UCLA
UCLA Previously Published Works

Title

Quantifying the phenotypic information in mRNA abundance

Permalink

Journal

Molecular Systems Biology, 18(8)

ISSN

1744-4292

Authors

Maltz, Evan
Wollman, Roy

Publication Date

2022-08-01

DOI

10.15252/msb.202211001

Copyright Information

Peer reviewed

Article

molecular
systems
biology

Quantifying the phenotypic information in mRNA abundance

Evan Maltz[1,2] & Roy Wollman[1,2,3,*] ○iD

Abstract

Quantifying the dependency between mRNA abundance and downstream cellular phenotypes is a fundamental open problem in biology. Advances in multimodal single-cell measurement technologies provide an opportunity to apply new computational frameworks to dissect the contribution of individual genes and gene combinations to a given phenotype. Using an information theory approach, we analyzed multimodal data of the expression of 83 genes in the Ca^{2+} signaling network and the dynamic Ca^{2+} response in the same cell. We found that the overall expression levels of these 83 genes explain approximately 60% of Ca^{2+} signal entropy. The average contribution of each single gene was 17%, revealing a large degree of redundancy between genes. Using different heuristics, we estimated the dependency between the size of a gene set and its information content, revealing that on average, a set of 53 genes contains 54% of the information about Ca^{2+} signaling. Our results provide the first direct quantification of information content about complex cellular phenotype that exists in mRNA abundance measurements.

Introduction

Cellular phenotypes emerge from many regulated interactions between various components. Rates of synthesis and degradation determine the instantaneous abundances of different biological molecules. These kinetic rates are themselves a property of regulatory interactions between biomolecules creating multilayered feedback networks (El-Samad, 2021). Both the dynamic and instantaneous abundances of biomolecules are key determinants of cellular phenotypes, underlying their ability to make different decisions given the same stimulus (Perkins & Swain, 2009; Cheong *et al*,

2011; Purvis & Lahav, 2013). The ability to systematically measure the abundance of large sets of different biomolecules such as mRNA and proteins enables the determination of regulatory strengths across different nodes of these complex networks. Pioneering work in *Escherichia coli* based on instantaneous single-cell measurements of mRNA and protein copy numbers reveals a surprisingly low correlation coefficient of $r = 0.01 \pm 0.03$ across 129 highly expressed genes (Taniguchi *et al*, 2010). The lack of correlation between mRNA and protein in *E. coli* might be due to their small size and magnitude of temporal fluctuation in mRNA levels. However, more recent advances in multimodal assays in mammalian cells also identified low correlations between the abundances of most proteins and corresponding mRNAs (Darmanis *et al*, 2016; Gong *et al*, 2017; Stoeckius *et al*, 2017; Schulz *et al*, 2018; Mair *et al*, 2020). This low correlation appears to contradict intuition that protein and mRNA levels should strongly correspond within cells because of the dependency suggested by the central dogma (Liu *et al*, 2016). An alternative hypothesis is that the majority of regulatory steps and phenotypically relevant information lie in posttranscriptional processes. Posttranscriptional regulation can modulate both protein activity and abundance via protein interactions, posttranslational modifications, RNA interactions/ structure, and more. Stochastic processes also obscure the importance of molecular composition to phenotypic outcomes (Perkins & Swain, 2009; Balázsi *et al*, 2011; Cheong *et al*, 2011). Yet, many studies have pointed to differences in mRNA levels among clonal cells to explain differences in cellular phenotypes (Shaffer *et al*, 2020; Emert *et al*, 2021). These observations highlight a need for a better framework to address fundamental questions: Does mRNA abundance matter? What fraction of the information about cellular phenotype is determined by mRNA abundance, and what fraction is due to posttranscriptional regulation?

Quantifying the information content in mRNA abundance about cellular phenotypes is technically and computationally challenging due to the many layers of complex interactions in cellular networks (Macaulay *et al*, 2017). Phenotypic information displayed in clonal cells is controlled by molecular composition, stochastic factors, intermediate regulation, and crosstalk (Azeloglu & Iyengar, 2015). Many approaches have been developed to disentangle these complex and distributed dependencies. Feature engineering has been one powerful tool to reveal interpretable characteristics of signaling

---

1 Department of Chemistry and Biochemistry, UCLA, Los Angeles, CA, USA
2 Institute of Quantitative and Computational Bioscience, UCLA, Los Angeles, CA, USA
3 Department of Integrative Biology and Physiology, UCLA, Los Angeles, CA, USA
   *Corresponding author. Tel: 310 8254321; E-mail: rwollman@ucla.edu

dynamics, finding multiple motifs that encode information about stimulus dose and type (Nelson *et al*, 2004; Hafner *et al*, 2017; Zhang *et al*, 2017; Wong *et al*, 2019; Adelaja *et al*, 2021). However, these features do not capture all the information in complex dynamics, which are difficult to study and fully recapitulate in mechanistic models (Myers *et al*, 2021). Another common approach is to perform dimensionality reduction and/or clustering to integrate different modalities (Subramanian *et al*, 2020; Kinnunen *et al*, 2021). Several studies have clustered groups of genes or cells based on signal patterns to reveal general mechanisms of how signaling dynamics affect transcription (Hafner *et al*, 2017; Lane *et al*, 2017). However, it is still unknown how differences in arbitrary sets of transcripts relate to dynamic signals in the same cells. Signaling phenomena may emerge due to differences among many combinations of genes, which may be missed when simplified to either individual genes or gene clusters. Single-cell network states are notoriously difficult to fully measure, and insights into the relationships between many components require high-dimensional and multimodal data from the same cells (Spencer *et al*, 2009; Azeloglu & Iyengar, 2015; Macaulay *et al*, 2017; Adelaja *et al*, 2021). Although useful in many contexts, feature engineering, clustering, and dimensionality reduction are not guaranteed to capture all useful information.

Directly quantifying the relationship between many transcripts and a phenotype via an information theoretic approach can provide a direct measure of the importance of mRNA abundance. However, three challenges prevent the general use of information theory in quantifying information content in RNA abundance. (i) Biological feedbacks entangle mRNA abundance and cellular phenotype. Cellular phenotypes that emerge over long timescale, for example, cellular differentiations, have a longer timescale than the lifetime of mRNA molecules that potentially determine the emerging phenotypes. In these cases, mRNA abundances themselves change dynamically adding additional complexities. (ii) Quantification of importance of mRNA abundance requires paired measurements of mRNA and the emerging cellular phenotype in question, measurements that are technically challenging due to the destructive nature of mRNA quantification methods. (iii) The statistical measures needed to answer these questions, entropy and mutual information, are notoriously hard to infer. Below, we discuss how these challenges could be addressed to provide direct quantification of the information content in mRNA abundance.

Ca^{2+} signaling is a useful model system to quantify the dependency between mRNA abundance and emerging cellular phenotypes. Ca^{2+} signaling is a system in which the emerging phenotype is faster than changes in mRNA abundance. This timescale separation allows us to assume mRNA abundances are at a quasi-steady state and do not change significantly during the experiment. The dynamics of the Ca^{2+} signaling response to ATP is a well-studied model system for environmental sensing, featuring one of the most ubiquitous and multifunctional pathways across cell types. An important role of Ca^{2+} signaling is the coordination of responses to changes in extracellular environment. In the physiological context of tissues, cell lysis causes an unusual local increase in extracellular ATP, among other molecules. This type of damage sensing relies on the purinergic cell surface receptors, P1 and P2, which detect adenosine and ATP, respectively (Alves *et al*, 2018). The P2Y GPCR triggers a downstream signaling cascade via protein interactions.

