

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Modeling and Measuring Dynamic Cellular Signaling States

Permalink

<https://escholarship.org/uc/item/8kj87750>

Author

Yao, Jason

Publication Date

2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Modeling and Measuring Dynamic Cellular Signaling States

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

in

Bioinformatics and Systems Biology

by

Jason Chi Hsuan Yao

Committee in charge:

Professor Roy Wollman, Chair
Professor Shankar Subramaniam, Co-Chair
Professor Gourisankar Ghosh
Professor Nan Hao
Professor Padmini Rangamani
Professor Scott Alan Rifkin

2017

Copyright

Jason Chi Hsuan Yao, 2017

All rights reserved.

The Dissertation of Jason Chi Hsuan Yao is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2017

DEDICATION

This work is firstly dedicated to my then 15-year-old self who came by himself to the States with much uncertainty of the future- Congratulations, your gamble is starting to pay off.

To myself in the parallel universe, who chose to play it safe and stayed in Taiwan, went to a good-enough college and is now working in a job he hates- You know you are better than this.

EPIGRAPH

Essentially, all models are wrong, but some are useful.

George E. P. Box

TABLE OF CONTENTS

SIGNATURE PAGE.....	iii
DEDICATION	iv
EPIGRAPH.....	v
TABLE OF CONTENTS	vi
LIST OF FIGURES	ix
LIST OF SCHEMES	x
LIST OF TABLES	xi
ACKNOWLEDGEMENTS	xii
VITA	xiii
ABSTRACT OF THE DISSERTATION	xiv
Chapter 1: AKUMA: A Toolbox for Single Cell Parameter Fitting of Biological Kinetic Models.....	1
Chapter 1.1: Abstract	1
Chapter 1.1.1: Background	1
Chapter 1.1.2: Results	1
Chapter 1.1.3: Conclusions.....	2
Chapter 1.2: Background.....	2
Chapter 1.3: Implementation	3
Chapter 1.3.1: Dataflow Overview.....	3
Chapter 1.3.2: Preliminary Setup	4
Chapter 1.3.3: Submit Jobs.....	4
Chapter 1.3.4: Algorithm Implementation	5

Chapter 1.4: Results.....	7
Chapter 1.5: Discussion	8
Chapter 1.6: Conclusion	9
Chapter 1.7: Availability and requirements	9
Chapter 1.8: Author Contributions	10
Chapter 1.9: Acknowledgements.....	10
Chapter 1.10: Figures.....	11
Chapter 1.11: References	16
Chapter 2: Distinct Cellular States Determine Calcium Signaling Response.....	17
Chapter 2.1: Abstract	17
Chapter 2.2: Introduction.....	17
Chapter 2.3: Results.....	21
Chapter 2.4: Discussion	27
Chapter 2.5: Materials and Methods.....	31
Chapter 2.5.1: Cell plating.....	31
Chapter 2.5.2: Preparation of calibration buffer	31
Chapter 2.5.3: Measurements of single-cell calcium response to an increase in extracellular ATP perturbation.....	31
Chapter 2.5.4: Longitudinal dual-stimulation experiment.....	32
Chapter 2.5.5: Model of calcium signaling.....	33
Chapter 2.5.6: Parameter estimation: pre-processing of calcium response	33
Chapter 2.5.7: Parameter estimation: calculating goodness of fit	34
Chapter 2.5.8: Parameter estimation: rejection sampling	34
Chapter 2.5.9: Parameter estimation: approximate Bayesian computation sequential Monte Carlo	35

Chapter 2.5.10: Parameter estimation: convergence test.....	36
Chapter 2.5.11: Clustering of the parameter ensemble	36
Chapter 2.5.12: Data availability	37
Chapter 2.6: Author contributions	37
Chapter 2.7: Conflict of interest	38
Chapter 2.8: Acknowledgements	38
Chapter 2.9: References	39
Chapter 2.10: Figures.....	42
Chapter 2.11: Schemes.....	56
Chapter 2.12: Tables.....	57

LIST OF FIGURES

Figure 1.1: Dataflow	11
Figure 1.2: Bayesian Optimization Algorithm Schematic Diagram	12
Figure 1.3: Simple Central Dogma Model	13
Figure 1.4: Results of Bayesian Optimization for A Synthetic Data of Central Dogma Model	14
Figure 1.5: 'Real' Synthetic Parameter vs. Fitted Parameters Distributions	15
Figure 1.6: Results of Single Cell Calcium Data Fitting	15
Figure 2.1: Calcium signal response heterogeneity	42
Figure EV2.1: Clustering results for cells based on only time-course data.....	44
Figure 2.2: Calcium model.....	46
Figure 2.3: Process of single-cell parameter fitting	47
Figure EV2.2: Quality of convergent single-cell model fitting	49
Figure 2.4: Clustering parameter distributions	51
Figure EV 2.3: Distribution of cell physical features among clusters	51
Figure 2.5: Mechanistic difference between clustered data.....	52
Figure EV 2.4: Boxplot of kinetic parameter distributions in the major clusters	54

LIST OF SCHEMES

Scheme 2.1: Rejection Sampling.....	56
-------------------------------------	----

LIST OF TABLES

Table EV 2.1: Model Parameters.....	57
-------------------------------------	----

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Roy Wollman for his support as the chair of my committee.

Chapter 1, in full is a work being prepared for publication with the title “AKUMA: A Toolbox for Single Cell Parameter Fitting of Biological Kinetic Models.” Jason Yao, Roy Wollman. The dissertation author is the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of “Distinct cellular states determine calcium signaling response” which was published in *Molecular Systems Biology*, 2016. Jason Yao, Anna Pilko, Roy Wollman. The dissertation author was the primary investigator and author of this paper.

VITA

2011 Bachelor of Science, University of Texas at Austin

2013-2017 Research Assistant, University of California, San Diego

2017 Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

Yao, J., & Wollman, R. (2017). AKUMA: A Toolbox for Single Cell Parameter Fitting of Biological Kinetic Models. **In Progress.**

Yao, J., Pilko, A., & Wollman, R. (2016). Distinct cellular states determine calcium signaling response. *Molecular Systems Biology*, 12(12), 894.

Handly, L. N., **Yao, J.**, & Wollman, R. (2016). Signal Transduction at the Single-Cell Level: Approaches to Study the Dynamic Nature of Signaling Networks. *Journal of Molecular Biology*, 428(19), 3669-3682.

Selimkhanov, J., Taylor, B., **Yao, J.**, Pilko, A., Albeck, J., Hoffmann, A., ... & Wollman, R. (2014). Accurate information transmission through dynamic biochemical signaling networks. *Science*, 346(6215), 1370-1373.

