, 964–969.

Spencer, S.L., Gaudet, S., Albeck, J.G., Burke, J.M., Sorger, P.K., 2009. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. Nature 459, 428–432.

Stewart-Ornstein, J., Weissman, J.S., El-Samad, H., 2012. Cellular noise regulons underlie fluctuations in Saccharomyces cerevisiae. Mol. Cell 45, 483–493.

Stoeger, T., Battich, N., Pelkmans, L., 2016. Passive Noise Filtering by Cellular Compartmentalization. Cell 164, 1151–1161.

Strebinger, D., Friman, E.T., Deluz, C., Govindan, S., Alber, A.B., Suter, D.M., 2018. Endogenous fluctuations of OCT4 and SOX2 bias pluripotent cell fate decisions. bioRxiv.

Suo, S., Zhu, Q., Saadatpour, A., Fei, L., Guo, G., Yuan, G.-C., 2018. Revealing the Critical Regulators of Cell Identity in the Mouse Cell Atlas. Cell Rep. 25, 1436–1445.e3.

Suter, D.M., Molina, N., Gatfield, D., Schneider, K., Schibler, U., Naef, F., 2011. Mammalian genes are transcribed with widely different bursting kinetics. Science 332, 472–474.

Symmons, O., Raj, A., 2016. What's Luck Got to Do with It: Single Cells, Multiple Fates, and Biological Nondeterminism. Mol. Cell 62, 788–802.

Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, Principal investigators, 2018. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature 562, 367–372.

Tanaka, G., Morino, K., Aihara, K., 2015. Dynamical Robustness of Complex Biological Networks. In: Ohira, T., Uzawa, T. (Eds.), Mathematical Approaches to Biological Systems: Networks, Oscillations, and Collective Motions. Springer Japan, Tokyo, pp. 29–53.

Tantale, K., Mueller, F., Kozulic-Pirher, A., Lesne, A., Victor, J.-M., Robert, M.-C., Capozi, S., Chouaib, R., Bäcker, V., Mateos-Langerak, J., Darzacq, X., Zimmer, C., Basyuk, E., Bertrand, E., 2016. A single-molecule view of transcription reveals convoys of RNA polymerases and multi-scale bursting. Nat. Commun. 7, 12248.

Taylor, M.S., Francis, M., 2014. Decoding dynamic Ca2+ signaling in the vascular endothelium. Front. Physiol. 5, 332.

Thattai, M., van Oudenaarden, A., 2004. Stochastic gene expression in fluctuating environments. Genetics 167, 523–530.

Toettcher, J.E., Weiner, O.D., Lim, W.A., 2013. Using optogenetics to interrogate the dynamic control of signal transmission by the Ras/Erk module. Cell 155, 1422–1434.

Trapnell, C., 2015. Defining cell types and states with single-cell genomics. Genome Res. 25, 1491–1498.

Veening, J.-W., Stewart, E.J., Berngruber, T.W., Taddei, F., Kuipers, O.P., Hamoen, L.W., 2008. Bet-hedging and epigenetic inheritance in bacterial cell development. Proc. Natl. Acad. Sci. U. S. A. 105, 4393–4398.

Vega, N.M., Gore, J., 2014. Collective antibiotic resistance: mechanisms and implications. Curr. Opin. Microbiol. 21, 28–34.

Wagner, A., Regev, A., Yosef, N., 2016. Revealing the vectors of cellular identity with single-cell genomics. Nat. Biotechnol. 34, 1145–1160.

Wang, G., Moffitt, J.R., Zhuang, X., 2018. Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. Sci. Rep. 8, 4847.

Whitacre, J.M., 2012. Biological robustness: paradigms, mechanisms, and systems principles. Front. Genet. 3, 67.

Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O., Botstein, D., 2002. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol. Biol. Cell 13, 1977–2000.

Xin, Y., Kim, J., Ni, M., Wei, Y., Okamoto, H., Lee, J., Adler, C., Cavino, K., Murphy, A.J., Yancopoulos, G.D., Lin, H.C., Gromada, J., 2016. Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells. Proc. Natl. Acad. Sci. U. S. A. 113, 3293–3298.

Yao, J., Pilko, A., Wollman, R., 2016. Distinct cellular states determine calcium signaling response. Mol. Syst. Biol. 12, 894.

Zhang, H., Liu, J., Sun, S., Pchitskaya, E., Popugaeva, E., Bezprozvanny, I., 2015. Calcium signaling, excitability, and synaptic plasticity defects in a mouse model of Alzheimer's disease. J. Alzheimers. Dis. 45, 561–580.

Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., Huang, Y., Wang, J., 2019. Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. Mol. Cell 73, 130–142.e5.

Zilionis, R., Nainys, J., Veres, A., Savova, V., Zemmour, D., Klein, A.M., Mazutis, L., 2017. Single-cell barcoding and sequencing using droplet microfluidics. Nat. Protoc. 12, 44–73.

Zoller, B., Nicolas, D., Molina, N., Naef, F., 2015. Structure of silent transcription intervals and noise characteristics of mammalian genes. Mol. Syst. Biol. 11, 823.