G_q-GTP is released from the P2Y receptor where it can then bind to and activate phospholipase C (PLCβ). PLCβ cleaves PIP_2 into IP_3 and DAG, which facilitate signaling by binding to their respective receptors, IP3R and DAGR. The IP3R is embedded in the membrane of the endoplasmic reticulum and functions as a gated Ca^{2+} channel that releases Ca^{2+} into the cytoplasm upon IP_3 binding. Cytoplasmic Ca^{2+} concentrations are kept relatively low at 50–100 nM and spike up to 1uM during signaling with significant and rapid fluctuations producing unique dynamics in every cell (Bagur & Hajnóczky, 2017). Changes in cytoplasmic Ca^{2+} concentration over time (i.e., its signaling dynamics) have many emergent features such as oscillations caused by coupling between positive and negative feedback loops (Azeloglu & Iyengar, 2015). Studies have shown these dynamics specifically propagate relevant environmental and stimulus information (Selimkhanov *et al*, 2014). While in the cytoplasm, Ca^{2+} regulates many signaling molecules, for example, kinases and phosphatases, through direct binding to Ca^{2+} binding domains such as the EF-hand and through binding to calmodulin isoforms that enables it to activate kinases such as protein kinase C. These kinases affect many downstream transcriptional and protein-mediated responses that ultimately regulate cell behavior. The timescale of Ca^{2+} dynamics is significantly faster than the timescale of gene expression differentiation, allowing us to interpret a symmetric measure of dependency, such as mutual information, in a directed manner (Putney, 2012). Overall, the Ca^{2+} signaling pathway is a complex network with regulation at transcriptional and posttranscriptional levels, providing us with a great model system to dissect the phenotypic information content in mRNA abundances.

Precise measurements of dynamic single-cell, multimodal data have been collected to address these questions. Studies featuring multiomic image-based measurements have mostly focused on fixed cell measurements such as immunofluorescence, spatial arrangement of cells in tissues, and chromatin structure (Wang *et al*, 2018; Liu *et al*, 2021; Zhang *et al*, 2021). Studies that have involved dynamic phenotypes were limited by the low sensitivity of scRNAseq (Lane *et al*, 2017) or had to focus on only a handful of genes (Lee *et al*, 2014). Nonetheless, methods are being developed to integrate live cell dynamics and reliable RNA quantification of hundreds of genes (Foreman & Wollman, 2020; Genshaft *et al*, 2021). Measuring the transcriptional state of the Ca^{2+} signaling pathway requires the quantification of the abundance of hundreds of genes. Multiplexed error-robust fluorescence in situ hybridization (MERFISH) has been developed as a high-throughput, single-cell method for accurately counting large numbers of transcripts (Moffitt *et al*, 2016). Because it is performed in situ, MERFISH can be combined with other imaging methods to create high-dimensional, multimodal data consisting of both dynamic and instantaneous measurements. Combining transcriptomic and live-cell data offers unique insights into the role of dynamic regulation and sources of phenotypic information. The challenge of collecting high-dimensional, single cell, paired transcriptomic, and signaling dynamics data has been successfully addressed in recent work (Foreman & Wollman, 2020). There, we demonstrated a single-cell method for collecting paired measurements of live Ca^{2+} signaling dynamics and relevant gene expression using MERFISH. In that work, nontransformed epithelial cells that express a Ca^{2+} biosensor were activated with extracellular ATP, imaged for 13 min, and fixed for mRNA abundance quantification using MERFISH. Pairing of cells

between the two modalities of the experiment created a unique dataset of 5,128 cells with 314 timepoints of Ca^{2+} signaling dynamics and counts for 83 transcripts. This dataset was the basis for the work described here.

New analytical frameworks have emerged for understanding complex dependencies in multimodal measurements with intractable data distributions. Information theory provides a powerful analytical framework for understanding the relationship between system structure and output (Brennan *et al*, 2012). Shannon's mutual information is a statistical approach for measuring the magnitude of shared, or symmetric, information between two random variables. This framework is powerful because it captures nonlinear relationships and measures true dependence in absolute terms, although it has been difficult to apply to biochemical systems without strict assumptions about the data distribution (Tostevin & ten Wolde, 2010; Uda *et al*, 2013). However, multiomic measurements of single cells often involve different data types that are difficult to relate, that is to define a joint probability distribution. In the case of Ca^{2+} signaling networks, signaling data are sampled from a continuous process, whereas RNA abundances are discrete. Many paradigms rely on separate analysis of each data type, often via dimensionality reduction or clustering, before relationships can be quantified (Welch *et al*, 2019; Lee *et al*, 2020; Fang *et al*, 2021). While other approaches (e.g., binning or kernel-density estimation) exist for defining a joint probability distribution over some data types, they fail to perform well outside of strict assumptions about the distributions (e.g., gaussianity) or limited dimensionality; a general, scalable approach is necessary to reduce the need for complex and highly specific analytical pipelines that have emerged (Gayoso *et al*, 2021; Zuo & Chen, 2021). Highly flexible neural networks have demonstrated their ability to estimate characteristics of these probability distributions to allow a deeper understanding of the statistical and information theoretic properties of the data. Deep learning has proven useful for classification of and feature generation from ERK and Akt signaling dynamics (Jacques *et al*, 2021). However, direct interpretation of latent embeddings in these neural network outputs is challenging. An alternative use of deep learning methodologies is a universal functional approximator where neural networks are used to approximate unknown functions to achieve different objective functions. This approach was codified within variational inference and has been proven very useful in probability estimates. For complex data of mixed types where mutual information is analytically intractable, optimization of neural network functional approximator could be used to find a lower bound. This approach was recently demonstrated under the name mutual information neural estimator (MINE), which uses a deep neural network to learn a function that can encode the data and find a tight lower bound on the mutual information (preprint: Belghazi *et al*, 2018). Briefly, MINE is a universal function approximator that searches for a mapping function T in a large space of encoder functions parameterized by θ. T maps the data consisting of mRNA counts and Ca^{2+} signal dynamics, G and Ca^{2+} respectively, of arbitrary dimensionality such that $T_\theta : (G, Ca^{2+}) \rightarrow \mathbb{R}$. Remarkably, the data require no significant transformations, MINE is simply trained on the raw data because all transformations required for an efficient mapping are theoretically learnable by a model with enough parameters and samples. Letting $\mathbb{P} := \mathbb{P}_{G,Ca^{2+}}$ represent the joint probability, that is, paired data, and $\mathbb{Q} := \mathbb{P}_G \otimes \mathbb{P}_{Ca^{2+}}$ represent the

product of the marginal probabilities, that is, independently sampled data, the mutual information between G and Ca^{2+} is the distance between the joint and marginal distributions. This distance is measured using the Kullback–Leibler divergence (D_{KL}), and a stronger relationship between G and Ca^{2+} is equivalent to a greater distance between the joint and marginals: $I(G, Ca^{2+}) = D_{KL}(\mathbb{P} \parallel \mathbb{Q})$. Using the Donsker-Varadhan representation of the D_{KL}, the model parameters θ are optimal when gradient ascent has maximized $\mathbb{E}_{\mathbb{P}}[T_\theta] - \log(\mathbb{E}_{\mathbb{Q}}[e^{T_\theta}]) \leq D_{KL}(\mathbb{P} \parallel \mathbb{Q})$, where $\mathbb{E}$ denotes the expected value. This estimate represents a lower bound on the mutual information. MINE is highly flexible because it makes almost no assumptions about the structure of the data. MINE searches through a large function space for the optimal transformation function to encode the data types assuming there are enough samples to constrain the model. The result is a lower-bound estimate of the mutual information between paired modalities of almost any dimensionality and complexity.