FIELDS OF STUDY

Major Field: Systems Biology

Studies in Computational Modeling of Signaling Pathway

Professor Roy Wollman

ABSTRACT OF THE DISSERTATION

Modeling and Measuring Dynamic Cellular Signaling States

by

Jason Chi Hsuan Yao

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2017

Professor Roy Wollman, Chair

Professor Shankar Subramaniam, Co-Chair

The recent progress of single cell quantitative microscopy produced a plethora of data that demonstrated cell to cell variability in signal transduction, which provoked many possible theories that sought to explain the phenomenon of cellular variability. In order to investigate the mechanism of cell to cell variability, mathematical modeling needs to be combined with experimental approach to gain systems-level understanding. However, until now there have been few studies which applied mathematical modeling to study single cell data because of the numerous practical and theoretical hurdles. Here I develop a computational toolbox specifically designed to perform single cell model fitting efficiently on a population scale by utilizing big data platform. The toolbox fits time series data using Bayesian parameter optimization that produces a parameter distribution for the mathematical model. The model fitting is carried out on the Google Cloud Platform which scales well with the increasing number of single cell data and allows for fitting a population of single cell data at fast pace and extreme low cost. I combine the same model fitting algorithm with single cell fluorescent microscopy experiment to study cell to cell variability in calcium signaling pathway. By analyzing the parameter distributions from single cell fitting using clustering methods, I discover distinct clusters within the parameter distributions of cell population that vary the most at IP3 receptor within the calcium signaling pathway. The theoretical findings are corroborated by subsequent experiments which demonstrated that IP3 receptors played significant role in defining calcium response identities in cells.

Chapter 1: AKUMA: A Toolbox for Single Cell Parameter Fitting of Biological Kinetic Models

Chapter 1.1: Abstract

Chapter 1.1.1: Background:

In recent years, we have seen rapid improvement in the single cell imaging and dynamic signaling data measurement. Despite developments in experimental technologies on numerous fronts, the progress in computational modeling remains relatively slow. One of the main challenges with computational modeling is that it is difficult to simulate the models with limited data on kinetic parameters. To overcome the lack of experimental kinetic parameter values, many computational methods have been developed to calibrate and optimize the fit of the parameters to the mathematical models. However, many of the existing methods are computationally intensive; in addition, few, if any, of the methods are built with single cell data in mind and do not scale well with fitting multiple sets of data to the model.

Chapter 1.1.2: Results:

In this study, we introduce AKUMA (**A**pproximate-Bayesian '**K**'loud-based **M**odel Fitting **A**pparatus), Python toolbox that employs Bayesian parameter optimization for mathematical model fitting. The toolbox is designed to utilize the Google Cloud Platform and allow users to easily scale up the model fitting with population of single cell data. The library accepts mathematical models in the standardized SBML format (Hucka *et al*, 2003) which ensures its shareability. The toolbox was tested on synthetic data generated for

simple gene transcription-translation model as well as on actual single cell experimental data for calcium signaling pathway. The results demonstrated that given reasonable fitness criteria, the toolbox was extremely economical, both in time and monetary cost, to produce good fits to the data.

Chapter 1.1.3: Conclusions:

With the high-performance capability of AKUMA, we expect the toolbox to be a significant contribution to the community and bring in a new era of computational systems biology that extends well beyond the population average model and into the realm of single cell study.

Chapter 1.2: Background

The field of systems biology relies on both the experimental and computational methods to test and refine quantitative hypotheses of causal relationships between components in a biological system (Chuang *et al*, 2010). The principles of systems biology have been applied to various areas of biological study and received great success. For example, in the area of cell signaling, the development of mathematical models for various canonical signaling pathways allow researchers to gain quantitative insights into cellular response (Huang & Ferrell, 1996; Hoffmann *et al*, 2002). In recent years, single cell measurement technologies have been gaining significant strides in areas such as single cell RNA-Seq and quantitative microscopy, fostering the need to extend computational biological modeling to single cell data (Elowitz *et al*, 2002; Patel *et al*, 2014). Conventionally, when fitting data to a model, due to the lack of actual measured kinetic parameter values, the parameterization was done through a series of calibrations and

optimizations (Moles *et al*, 2003; Kim *et al*, 2010; Aldridge *et al*, 2006). Many algorithms and software libraries have been developed to perform model fitting (Lopez *et al*, 2013; Egea *et al*, 2014; Liepe *et al*, 2014). However, the majority of the methods suffer from at least one of the two common issues. First, many of the methods are time consuming. In addition, many of the methods only optimize for a single best fit, which is not valuable when the mathematical model has high dimensions and sloppiness or potential non-identifiability. Both of those issues are amplified when users try to extend model fitting to single cell data, because not only do the algorithms need to scale well with fitting the model multiple times, the algorithms also need to address non-identifiability of the model to allow separation between different cell identities. Here we introduce AKUMA, a high-performance Python toolbox using Bayesian optimization algorithms and built within the Google Cloud Platform ecosystem. The toolbox addresses the concern of runtime by utilizing cloud computing, while the issue of sloppiness is addressed through Bayesian parameter optimization which estimates the uncertainty intervals of the kinetic parameters by producing a population of parameter vectors.

Chapter 1.3: Implementation

Chapter 1.3.1: Dataflow Overview:

The architecture of the dataflow of the toolbox consists of four main components as shown in Figure 1.1: data staging station for the user, Google Cloud instance group, Google Drive and the Git repository. The data staging station is the only component that the user will directly interface with and it will automatically handle the operation of all other components; the Google cloud instance group contains the Virtual Machine instances that

perform the actual data fitting to the model and transfer the results files to Google Drive while the staging station monitors Google Drive for finished results files to download from.

Chapter 1.3.2: Preliminary Setup:

After installing the required packages in the “Availability and requirements” section, the user will first clone the git repository of the library and then download the `client_secrets.json` file from the Google Service Account into the cloned local repository. The local repository will be the working folder. The user will then initialize the Google Cloud SDK so that the user account on the data staging station will have the proper authentication credentials to interface with Google Cloud Platform.

Chapter 1.3.3: Submit Jobs:

After setting up and configuring all the necessary software application, submitting data to be optimized is extremely simple. The user would provide files required for the running of algorithm, which include: 1. Input text file specifying the names of all other data files as well as tuning parameters of the fitting algorithm. 2. Array of unique ID's for data of each of the individual instances. 3. File of a data matrix which includes all the time course data to be fitted that correspond to the array of unique ID's. 4. Array of fitness scores convergence threshold 5. The SBML file of the mathematical model 6. The array of corresponding time points. The user will then execute the script *SubmitJobs.sh* which will proceed to create a Google Cloud Compute Engine instance group with the number of individual Virtual Machine (VM) instances as specified by the user. After the instance group is created, *SubmitJobs.sh* will copy the files over to the worker instances along with the indices into the data file for each of the individual VM instances in the group. For example, if the user specifies n instances to be created, *SubmitJobs.sh* will initialize model

fitting procedures for the first n time course data. Each of the VM's in the group will initialize by installing necessary packages as well as the git repository of the tool box. After finishing initialization and installation, each of the virtual machine instances will begin running Bayesian parameter optimization to fit their respectively assigned single cell data to the mathematical model. As soon as optimization in a VM is finished, the individual VM will upload the fit result data to the Google Drive associated with the Google service account to be download by the data staging station that the user interfaces with, thereby completing the dataflow.

Chapter 1.3.4: Algorithm Implementation:

The main implementation to optimize parameters is divided into the following classes: *abcsmc*, *simulationandscoreing*, *parametersampling* and *sampleweighting*. The class *abcsmc* implements two inner classes *FittingData* and *ModelFitting*, as well as two methods *RejectionSampling* and *SequentialMonteCarlo*. The class *FittingData* encapsulates all the results associated with the optimization, such as parameter ensembles and simulated data, along the iterations of the algorithm. The class also writes the data to format readable with standard Python library. The *ModelFitting* class encapsulates a *FittingData* along with the information regarding the tuning parameters of the parameter fitting algorithm, such as the score thresholds for convergence along each of the iterations, goodness of fit as well as parameters for the model simulation solver. The class method *SequentialMonteCarlo* is the main workhorse of the toolbox which carries out the entire Approximate Bayesian Computation Sequential Monte Carlo (ABC-SMC)(Toni *et al*, 2009) algorithm as described in Figure 1.2. The algorithm consists of filtering the parameter distributions through a sequence of rejection sampling that would gradually optimize the distributions to reach the target final convergence criteria. Within a

given iteration, the class method *RejectionSampling* repeatedly draws random parameter vectors according to a given prior sampling function and decides to either accept or reject the parameter vectors based on a given fitness scoring function combined with convergence criteria. Once the collected population of parameter vectors reaches the convergence criteria, the rejection sampling will terminate and return the collected parameter vector population as the estimate of its posterior distribution. The posterior from one rejection sampling will then become the input prior distribution for the next iteration of rejection sampling.

The user has a number of options to determine the convergence criteria for the sequence of rejection sampling procedure. For example, the user can set a sequence of fitness score thresholds, according to which the rejection sampling has to terminate at each iteration. Alternatively, the user can also set a single or a sequence of relative fitness score thresholds rather than based on actual scores. For example, the user can elect that for each of the iterations that the algorithm only collects parameter vectors that score in a given top percentile compared to the prior population, thus alleviating the need for the user to pre-define the actual values of the convergence thresholds. Once the fitness of a rejection sampling procedure posterior reaches the final score threshold, the entire Sequential Monte Carlo algorithm will terminate, although the user also has the flexibility to elect the fraction of posterior that would need to reach the final fitness threshold for Sequential Monte Carlo to terminate.

The other three classes, *simulationandscoreing*, *parametersampling* and *sampleweighting* exist to help tune the Sequential Monte Carlo algorithm. The class *simulationandscoreing* contains an inner class *FncOutput* which encapsulates the output of model simulation such as the time course data of state variables and the observable. The

simulationandscore class also contains a collection of methods that simulate the model, score the model, and evaluate the model for acceptance into posterior given the input kinetic parameter values. The class *parametersampling* contains a collection of methods that construct sampling function based either on prior specified distribution or from density estimation of a given parameter vector ensemble. The class *sampleweighting* includes methods to assign sampling weights to each of the particles in a population of parameter vectors based on either a pre-specified probability distribution or on density estimation of the given population.

Chapter 1.4: Results

The effectiveness of the library was first tested on a simple gene transcription and translation mathematical model which is governed by a system of linear differential equations as described in Figure 1.3. A randomized parameter vector was generated in order to simulate a time course of synthetic data of protein amount time series. The generated synthetic data was fit to the model using the toolbox to evaluate the efficiency of model fitting to produce the results shown in Figure 1.4. The optimization itself finished in 34 minutes and the final posterior formed a reasonable estimate of the uncertainty interval for the kinetic parameters. From the select pairwise scatter plots of the parameters the parameters are shown to be increasingly congregated along manifolds as the optimization progresses, and it shows how the Bayesian fitting algorithm could be used to investigate the sloppiness of the model and the inter dependence of the parameters on model fitting. To further investigate the performance of the model with a population of single cell data, synthetic data was generated for 39 single cells and fit to the model successfully. Figure 1.5 shows the correspondence between the true values of the

parameter α and their related distributions of fitted values in the posterior of each single cells. The results show that the method was generally robust over a range of synthetic data.

In order to assess the usefulness of the library on realistically complex models, the single cell calcium data and mathematical model from a previous study (Yao *et al*, 2016). were used to perform model fitting. The results of the fitting over the population of calcium single cell data are shown in Figure 1.6. For just over 250 single cell data, the plot shows that when using the computational infrastructure and methods in the library, it would take only less than 30 cents to fit the data. Taking into account all the overhead time including dynamically setting up all the computational infrastructure and handling all the data transferring, it would take under 45 minutes to fit all the data. The histogram plot also shows that for the overwhelming majority of the calcium single cell data, the actual running of the Bayesian fitting algorithm takes less than 10 minutes. In summary, the computational methods and infrastructure described in this project allows the user to carry out single cell data fitting reliably at short time scale and extremely low cost.

Chapter 1.5: Discussion

The library we have introduced in this study utilizes state-of-the-art Bayesian parameter optimization methods with Google Cloud Platform which accepts the user supplied mathematical model in the SBML format, which allows users to perform computational experiments that are easy to implement and readily sharable. While there are already multiple software tools which allow for biological model fitting using either probabilistic or deterministic optimization scheme, those tools generally do not have built in integration with big data infrastructure or are difficult to implement. The method we

introduce in this study has built-in integration with Google's commercial infrastructure which is highly sustainable that scales well with population of single cell time course data, thus ensuring the longevity of the library. In addition, the toolbox requires minimal customization of the working station of the user, since the actual optimization is carried out in the Google Cloud Platform, which makes the implementation very easy to use and largely independent of the specification of the user working station.

Chapter 1.6: Conclusion

In this study, we introduce an effective model fitting toolbox that carries out efficient Bayesian parameter optimization of the model in the standardized SBML format. The toolbox has been tested on both a simple gene transcription-translation model as well as in high dimensional nonlinear cell signaling model and is able to produce well fitted parameter values at fast time scale and extremely low cost. With the rapid advancement of single cell measurement technologies, we are confident that the tool we present here will have much use to uncover the mechanistic insights through single cell model fitting.