The recent technological development in multiplexed single-cell measurements and machine learning approaches for the inference of mutual information could be integrated to provide direct quantification of the phenotypic information content of mRNA abundances. Here, we utilize these developments and focus on a model with timescale separation between an emerging phenotype and mRNA abundance. We relied on highly multiplexed FISH-based quantification of mRNA levels that is more accurate than sequencing-based approaches and also allows integration with other imaging modalities. Inference of mutual information was done using the Ca^{2+} signaling network as a model; we fit MINE on various subsets of 83 genes and 314 Ca^{2+} timepoints to quantify the contribution of transcript abundance to signaling dynamics. To establish a baseline, we first calculated the dependency between individual genes and Ca^{2+} signals. We then calculated the mutual information between gene pairs and Ca^{2+} signals to account for redundancy. Gene sets of all sizes were then sampled using various strategies to measure how information changes with set size. Using PCA, we evaluate how useful phenotypic information accounts for transcript-level variance. Overall, we demonstrate a new information theoretic framework for analyzing paired single-cell data that provide a quantification of the dependency between sets of mRNAs and an emergent cell-scale dynamic phenotype.

Results

To investigate the information content of transcript counts and dynamic Ca^{2+} signals, we first analyzed each modality on their own. Ca^{2+} signals display significant heterogeneity across cells (Fig 1A). Likewise, most transcripts had a large range of abundances across cells, although distributions varied depending on the transcript. Pairwise correlation coefficients were calculated for 83 genes across 5,128 cells (Fig 1B). The magnitude of the correlations for all gene pairs was relatively low with an average of $r = 0.16$ compared to just cell cycle genes at $r = 0.44$. Assuming most genes are generally informative, one interpretation of the low correlations and heterogeneous transcript distributions is that transcripts contain unique information. To test this hypothesis, we quantified the transcriptional information by performing PCA then calculating the differential entropy, a measure of information for
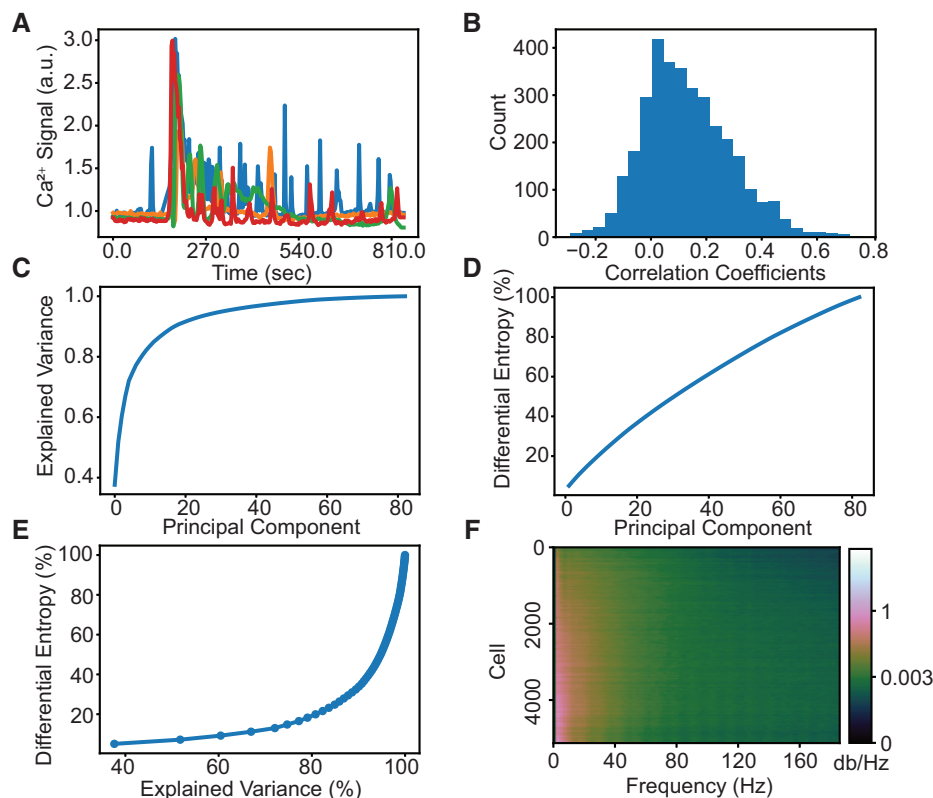
Figure 1. Structure of gene and Ca²⁺ data.

A Representative examples of Ca²⁺ dynamics of four cells in the dataset.
B A histogram of the pairwise gene correlation matrix (tri-up) which highlights the relatively low correlations.
C Explained variance of mRNA transcript counts from PCA.
D Differential entropy of transcripts estimated by PCA.
E Plot of explained variance (panel C) vs differential entropy (panel D) with an increasing number of principal components.
F Dynamic Ca²⁺ signal periodogram (cropped to show only the lower wavelength, higher power frequencies). Ca²⁺ dynamic signals were found to contain a spectral entropy of 4.2 bits.

Data information: Collectively, panels (C–E) show that most of the entropy comes from components that do not explain much of the variance.

continuous probability distributions, of the components (equation 2, Fig 1C–E). Because each principal component is an independent, weighted sum of the row vectors of the data, we can approximate the differential entropy among orthogonal components assuming normality via the central limit theorem. Differential entropy across principal components does not measure the information in absolute terms but can describe how the information is distributed relative to the explained variance. We found that six principal components explain 75% of the variance, but only 15% of the gene entropy. The contrast between Fig 1C and D appears contradictory in that few orthogonal components explain most of the variance, yet entropy is steadily added across components with no obvious plateau. This analysis shows that simple measures such as explained variance that are often used for dimensionality reduction are not necessarily appropriate proxies of information content. Accounting for relevant, phenotypic information could help resolve discrepancies between explained variance and differential entropy. To estimate the signal entropy, that is, information content in Ca²⁺ signaling, we took advantage of its dynamic patterns. Differential entropy of Ca²⁺ can be estimated

using spectral entropy (equation 1), a scale-invariant measure of information (Burg, 1975). The periodogram shows a continuum of signal frequencies with apparently low variance (Fig 1F). Signal entropy can be calculated using this distribution of frequencies, which we found to be 4.2 bits or ~18 signaling states. The mutual information between mRNA abundance and Ca²⁺ signaling is bounded by the distribution with the lowest entropy. Thus, 4.2 bits provides a likely upper bound to the true mutual information between transcripts and Ca²⁺ signals.

To quantify how useful phenotypic information is distributed across genes, we estimated mutual information between individual genes and Ca²⁺ signals (Fig 2A). To help choose hyperparameters and evaluate MINE's performance, we tested the model on multivariate gaussian distributions and found a mean residual of 0.37 bits with a Pearson's correlation coefficient of 0.97 to the ground truth (Appendix Fig S2). Applied to the experimental data, we measured the total mutual information between all genes and Ca²⁺ at 2.5 ± 0.4 bits. Most individual genes contain significant information about Ca²⁺ signals, an average of 0.7 bits, and the most informative gene accounts for 34% of the 4.2 bits of signal entropy. Cell cycle-
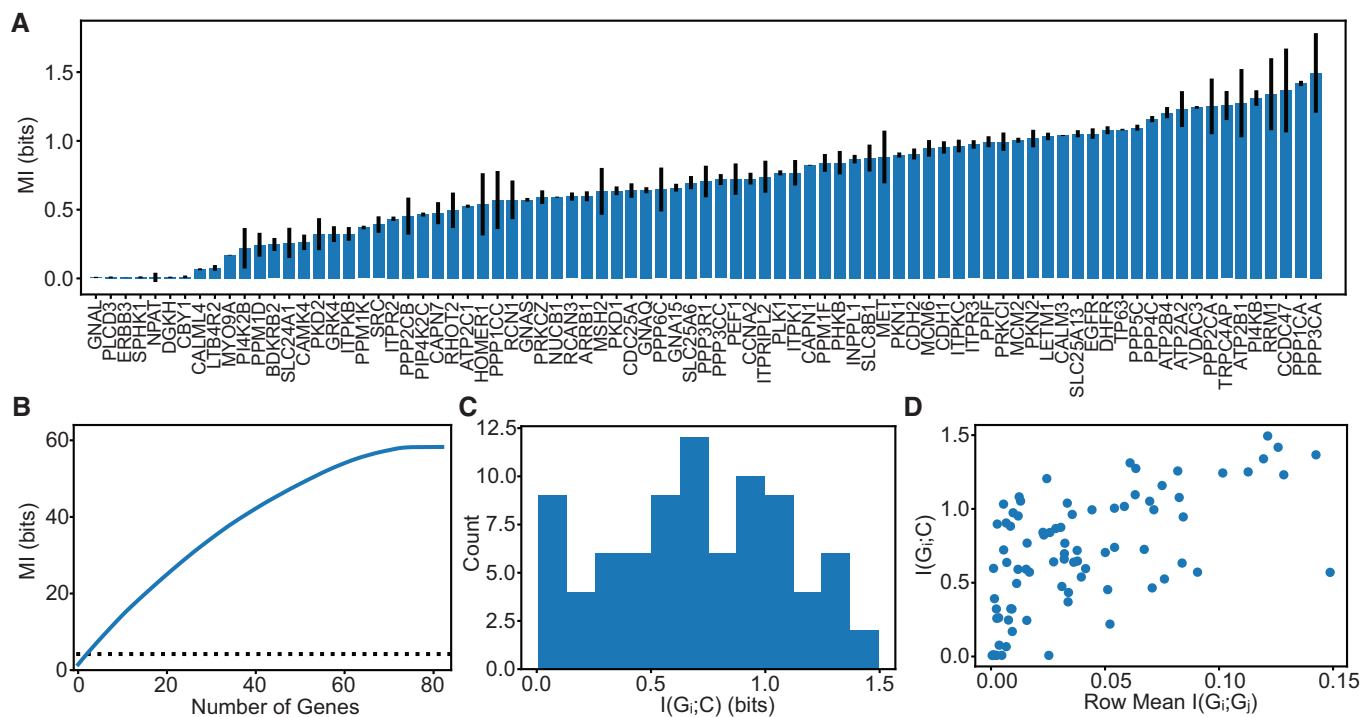
Figure 2. Mutual information between individual genes and Ca²⁺ signals.

A $I(G_i;Ca^{2+})$ sorted from least to greatest. Error bars show standard deviation of three technical replicates, that is three independent runs of the MINE algorithms on the same data.

B   The blue line shows the cumulative sum of $I(G_i;Ca^{2+})$ from (A) sorted from greatest to least, and individual genes appear to contain a lot of information (56 bits) about Ca²⁺ signals. The black dashed line shows mutual information between all 83 genes and Ca²⁺ dynamics estimated to be 2.5 bits.

C   Histogram of (A) showing the mean $I(G_i;Ca^{2+})$ is 0.7 bits.

D $I(G_i;G_j)$ represents the pairwise mutual information between genes, the information that genes have about each other. This plot shows that genes that are more informative about other genes tend to be more informative about Ca²⁺ dynamics.

associated genes were well distributed throughout the list, whereas genes coding for Ca²⁺ and/ or calmodulin-dependent proteins such as *PPP3CA* and *CCDC47* were at the top of the list. If each gene contained completely unique information, then the sum of the phenotypic information in each gene should add up to the total of 2.5 ± 0.4 bits. Interestingly, this sum is significantly larger than the total $I(G;Ca^{2+})$, indicating a high degree of redundancy (Fig 2B). The average mutual information between a single gene and Ca²⁺ signals is 0.7 bits, which is 17% of the signal entropy (Fig 2C). How the mutual information is shared across genes is not immediately clear. We further tested whether informative genes, that is, genes that have high average pairwise mutual information to other genes, are also informative about Ca²⁺ dynamics (Fig 2D). Overall, genes that are more informative about Ca²⁺ signaling are also more informative about the expression of other transcripts. These genes that are highly informative about Ca²⁺ and many other transcripts may be interpreted as summary genes containing redundant, but distributed information. The second most informative single gene, *PPP1CA*, exemplifies this effect, as it codes for a subunit of PP1 that interacts with >200 regulatory proteins involved in a myriad of critical cell processes. Notably, the top two most informative genes, *PPP3CA* and *PPP1CA*, are both broadly connected phosphatases; kinases and phosphatases were consistently informative and concentrated toward the top of the list. However, from this analysis alone, it is

not clear how many genes contain redundant information and to what extent.

To better understand how the superfluous information in Fig 2B is distributed among genes, we calculated the synergy redundancy index (SRI) between gene pairs with respect to Ca²⁺ (Dietterich *et al*, 2002; Schneidman *et al*, 2003). $SRI(G_i,G_j \mid Ca^{2+})$ measures the information overlap between genes by subtracting $I(G_i;Ca^{2+})$ and $I(G_j;Ca^{2+})$ from $I(\{G_i,G_j\};Ca^{2+})$. A gene pair with negative SRI means that the sum of the mutual information between each gene and Ca²⁺ was greater than the gene pair, so the genes must contain some of the same information (redundant). A positive SRI indicates that there is more information about Ca²⁺ in the gene pair than in the sum of the individual genes (synergistic). An SRI of 0 describes a pair of genes that are either generally uninformative or are independent, containing unique and nonoverlapping information about Ca²⁺. Calculating SRI between all gene pairs reveals that most pairs are significantly redundant (Fig 3A and B). On average, gene pairs share 0.43 bits which accounts for 61% of the 0.7 bits of phenotypic information contained in the average individual gene. Furthermore, the more informative a gene is about Ca²⁺, the more redundant it is with other genes (Fig 3C). This finding supports that some genes aggregate information from many others and the more information a gene has, the more it shares. Consistent with findings in Fig 2A, phosphatases and kinases such as *PPP3CA, PRKCI, PPP2CA, PI4KB,*
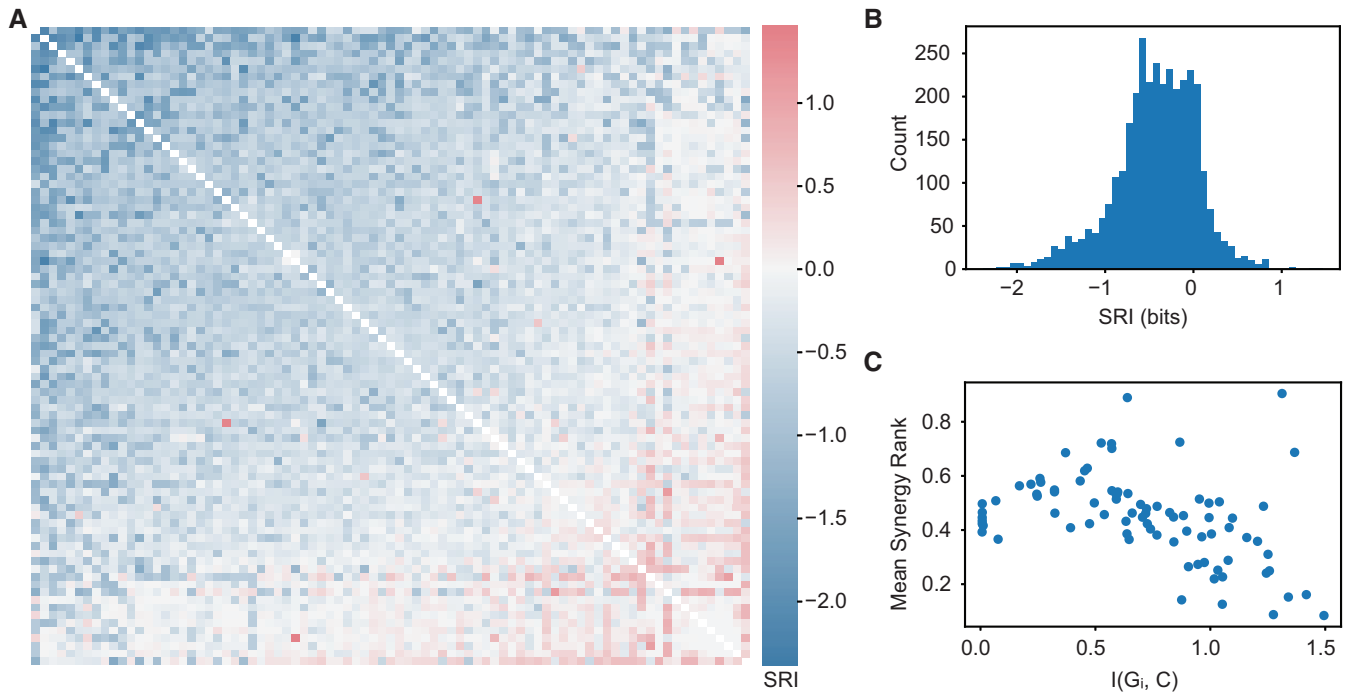
Figure 3. Synergy and redundancy of gene pairs with respect to Ca^{2+}.