Chapter 1.7: Availability and requirements

Currently the toolbox has been tested to work in Linux 14.04 LTS

Working internet browser

Python 2.7+

Pip(<https://pip.pypa.io/en/stable/installing>)

Oauth2client (<https://oauth2client.readthedocs.io/en/latest/>)

PyDrive (<https://pythonhosted.org/PyDrive/>)

Git (<https://git-scm.com/book/en/v2/Getting-Started-Installing-Git>)

Google Cloud SDK (<https://cloud.google.com/sdk/downloads>)

Google service account (<https://developers.google.com/api-client-library/python/auth/service-accounts>)

Chapter 1.8 Author Contributions

JY and RW conceived of the study. JY implemented the algorithm and performed all necessary computational studies as well as subsequent statistical analyses. JY and RW wrote the manuscript.

Chapter 1.9 Acknowledgements

The work was supported by GM111404 & EY024960 (RW), and a training grant (GM007240) for LNH. Chapter 1, in full, is a work being prepared for publication with the title “AKUMA: A Toolbox for Single Cell Parameter Fitting of Biological Kinetic Models.” Jason Yao, Roy Wollman. The dissertation author was the primary investigator and author of this paper.

Chapter 1.10: Figures

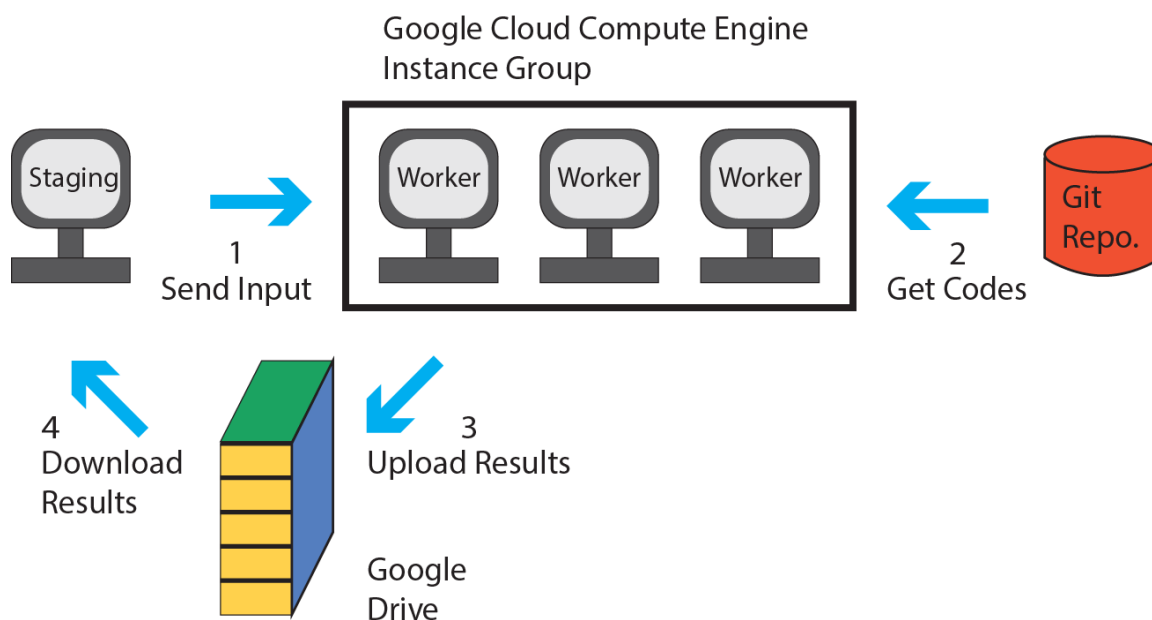


Figure 1.1. Dataflow. There are four components in the entire workflow: The data staging station, Google Cloud Compute Engine Instance Group, Git Repository of code and Service Account Google Drive. In the first step, the data staging station executes the command to create GCE instance group and then transfers necessary data to the instance group; in the second step, as the GCE instance group initializes its setup, it downloads the AKUMA Toolbox from the git repository to execute the Bayesian parameter optimization. As soon as the individual GCE instances are finished, they upload the results files to the Google Drive linked to the service account. From the Google Drive, the result files are downloaded back into the data staging station.

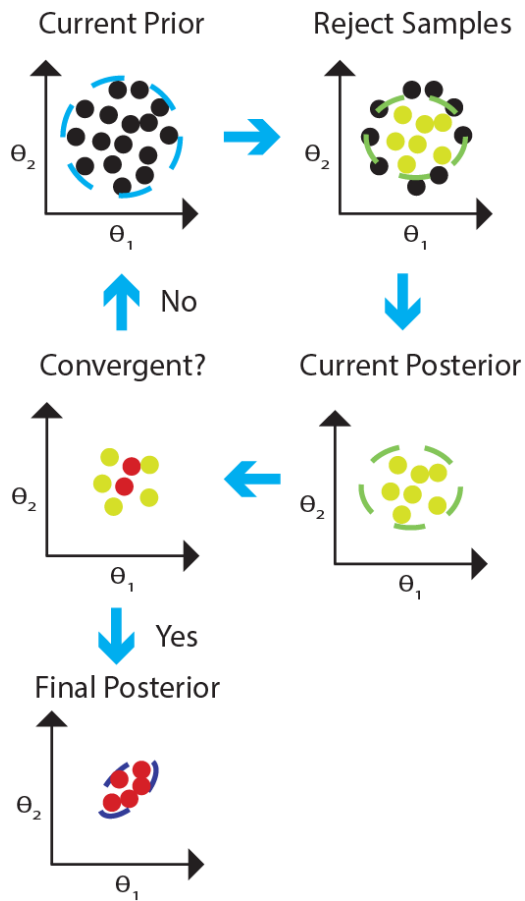


Figure 1.2. Bayesian Optimization Algorithm Schematic Diagram. The optimization algorithm consists of a progression of rejection samplings which iteratively refines the parameter vectors. In each of the rejection sampling rounds, the algorithm filters the parameter vectors through increasingly more stringent cutoff thresholds until the parameter populations converge under a final cutoff threshold.

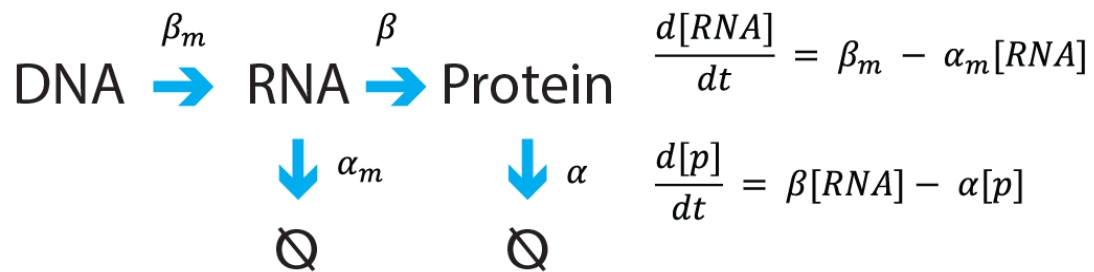


Figure 1.3. Simple Central Dogma Model. The model has two state variables: RNA and protein (p). The rate of change of RNA is determined by a zero order RNA production rate(β_m) with a first order RNA degradation rate(α_m). The protein rate of change is determined by a first order rate constant (β) and a first order protein degradation rate (α).