A SRI(G$_i$,G$_j$|C) sorted by average SRI.
B Histogram of SRI showing that most gene pairs are highly redundant with an average score of −0.43 bits.
C The mean rank of all synergistic pairs compared to the mutual information between that gene and Ca^{2+} signals, (spearman *r* = 0.5, *P* < 2e-6), indicating that genes with more information about Ca^{2+} are also more redundant.

and *PPP1CA* were among the most redundant. Interestingly, some genes are highly synergistic on average. One such synergistic gene is *PLCD3*, which appears to have no information about Ca^{2+} on its own, but suddenly becomes informative in gene pairs. *PLCD3* codes for an isoform of phospholipase C, a critical step in the Ca^{2+} in the signal transduction of extracellular ATP. It is surprising that *PLCD3* expression appears to contain little information about Ca^{2+} on its own considering its relevance to stimulus sensing, but this apparent paradox is reconciled by its high synergy. The most synergistic gene on average was *ATP2C1*, which codes for a calcium-transporting ATPase that couples ATP hydrolysis with Ca^{2+} transport into the Golgi lumen. Genes which are critical for modulating Ca^{2+} concentrations in the cytoplasm represent steps in linear processes, rather than cooperating with many other components to achieve their function. Generally, the most synergistic genes were not very informative on their own (Fig 2A) but became informative in a group of 2 (Fig 3C). The high degree of synergy suggests that these genes provide contextual or conditional information that is absent from most other genes, even genes that were independently informative. Most of the genes with near zero SRI were generally uninformative about signaling based on Fig 2A. Although, genes typically function in larger sets beyond pairs, and a thorough understanding of transcriptional information requires evaluation of higher-order interactions.

To explore the mutual information between Ca^{2+} and gene sets of various sizes, we tested various sets using gene annotations, a sequential search, and PCA. To quantify set level information at a functional level, we summarized pairwise SRIs based on gene annotations (Fig 4A). Calculating the mean SRI for combinations of annotations revealed how different functional gene sets contain phenotypic information. The Ca^{2+}/ER annotation contains the most redundancies by a large margin, whereas the miscellaneous category "Other" is the most synergistic which can be explained by the functional diversity in this group. The Ca^{2+}/ER annotation contains the genes most relevant to the stimulus and appear to provide similar information. To understand how phenotypic information depends on gene set size, we calculated the mutual information between Ca^{2+} signals and gene sets of all sizes. Because testing all possible sets is prohibitively computationally expensive, we first sampled random sets of all possible sizes (4B). Each set size was sampled four times. For random sets, 53 genes contained 54% of the phenotypic information. To understand the upper and lower bounds on information in each set, we performed two directed heuristic searches. The directed searches first picked the most (least) informative gene, and then tested every possible addition to the set to add the member that contributed the most (least) information until the sets were of maximum size. The upper bound in green shows that the information quickly plateaus as the best 12 genes contain 54% of the phenotypic information, and all further additions contribute minimal additional information. Compared to random gene sets, using only the most informative combination of genes dramatically reduces the number of genes required to recapitulate most of the phenotypic information from 53 to 12. The lower bound in purple shows the unique information per gene given the
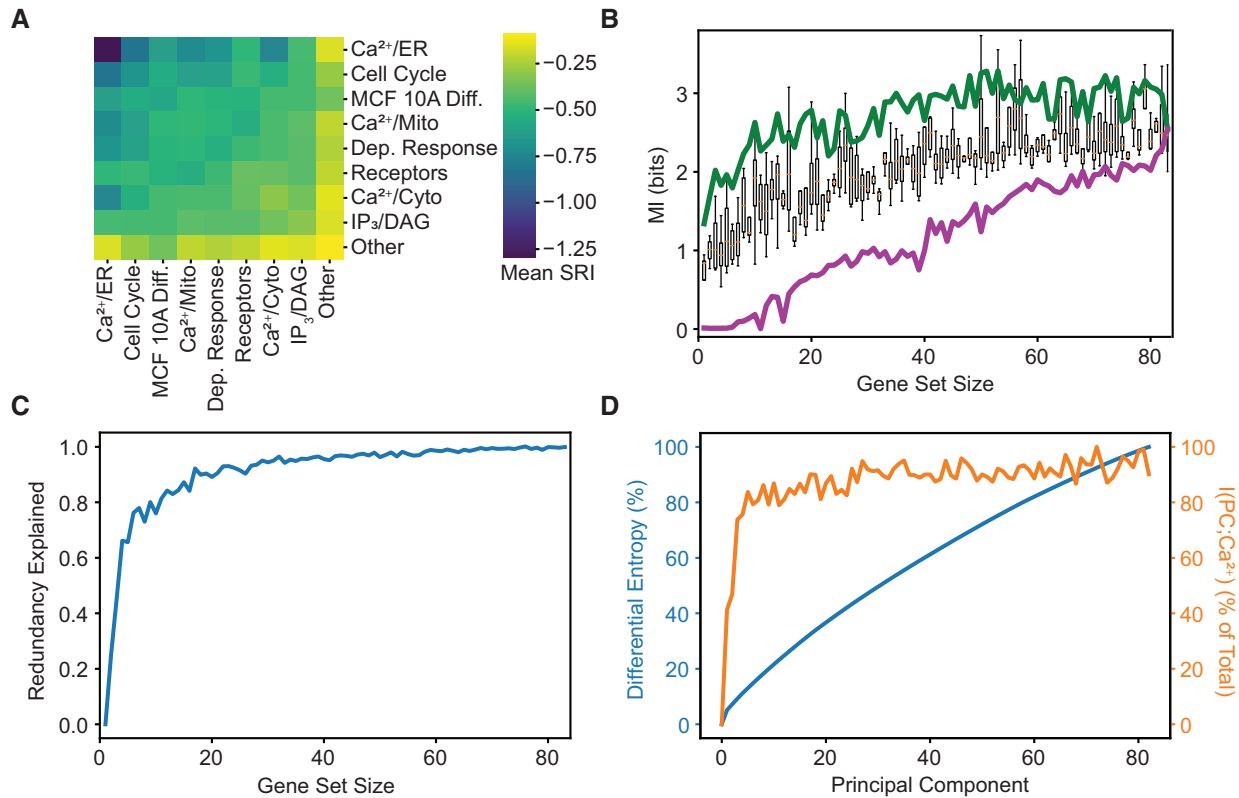
Figure 4. Mutual information between gene sets and Ca²⁺ signals.