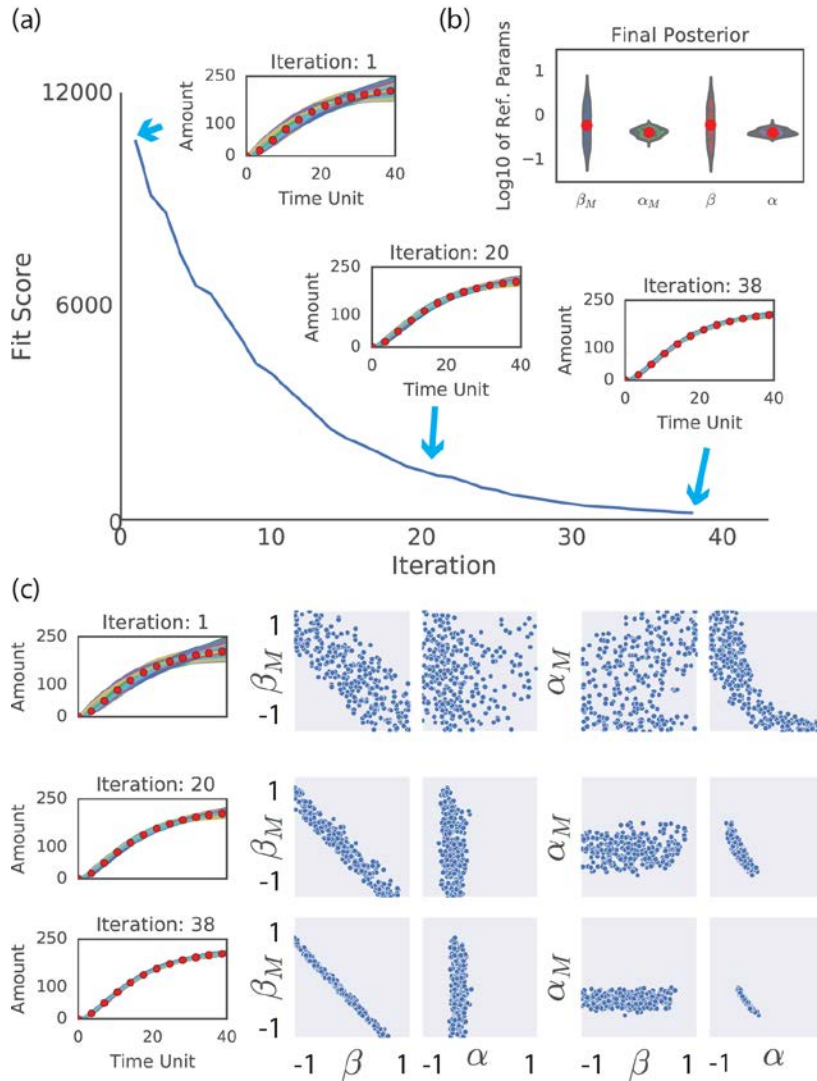


Figure 1.4. Results of Bayesian Optimization for A Synthetic Data of Central Dogma Model. (a) The algorithm terminated successfully within 38 iterations of rejection samplings. The fit score between two time series is based on the combination of two quantities: the sum of squares of the difference between the two time series and the sum of squares of the difference between the vectors of adjacent differences of the two time series. For the optimization, the final fitting threshold is determined by introducing a white noise to the synthetic data and scoring the fitness between the synthetic data and noisy data, so that the simulated data cannot fit worse than the noisy data. The number of parameter vector population in each of the iterations is 300. (b) The violin plots of the final posterior. The parameter values are displayed as the inverse log of 10 of the ratio of simulated parameter to reference parameter. The reference values for the four kinetic parameters are [100, 0.5, 0.1, 0.2]. (c) Plot of simulated data vs synthetic data and pairwise scatterplots of selected iterations. The axis of the scatterplots is selected the same way as in (b).

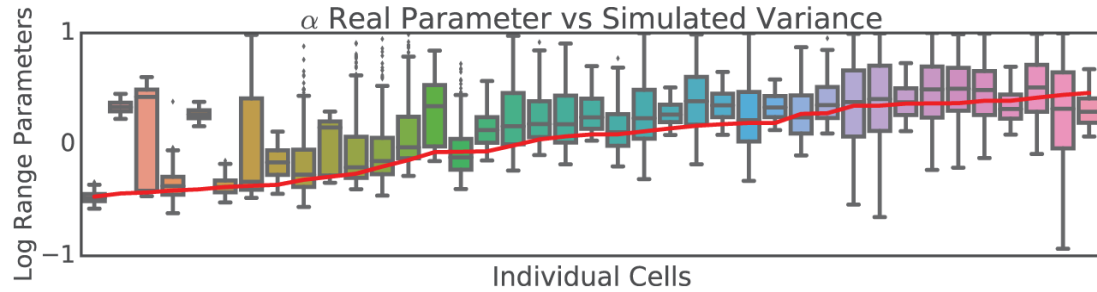


Figure 1.5. 'Real' Synthetic Parameter vs. Fitted Parameters Distributions

The comparison between the actual values of α vs distributions of fitted values of α for 39 synthetically generated single cell data. The red line shows the real values of α that are used to generate the population of single cell data, whereas the boxplots are the distribution of the values of α from the fitted simulated data.

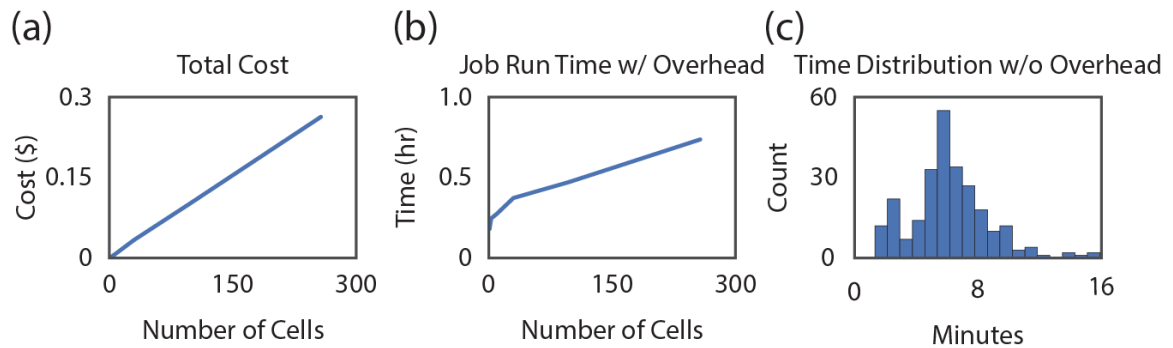


Figure 1.6. Results of Single Cell Calcium Data Fitting. Fit the data to real actual experimental data with complex nonlinear high dimensional signaling model. (a) The total cost in dollars vs. the number of single cell data fitted. (b) Total run time of single cell fitting that includes overhead (setting up GCE instances, uploading finished results and downloading results). (c) Histogram of run time of actual fitting algorithm of individual single cell data without overhead.

Chapter 1.11: References

- Aldridge BB, Burke JM, Lauffenburger DA & Sorger PK (2006) Physicochemical modelling of cell signalling pathways. *Nat. Cell Biol.* 8: 1195–1203
- Chuang H-Y, Hofree M & Ideker T (2010) A decade of systems biology. *Annu. Rev. Cell Dev. Biol.* 26: 721–744
- Egea JA, Henriques D, Cokelaer T, Villaverde AF, MacNamara A, Danciu D-P, Banga JR & Saez-Rodriguez J (2014) MEIGO: an open-source software suite based on metaheuristics for global optimization in systems biology and bioinformatics. *BMC Bioinformatics* 15: 136
- Elowitz MB, Levine AJ, Siggia ED & Swain PS (2002) Stochastic gene expression in a single cell. *Science* 297: 1183–1186
- Hoffmann A, Levchenko A, Scott ML & Baltimore D (2002) The I κ B-NF- κ B Signaling Module: Temporal Control and Selective Gene Activation. *Science* 298: 1241–1245
- Huang CY & Ferrell JE Jr (1996) Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc. Natl. Acad. Sci. U. S. A.* 93: 10078–10083
- Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr J-H, Hunter PJ, et al (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19: 524–531
- Kim KA, Spencer SL, Albeck JG, Burke JM, Sorger PK, Gaudet S & Kim DH (2010) Systematic calibration of a cell signaling network model. *BMC Bioinformatics* 11: 202
- Liepe J, Kirk P, Filippi S, Toni T, Barnes CP & Stumpf MPH (2014) A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nat. Protoc.* 9: 439–456
- Lopez CF, Muhlich JL, Bachman JA & Sorger PK (2013) Programming biological models in Python using PySB. *Mol. Syst. Biol.* 9: 646
- Moles CG, Mendes P & Banga JR (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res.* 13: 2467–2474
- Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL, Louis DN, Rozenblatt-Rosen O, Suvà ML, Regev A & Bernstein BE (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344: 1396–1401
- Toni T, Welch D, Strelkowa N, Ipsen A & Stumpf MPH (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* 6: 187–202
- Yao J, Pilko A & Wollman R (2016) Distinct cellular states determine calcium signaling response. *Mol. Syst. Biol.* 12: 894

Chapter 2: Distinct Cellular States Determine Calcium Signaling Response

Chapter 2.1 Abstract

The heterogeneity in mammalian cells signaling response is largely a result of pre-existing cell-to-cell variability. It is unknown whether cell-to-cell variability rises from biochemical stochastic fluctuations or distinct cellular states. Here, we utilize calcium response to adenosine trisphosphate as a model for investigating the structure of heterogeneity within a population of cells and analyze whether distinct cellular response states coexist. We use a functional definition of cellular state that is based on a mechanistic dynamical systems model of calcium signaling. Using Bayesian parameter inference, we obtain high confidence parameter value distributions for several hundred cells, each fitted individually. Clustering the inferred parameter distributions revealed three major distinct cellular states within the population. The existence of distinct cellular states raises the possibility that the observed variability in response is a result of structured heterogeneity between cells. The inferred parameter distribution predicts, and experiments confirm that variability in IP3R response explains the majority of calcium heterogeneity. Our work shows how mechanistic models and single-cell parameter fitting can uncover hidden population structure and demonstrate the need for parameter inference at the single-cell level.

Chapter 2.2 Introduction

Cell-to-cell variability in dynamic responses to stimuli is observed ubiquitously (Geva-Zatorsky *et al*, 2006; Cohen-Saidon *et al*, 2009; Tay *et al*, 2010; Selimkhanov *et al*,

2014), yet its underlying causes are still unknown. Previous work decomposed biological processes in gene expression and signal transduction into intrinsic fluctuations and pre-existing variability (Elowitz *et al*, 2002; Rhee *et al*, 2014; Selimkhanov *et al*, 2014). Specific measurement of the relative contribution of stochastic fluctuations during signal transduction compared to cell-to-cell variability showed that 50–98% of the variability could be attributed to pre-existing cell-to-cell variability (Toettcher *et al*, 2013; Selimkhanov *et al*, 2014). While cell-to-cell variability was determined to be the dominant source of heterogeneity in cellular response, the cause of cell-to-cell variability is still unknown. Two hypotheses can explain the observed cell-to-cell variability: The first is that variability is a result of accumulation of stochastic fluctuations in gene expression (Shibata & Fujimoto, 2005; Ladbury & Arold, 2012; Rhee *et al*, 2014), organelle composition (Oates, 2011), and other cellular factors where small numbers of molecules increase biochemical randomness. Under this hypothesis, a clonal population will accumulate changes over time that would limit biochemical reaction accuracy. The second hypothesis is that the population of cells contains several deterministic cellular states, according to the surrounding microenvironment, or due to functional differences between cells, in which case cells would react in predictable patterns of behaviors (Snijder *et al*, 2009). While these two hypotheses seem contradictory, they are not mutually exclusive and each of these hypotheses can explain a different amount of the total observed variance. One approach that was used to test the different contributions of these two mechanisms is to predict the observed response variability based on measurements of large arrays of features for each cell (Snijder *et al*, 2009). However, it is possible that the measured features themselves include information that is indicative of the accumulation of the stochastic fluctuation. Therefore, while the measured features have high predictive power, it does not provide an indication of distinct functional states. Another approach to identify

functional cell state is based on a dynamical systems point of view, which argues that a cell can be represented as a dynamical system and that each cell state is an attractor within the cell state space (Furusawa & Kaneko, 2012). The hallmark of this multiple-attractor view is the existence of distinct clusters within cell state space. While distinct clusters are indicative of multiple distinct cell states, not every functional heterogeneity will show distinct clusters. It is possible that because cells need to comply with multiple objectives, they occupy a Pareto front and therefore functional heterogeneity will not necessarily cause distinct cell states (Sheftel *et al*, 2013). Nonetheless, the existence of distinct cell states can provide support for the existence of attractors and therefore the possibility that structured heterogeneity contributes to the observed cellular variability.