A   Mean pairwise SRI from Fig 3 for sets based on annotation. MCF 10A differentiation and Ca²⁺-dependent response are abbreviated.
B   Gene sets of various sizes were constructed using three different strategies: an upper bound (green) that always adds the most informative gene to the set given the genes already included, random strategy (boxes, center band shows the median, box covers the 25–75% and whiskers show 95% confidence interval. Boxes are based on 12 random samples of genes) that samples random sets of genes, and a lower bound (purple) that always adds the least informative gene to the set given the genes already included.
C   The blue line shows the fraction of redundant information using the expected value of $I(\{G_0, \ldots, G_n\};Ca^{2+})$ from equation (3).
D   A *y-y* plot of gene differential entropy in blue (same as Fig 1D) and the mutual information between gene principal components and Ca²⁺ in orange. Both values are normalized by their respective max values.

set, sorted from least to greatest. Because the lower bound always adds the least informative and most redundant genes first, the last genes contain the most unique information. *PPP3CA* is the first gene added to the upper bound and the last gene to be added to the lower bound, which means it must have both the most absolute information and the most unique information. Interestingly, the growth of information in the least informative set was approximately linear, meaning that there is always some unique information in every gene. The slope of the lower bound is 0.03 bits, which represents the average unique information per gene.

Using the mean mutual information between a gene set and Ca²⁺, we can also estimate the "redundancy explained" of all sets of genes of a given size according to equation (5) (Fig 4C). We found that sets of only three genes explain 66% of the redundancy. Small gene sets contain much more redundant phenotypic information than larger gene sets. The point at which this curve begins to level off can be interpreted as a fundamental set size above which most phenotypic information lies within the sets. We observe that small gene sets contain most of the information on Ca²⁺ dynamics suggesting that higher-order interactions in larger sets are not

required to capture the full dependency between mRNA abundance and Ca²⁺ dynamics.

Finally, we calculated the mutual information between transcript principal components and Ca²⁺ signals to compare with differential entropy (from Fig 1D) and understand how useful phenotypic information is distributed (Fig 4D). In agreement with Fig 4C, phenotypic information saturates quickly with only three principal components accounting for 74% of the 2.5 bits of mutual information between transcripts and signals. This result starkly contrasts with the differential entropy of gene principal components independent of Ca²⁺ signals which rises slowly and does not appear to plateau. By accounting for phenotypic information, far fewer orthogonal components are required to preserve the useful information. The difference between these curves indicates that focusing on phenotypic information may filter or compress transcriptional information. $I(PC;Ca^{2+})$ resembles the curve in Fig 1C, although still plateaus more quickly. These results confirm that phenotypic information is mostly explainable by a few components and higher-order interactions do not significantly contribute.

Discussion

The complexity of biological regulation is staggering. While many details about biological networks are known, the gaps in our knowledge make some simple questions very challenging to answer. For example, to what degree does abundance of one set of molecules matter? Specifically, does the abundance of mRNAs matter for the regulation of complex cellular phenotypes such as signaling response to a ligand? Here, we provide a framework for answering such questions through the combination of paired single-cell data and application of recent advances at the interface between machine learning and information theory (preprint: Belghazi *et al*, 2018). Applying a recently developed framework for mutual information estimation to single-cell data of multiplexed mRNA levels paired with live cell imaging allowed us to quantify the strength of the causal connection between mRNAs and Ca^{2+} signaling. We found that approximately 60% of Ca^{2+} signal information exists in transcript counts, which is 2.5 (± 0.4) bits. Furthermore, the framework we developed provides key information about information synergy and redundancy can be used to quantify the joint information in sets of genes, and reveals how overall dependency changes with the size of the set. On average, genes were found to contain 61% redundant information with each other, although nearly all genes contained some unique information. Genes that appeared to contain little phenotypic information individually were in fact the most synergistic and became informative in pairs. The unique information present in gene sets is best visualized in Fig 4B, which illustrates the difference in information among the most, least, and average set. In the best case, only 12 genes contain 54% of signal information, which is significantly fewer than an equally informative 53 random genes. While all genes contain unique information, some sets are still significantly more informative than others likely due to their role in the signaling network. Decomposition by principal components (Fig 4D) revealed a rapid plateau in phenotypic information, starkly contrasting the increasing growth in differential entropy. These results demonstrate the utility of information theoretic analysis in quantifying the phenotypic information of mRNA abundance.

The framework we propose is very general and can be applied to any two "slices" within a complex biological regulatory network. Our numerical experiments (Supplementary Material) demonstrate that with minor adaptations for bias removal, MINE can robustly estimate mutual information between two high-dimensional vectors containing 100+ features. The generalizability of this framework provides a new tool to put weights and interpretable numbers on different "arrows" within complex biological regulatory networks. Importantly, such "arrows" do not necessarily represent direct mechanistic steps. There are numerous reactions that occur post-transcriptionally to determine Ca^{2+} signaling responses. Yet, using MINE, we were able to infer the individual contribution of each gene in controlling the emergent phenotypes. Furthermore, using pairs of genes and estimation of the effect of gene set size, we determined how information between multiple mRNA types is integrated. This inference showed that despite the information having to propagate through multiple layers of regulation, it still shows significant dependency. Even though correlations between mRNA and protein levels are generally low, a substantial amount of phenotypically relevant information is still preserved in the transcriptome. Our results support the use of mRNA measurements to infer and predict useful phenotypic characteristics of cell populations. One interpretation of the 2.5 (± 0.4) bits of mutual information is that transcripts can differentiate approximately six distinct states of Ca^{2+} signaling dynamics. An important feature of our analysis is that all inference was done relaying on natural heterogeneity without any experimental perturbation to gene expression circumventing compensation and nonlinear dependencies that are common pitfalls of perturbation analysis (Welf & Danuser, 2014).

While the framework we propose is very general, our findings are systems specific and will change depending on the set of genes and measured phenotypes. Here we focused on Ca^{2+} signaling in response to activation of GPCR in a clonal population of MCF 10A cells. In previous work, we estimated that a cellular population is composed of multiple subtypes (Yao *et al*, 2016) and have shown that mRNA variability is dominated by cell state differences with a minor contribution from transcriptional bursting (Foreman & Wollman, 2020). Our current finding that 60% of information in the emerging Ca^{2+} signaling phenotypes can be attributed to cellular transcriptional state largely agrees with these previous findings. It is likely that in other systems, decomposition of information content will differ from the 60% transcriptional and 40% posttranscriptional measured here. For example, broad phenotypes such as cell type classification that often correspond to larger and highly patterned transcriptional differences will likely show higher levels of transcriptional dependency. Additionally, it is possible that the full transcriptome may contain more phenotypic information than is found in just the 83 genes measured in this study. While only 12 genes accounted for most of the shared information, the apparent plateau and informational redundancy may result from the strong functional relationships and dependencies in the selected gene set. Including significantly more genes related to other cellular processes may provide more information about the observed phenotype by better defining the transcriptional state or revealing indirect dependencies to other cellular processes.

Our approach has several limitations, experimental and computational, that will need to be addressed in future work. Experimentally, gene selection, that is, the expression of which genes are measured, is limited due to gene length, specific sequence, and other experimental constraints that are continuously improving. Furthermore, the approach could be applied to tissue samples with much higher population diversity where the relationship between transcripts and phenotype is more relevant. Computationally, because of stochastic gradient ascent, the model's estimates are somewhat noisy and required multiple replicates. Additionally, we were limited to explore the effect of set sizes with search strategies and only exhaustively examined pairwise dependencies because the model was computationally expensive to run. None of the search strategies are guaranteed to find the truly most or least informative set because doing so would require a prohibitively time-consuming, exhaustive search. Despite these limitations, MINE was able to provide an interpretable and scalable quantification of dependency between transcript sets and Ca^{2+} signaling.

Recent advances in single-cell technologies are making high-dimensional, multimodal measurements feasible. Statistical descriptions of complex phenotypes will become increasingly useful as single-cell experiments generate more multimodal and multiomic data. Integrating multiple different data types is still a challenge in

the field, and this work represents a new approach to synthesize statistical descriptions of high-dimensional, multimodal data that does not make any assumptions about the underlying functional relationships. This unbiased approach will enable a deeper understanding of complex, multidimensional data by quantifying the dependency between any single cell phenomena.

Materials and Methods

Data selection

Data collection is described in previous work (Foreman & Wollman, 2020). Of the 336 genes measured, 150 genes were measurably expressed and the top 83 were chosen by the highest magnitude z scores from multiple linear regression.

Preprocessing

Transcript counts from the 83 genes and 314 timepoints of Ca^{2+} signals were independently *z*-score normalized. Normalization was applied to the entire matrix of all cells for each data type (e.g., 5128x83 for transcripts) and not to individual columns, preserving relative magnitude across genes and timepoints.

Spectral entropy

Spectral entropy of a signal is defined as the Shannon entropy (*H*) of the normalized power spectral density (*P*), calculated here using the Fourier transform. Although the calculation requires a sampling frequency (f_s), the result does not change above a sufficiently large value.

$$H(f_s) = -\sum_{f=1}^{f_s/2} P(f)\log_2 P(f) \tag{1}$$

Calculation of spectral entropy was robust to changes in scale and dimensionality of the input data (see Appendix Fig S4).

Differential entropy

Differential entropy was calculated via the determinant of the covariance matrix of the PCA-transformed data. This approach was used to estimate the entropy of the mRNA transcript counts.

$$\frac{n}{2} + \frac{n\log_2 2\pi}{2} + det\Sigma_{PC} \tag{2}$$

n = number of principal components, Σ = the covariance matrix, PC = PCA-transformed data for a given *n*.

MINE

Hyperparameters were chosen by fitting analytically tractable data from an additive white gaussian noise model of the data across a range of strengths of dependence (Appendix Fig S1). Additional bias correction was implemented by fitting, where $I_{obs}(t) = I_{true}(1-a) * e^{-b*t} + ct$, I_{true}, a, b and c are the fitting parameters and

is the number of iterations (Appendix Fig S2). Convergence tests were performed on the real data by comparing the residuals of the bias correction fit. The chosen hyperparameters of 600 hidden units and a learning rate = 3e-4 resulted in the highest yield, that is, fewest failed fits. We performed a jackknife bias correction on all MINE inferences as shows in Appendix Fig S3.

Synergy redundancy index (SRI)

The synergy redundancy index was developed to evaluate information about a stimulus shared among a small population of cells (Dieterich *et al*, 2002). Equation (3) describes the calculation, which involves comparing pairwise and individual mutual information between genes and Ca^{2+} signals.

$$SRI \ (G_i, G_j, Ca^{2+}) = I(G_i, G_j, Ca^{2+}) - I(G_i, Ca^{2+}) - I(G_j, Ca^{2+}) \tag{3}$$

First, the mutual information between each unique pair of genes and Ca^{2+} were estimated, $I(G_i, G_j; Ca^{2+})$. Then, $I(G_i; Ca^{2+})$ was calculated, and equation (3) was calculated for all genes.

Redundancy explained

This metric represents the amount of extra information assuming no redundancy between elements. Equation (4) first calculates an expected value by taking the mean of sampled gene sets of size *k*. The expected value is multiplied by the number of possible sets then divided the number of times an individual gene appears in all sets to calculate the nonredundant information (NRI) as if all individual sets contain unique information:

$$NRI = \frac{\mathbb{E}\big[I(\{G_1, \ldots, G_k\}, Ca^{2+})\big]\binom{n}{k}}{\binom{n-1}{k-1}} \tag{4}$$

$$= \mathbb{E}\big[I(\{G_1, \ldots, G_k\}, Ca^{2+})\big]\frac{n}{k}$$

k = set size, *n* = total number of genes.

From equation (4), we can calculate the fraction of purely redundant information for a set of size *k* out of the maximum, which is the NRI at *k* = 1 minus the full mutual information. The Redundancy Explained (RE) is calculated as follows:

$$RE = 1 - \frac{NRI_k - I(G, Ca^{2+})}{NRI_1 - I(G, Ca^{2+})} \tag{5}$$

Data availability

This study includes no data deposited in external repositories.

Author contributions

Evan Maltz: Conceptualization; software; formal analysis; validation; investigation; visualization; methodology; writing – original draft; writing – review and editing. **Roy Wollman:** Conceptualization; formal analysis; supervision; funding acquisition; validation; methodology; writing – original draft; project administration; writing – review and editing.

Disclosure and competing interests statement

The authors declare that they have no conflict of interest.

References

Adelaja A, Taylor B, Sheu KM, Liu Y, Luecke S, Hoffmann A (2021) Six distinct NFκB signaling codons convey discrete information to distinguish stimuli and enable appropriate macrophage responses. *Immunity* 54: 916–930.e7

Alves M, Beamer E, Engel T (2018) The metabotropic purinergic P2Y receptor family as novel drug target in epilepsy. *Front Pharmacol* 9: 193

Azeloglu EU, Iyengar R (2015) Signaling networks: information flow, computation, and decision making. *Cold Spring Harb Perspect Biol* 7: a005934

Bagur R, Hajnóczky G (2017) Intracellular Ca2+ sensing: its role in calcium homeostasis and signaling. *Mol Cell* 66: 780–788

Balázsi G, van Oudenaarden A, Collins JJ (2011) Cellular decision making and biological noise: from microbes to mammals. *Cell* 144: 910–925

Belghazi MI, Baratin A, Rajeswar S, Ozair S, Bengio Y, Courville A, Devon Hjelm R (2018) MINE: mutual information neural estimation. *arXiv [csLG]* https://doi.org/10.48550/arXiv.1801.04062 [PREPRINT]

Brennan MD, Cheong R, Levchenko A (2012) Systems biology. How information theory handles cell signaling and uncertainty. *Science* 338: 334–335

Burg JP (1975) Maximum entropy spectral analysis. PhD Thesis, Stanford University, Palo Alto, CA

Cheong R, Rhee A, Wang CJ, Nemenman I, Levchenko A (2011) Information transduction capacity of noisy biochemical signaling networks. *Science* 334: 354–358