The true complexity of cellular state space is unknown and can only be probed by examining specific signatures that indicate the underlying cell state. There are numerous forms for possible signatures, such as transcript abundance and protein concentrations. Many of these signatures have thousands of dimensions, and therefore, dimensionality reduction approaches are routinely used to reduce these large spaces into fewer dimensions (Buettner *et al*, 2015). However, statistical methods on their own do not provide any mechanistic insights toward the meaning of any identified clusters. An additional limitation with statistical dimensionality reduction approaches is that they do not take into account the substantial body of prior knowledge on the underlying biochemical mechanisms which contribute to the variation. The kinetic parameters of the reactions within the signaling network are an alternative and natural cell state signature. Kinetic parameters are aggregates of molecular activities from the genetic level upwards that provide a functional representation of cell-to-cell variability in the dynamics of signal transduction pathways. By employing differential equation modeling, we can simulate the dynamic output of each signature and connect the underlying state variability with the

observed response variability. Therefore, kinetic parameters could provide a useful signature of cellular state that could be used to determine the underlying distribution of cell heterogeneity.

Determining cellular kinetic parameters is challenging, because experimental measurement of kinetic parameters of biochemical reactions *in vivo* with high accuracy is extremely difficult. An alternative approach is to infer the kinetic parameter values from cellular responses (Lee *et al*, 2009). Parameter inference allows for the learning of parameter values for a differential equations model of signal transduction based on measured time-series data of state variables in the model (Eydgahi *et al*, 2013). Parameter inference of signal transduction networks is a difficult problem due to the large number of state variables and parameter values which would create non-identifiability issues (Raue *et al*, 2011) or challenges from model “sloppiness” where multiple sets of parameters could fit the data equally well (Gutenkunst *et al*, 2007). Thus, to utilize differential equation modeling to make meaningful inference about hidden cell states, it is important to account for the uncertainty associated with parameter inference. This could be done through the use of Bayesian inference to infer the full posterior parameter distribution for each parameter. The full distributions provide information on the accuracy of the inference and report on existence of any “sloppy” eigenvectors of parameters within goodness-of-fit space.

Through the use of full distributions of kinetic parameters as cellular signatures, it is possible to glean insights into cellular state distribution. Here, we inferred cell states of signaling dynamics by fitting a population of single-cell calcium trajectories to a mathematical model and produced a distribution of parameter ensembles. We first measured cellular calcium responses to adenosine triphosphate (ATP) stimulations using

fluorescence microscopy and MATLAB image processing to obtain single-cell calcium responses. The single-cell data were fitted to a system of differential equations of the calcium signaling pathway using Bayesian inference to obtain the posterior distribution of parameter values vectors within a certain level of tolerance for each of the fitted trajectories. We analyzed the distribution of parameter ensembles using the Kullback–Leibler divergence criteria and found three major distinct clusters of parameter ensembles. The identification of major clusters indicates the possibility that cell-to-cell variability could be interpreted not as a simple variability that stems from accumulation of stochastic fluctuations but rather from the existence of multiple distinct states within the population.

Chapter 2.3 Results

We measured calcium responses of a non-tumorigenic human mammary epithelial cell line (MCF10A) to stimulation with 10 μ M ATP using the Fluo-4 calcium indicator calibrated by using an established protocol (Bao *et al*, 2010). The fluorescence images were segmented, and cells were identified and tracked over time to measure single-cell dynamic calcium responses. Figure 2.1 shows the response matrix (Fig 2.1A), population average (Fig 2.1B) and a few representative examples (Fig 2.1C) of single-cell calcium responses to ATP. As was previously reported for this system (Selimkhanov *et al*, 2014), cellular responses were highly heterogeneous. We analyzed the following response dynamics: response in contrast with basal level before stimulation, latency to reach half maximal level, latency to reach maximum level, maximum level, time from maximal level to half maximal, and steady state after stimulation (Fig 2.1D) (Cohen-Saidon *et al*, 2009). The results showed wide heterogeneity with respect to the trajectory features (Fig 2.1E). The distribution of cells with similar trajectory features were not spatially correlated (Fig

2.1F and 2.1G). We initially attempted to cluster the single-cell calcium trajectories based on either the entire calcium responses or through their representative time-series features. The obtained clusters did not identify clear response archetypes (Fig EV2.1) potentially as a result of the fact that the ratio of variance between clusters to the intracellular variance was low (Fig EV2.1D). Therefore, while clustering based on time-series data is a valid and useful approach, in this case it did not lead to identification of cell states. To further explore the possibility that cell states exist, we hypothesize that insights regarding cellular states could be uncovered by studying the mechanistic structure of pathway using mathematical modeling that incorporate prior knowledge on the underlying signaling network.

We used the mathematical model of calcium signaling from Lemon *et al* (2003) and Li and Rinzel (1994) as a template to construct a model of the calcium signaling response. The model was simplified to improve parameter identifiability. Overall, the mathematical model included 17 parameters and four state variables. Several simplifications and assumptions were included as follows: there is no mitochondrion; the model does not include plasma membrane leakage or store operated channel for calcium therefore total cellular calcium is kept constant. While the model does not capture all the known biochemical details, it is simplified to attenuate model redundancy and increase model identifiability. The schematic diagram of calcium signaling is shown in Fig 2.2. According to our model, when the ligand ATP binds to the purinergic receptor, the binding action triggers a series of events leading to the activation of enzyme phospholipase C (PLC). PLC then hydrolyzes phosphatidylinositol 4,5-bisphosphate (PIP2) to produce inositol 1,4,5-trisphosphate (IP3) and diacylglycerol (DAG). IP3 activates the calcium channels on the endoplasmic reticulum (ER), which would release calcium from ER into the cytosol, producing a calcium spike. The sarco/endoplasmic reticulum Ca^{2+} -ATPase

(SERCA) channel on the ER then pumps the calcium in the cytosol back into the ER, thus completing the transient calcium spike (Lemon *et al*, 2003).

To uncover the functional signatures of cell states, we inferred the kinetic parameter distributions for each of the single-cell trajectories using a Bayesian algorithm. The entire workflow, including raw data signal processing, data fitting, convergence testing, and sampling parameter distributions, is outlined in Fig 2.3A. Initially, the data were processed to remove experimental noise and other high-frequency elements of calcium response that were not captured by the model. We found that pre-processing the data increased the quality of the identified model fits, likely because it effectively reduced the stochastic and technical noise. The square of the sum of difference between the processed and raw time-course calcium data was later used to determine the rejection cutoff of a plausible model. For parameter inference, we used a variant of approximate Bayesian computation-based sequential Monte Carlo (ABC-SMC) method for our single-cell trajectory fitting (Liepe *et al*, 2014). In the Bayesian framework, we inferred parameter values for the system of differential equations in terms of their posterior distributions (Liepe *et al*, 2014) corresponding to single-cell calcium responses. At the core of our method is the ABC rejection algorithm, which arrives at the posterior distribution through simulating the differential equation model, calculating the agreement between simulation results and the experimental data and rejecting parameter values that have low agreement with the data. The key to efficient ABC sampling is using a good proposal distribution such that the fraction of rejection parameters is low. The ABC-SMC algorithm iteratively updates the proposal distribution through repeated ABC sampling where a proposal distribution (i.e., prior) of one iteration is based on the posterior of the previous iteration. In each step, the proposal distribution and the rejection cutoff are updated to balance the need to explore

unseen regions of parameter space with the need to exploit regions that provide good fits. The iterations continue until the parameter distribution converges to within the experimental error that is determined through low-pass filtering of the data. During sampling iterations, the region that the parameter vectors occupy converges. An example of the projection of the parameter distributions from multiple iterations on 2D space is shown in Fig 2.3B. The parameters were sampled in log scale with the initial prior distribution ranging over two orders of magnitude and centered on the referenced values that were chosen based either on the literature or our estimates of biologically reasonable ranges.