Darmanis S, Gallant CJ, Marinescu VD, Niklasson M, Segerman A, Flamourakis G, Fredriksson S, Assarsson E, Lundberg M, Nelander S *et al* (2016) Simultaneous multiplexed measurement of RNA and proteins in single cells. *Cell Rep* 14: 380–389

Dietterich TG, Becker S, Ghahramani Z (2002) Advances in neural information processing systems. Proceedings of the 2001 Conference MIT Press

El-Samad H (2021) Biological feedback control-respect the loops. *Cell Syst* 12: 477–487

Emert BL, Cote CJ, Torre EA, Dardani IP, Jiang CL, Jain N, Shaffer SM, Raj A (2021) Variability within rare cell states enables multiple paths toward drug resistance. *Nat Biotechnol* 39: 865–876

Fang R, Preissl S, Li Y, Hou X, Lucero J, Wang X, Motamedi A, Shiau AK, Zhou X, Xie F *et al* (2021) Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat Commun* 12: 1337

Foreman R, Wollman R (2020) Mammalian gene expression variability is explained by underlying cell state. *Mol Syst Biol* 16: e9146

Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, Yosef N (2021) Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat Methods* 18: 272–282

Genshaft AS, Ziegler CGK, Tzouanas CN, Mead BE, Jaeger AM, Navia AW, King RP, Mana MD, Huang S, Mitsialis V *et al* (2021) Live cell tagging tracking and isolation for spatial transcriptomics using photoactivatable cell dyes. *Nat Commun* 12: 1–15

Gong H, Wang X, Liu B, Boutet S, Holcomb I, Dakshinamoorthy G, Ooi A, Sanada C, Sun G, Ramakrishnan R (2017) Single-cell protein-mRNA correlation analysis enabled by multiplexed dual-analyte co-detection. *Sci Rep* 7: 2776

Hafner A, Stewart-Ornstein J, Purvis JE, Forrester WC, Bulyk ML, Lahav G (2017) p53 pulses lead to distinct patterns of gene expression albeit similar DNA-binding dynamics. *Nat Struct Mol Biol* 24: 840–847

Jacques M-A, Dobrzyński M, Gagliardi PA, Sznitman R, Pertz O (2021) CODEX, a neural network approach to explore signaling dynamics landscapes. *Mol Syst Biol* 17: e10026

Kinnunen PC, Luker KE, Luker GD, Linderman JJ (2021) Computational methods for characterizing and learning from heterogeneous cell signaling data. *Curr Opin Syst Biol* 26: 98–108

Lane K, Van Valen D, DeFelice MM, Macklin DN, Kudo T, Jaimovich A, Carr A, Meyer T, Pe'er D, Boutet SC *et al* (2017) Measuring signaling and RNA-Seq in the same cell links gene expression to dynamic patterns of NF-κB activation. *Cell Syst* 4: 458–469.e5

Lee J, Hyeon DY, Hwang D (2020) Single-cell multiomics: technologies and data analysis methods. *Exp Mol Med* 52: 1428–1442

Lee RE, Walker SR, Savery K, Frank DA, Gaudet S (2014) Fold change of nuclear NF-κB determines TNF-induced transcription in single cells. *Mol Cell* 53: 867–879

Liu Y, Beyer A, Aebersold R (2016) On the dependency of cellular protein levels on mRNA abundance. *Cell* 165: 535–550

Liu M, Yang B, Hu M, Radda JSD, Chen Y, Jin S, Cheng Y, Wang S (2021) Chromatin tracing and multiplexed imaging of nucleome architectures (MINA) and RNAs in single mammalian cells and tissue. *Nat Protoc* 16: 2667–2697

Macaulay IC, Ponting CP, Voet T (2017) Single-cell multiomics: multiple measurements from single cells. *Trends Genet* 33: 155–168

Mair F, Erickson JR, Voillet V, Simoni Y, Bi T, Tyznik AJ, Martin J, Gottardo R, Newell EW, Prlic M (2020) A targeted multi-omic analysis approach measures protein expression and low-abundance transcripts on the single-cell level. *Cell Rep* 31: 107499

Moffitt JR, Hao J, Wang G, Chen KH, Babcock HP, Zhuang X (2016) High-throughput single-cell geneexpression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc Natl Acad Sci USA* 113: 11046–11051

Myers PJ, Lee SH, Lazzara MJ (2021) Mechanistic and data-driven models of cell signaling: tools for fundamental discovery and rational design of therapy. *Curr Opin Syst Biol* 28: 100349

Nelson DE, Ihekwaba AEC, Elliott M, Johnson JR, Gibney CA, Foreman BE, Nelson G, See V, Horton CA, Spiller DG *et al* (2004) Oscillations in NF-kappaB signaling control the dynamics of gene expression. *Science* 306: 704–708

Perkins TJ, Swain PS (2009) Strategies for cellular decision-making. *Mol Syst Biol* 5: 326

Purvis JE, Lahav G (2013) Encoding and decoding cellular information through signaling dynamics. *Cell* 152: 945–956

Putney JW (2012) Calcium signaling: deciphering the calcium-NFAT pathway. *Curr Biol* 22: R87–R89

Schneidman E, Bialek W, Berry MJ 2nd (2003) Synergy, redundancy, and independence in population codes. *J Neurosci* 23: 11539–11553

Schulz D, Zanotelli VRT, Fischer JR, Schapiro D, Engler S, Lun X-K, Jackson HW, Bodenmiller B (2018) Simultaneous multiplexed imaging of mRNA and proteins with subcellular resolution in breast cancer tissue samples by mass cytometry. *Cell Syst* 6: 25–36.e5

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

Shaffer SM, Emert BL, Reyes Hueros RA, Cote C, Harmange G, Schaff DL, Sizemore AE, Gupte R, Torre E, Singh A *et al* (2020) Memory sequencing reveals heritable single-cell gene expression programs associated with distinct cellular behaviors. *Cell* 182: 947–959.e17

Spencer SL, Gaudet S, Albeck JG, Burke JM, Sorger PK (2009) Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature* 459: 428–432

Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 14: 865–868

Subramanian I, Verma S, Kumar S, Jere A, Anamika K (2020) Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights* 14: 1177932219899051

Taniguchi Y, Choi PJ, Li G-W, Chen H, Babu M, Hearn J, Emili A, Xie XS (2010) Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329: 533–538

Tostevin F, ten Wolde PR (2010) Mutual information in time-varying biochemical systems. *Phys Rev E Stat Nonlin Soft Matter Phys* 81: 061917

Uda S, Saito TH, Kudo T, Kokaji T, Tsuchiya T, Kubota H, Komori Y, Ozaki Y-I, Kuroda S (2013) Robustness and compensation of information transmission of signaling pathways. *Science* 341: 558–561

Wang G, Moffitt JR, Zhuang X (2018) Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. *Sci Rep* 8: 4847

Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ (2019) Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* 177: 1873–1887.e17

Welf ES, Danuser G (2014) Using fluctuation analysis to establish causal relations between cellular events without experimental perturbation. *Biophys J* 107: 2492–2498

Wong VC, Mathew S, Ramji R, Gaudet S, Miller-Jensen K (2019) Fold-change detection of NF-κB at target genes with different transcript outputs. *Biophys J* 116: 709–724

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

Zhang Q, Gupta S, Schipper DL, Kowalczyk GJ, Mancini AE, Faeder JR, Lee REC (2017) NF-κB dynamics discriminate between TNF doses in single cells. *Cell Syst* 5: 638–645.e5

Zhang M, Eichhorn SW, Zingg B, Yao Z, Cotter K, Zeng H, Dong H, Zhuang X (2021) Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature* 598: 137–143

Zuo C, Chen L (2021) Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Brief Bioinform* 22: bbaa287