A key challenge in any optimization in a large parameter space is the tendency to identify only local solutions. This is especially a concern due to the probabilistic nature of our search algorithm. To verify that our search algorithm samples correctly from the target distribution, we used strict convergence criteria. We tested for convergence by fitting each of the single-cell responses twice. This means that one single-cell trajectory would have two populations of parameter vectors run #1 and run #2 which represented the posterior parameter distributions associated with the two independent runs of the algorithm. We reasoned that if the two posterior populations were from the same distribution, then the nearest neighbors of the parameter vectors from those two runs should come from either population with equal probability. This was the null hypothesis of our test. To test whether the nearest neighbor probability that we calculated shows a statistically significant difference from the null hypothesis of equal probability, we used a chi-square test. The two runs were considered to be convergent if the test failed to reject the null hypothesis. To demonstrate the overall parameter convergence, Fig 2.3C shows the multidimensional scaling projection onto a 2D plane from two sets of fitting to trajectories, one is convergent and the other non-convergent. Note that the 2D projection is presented for visualization

purposes only and the test was conducted in a 17-dimension parameter space. The result of the algorithm for a single trajectory produced a population of parameter vectors representing the individual parameters in terms of their confidence intervals (Fig 2.3D). We successfully identified 672 cellular responses with corresponding convergent parameter distributions (Fig EV2.2).

To determine whether distinct cell states exist, we clustered the identified parameter distributions for each of the cells (Fig 2.4). The key step in any cluster analysis is to determine a distance between any two cells in the cell state space. To take into account the multidimensional confidence intervals provided by the parameter distribution, we used a Kullback–Leibler (KL) divergence measure, an information theoretic measure that determines the similarity between parameter ensembles. For every two cells, we constructed a probability distribution of the likelihood of nearest neighbor parameter vector to be from the same cells. The distance measure was defined as the KL divergence between the constructed distribution and a null model where the parameter vector sets for the two cells are fully mixed. Combining the full probability distribution and the use of KL divergence resulted in a distance measure that integrates information on the difference between parameter values and on the confidence we have in that difference. Given the KL divergence distance measure, we used standard hierarchical cluster analysis with average linkage. The clustering results from the analysis showed three major distinct cell clusters. The cluster analysis based on parameter values showed ~150-fold better cluster separation comparing to clustering based on calcium response directly (Fig EV2.1C–E). The identification of these clusters suggests the existence of three distinct cell states within the population. These cell states are not dependent on cell cycles, cell size, and several other cellular morphology features (Fig EV2.3).

The use of kinetic parameters as signatures for cellular states enabled direct interpretation of mechanistic differences between the identified clusters. To examine how different parameters contributed to the clustering, we investigated the difference between the individual parameters among the three main clusters and how they related to the mechanism of calcium signaling. Figure 2.5A illustrates the degree of difference between clusters along different components of the pathway. We identified that the key parameters that separate these classes were related to the strength of positive and negative feedback between calcium and the IP3R channel, as well as the sensitivity of the IP3R channel to IP3 (Fig 2.5B). In our model, the IP3 channel is composed of three components: the IP3-activated subunit, the calcium-activated subunit, and the calcium-inactivated subunit. The significant parameters corresponded to the dissociation constants for these three subunits. Calcium has both activating and inhibiting effects on IP3R channels (Li & Rinzel, 1994). We found that the high magnitude of calcium responses in a cluster corresponded to stronger activation and lower inhibition. Overall, the three clusters of parameter distributions were separated into a cluster with strong positive calcium feedback, a cluster with weak positive calcium feedback, and a third with strong negative calcium feedback (Fig 2.5C). As expected, each of the identified feedback structure had distinct temporal features associated with it (Fig EV2.4B). Our analysis makes a specific prediction that majority of cellular response variability is determined by IP3R activity. To test this prediction, we performed longitudinal experiment where cells were first exposed to extracellular ATP followed by increase in intracellular IP3. Intracellular increase in IP3 was achieved by 365 nm uncaging of preloaded caged IP3 (Materials and Methods). Our experiments show that ~70% of variability in magnitude of calcium response to ATP can be explained by response to intracellular IP3 (Fig 2.5D).

Chapter 2.4 Discussion

We employed an inference-based approach to gain insights into the structure of cellular state distributions. While cell-to-cell variability in cellular signaling responses is commonly observed, little is known about the statistical distributions of cellular responses. By fitting single-cell calcium responses to a mechanistic model using a Bayesian algorithm, we were able to gain insights into the structure of cell state distribution. By analyzing the distributions of the identified kinetic parameters, we found that cellular state distribution for calcium signaling was best approximated by a mixture model of three distinct clusters, represented by parameter distributions. Such characterization of cellular state space distribution is a fundamental step toward further work aimed at understanding the determinant of the observed signaling variability and its functional role.

Previous work proposed two possible hypotheses that can explain cell-to-cell extrinsic variability (Shibata & Fujimoto, 2005; Snijder *et al*, 2009; Ladbury & Arold, 2012; Rhee *et al*, 2014). The first hypothesis states that the variability is due to the accumulation of stochastic differences in gene expression, organelle structure, and other cellular processes that cause cells to drift randomly in the cell state space. According to this hypothesis, cellular variability is best described using an unstructured population distribution. The second hypothesis states that cells converge onto certain specific attractors in the cell state space instead of drifting randomly. Under this hypothesis, the basis for differences between cells depends on the different functional needs of the cell population instead of the stochastic fluctuations in gene expression. In other words, the second hypothesis states that cellular heterogeneity has a definite structure. Our approach of single-cell data model fitting showed that the kinetic parameters which define signaling are composed of a mixture of parameter distributions, supporting the second hypothesis.

The different clusters of parameter distributions correspond to different fundamental signaling behaviors. The identification of these clusters raises interesting questions such as what is the physiological significance of the identified clusters? And what are the molecular mechanisms that enable the identified functional diversity?

In all likelihood, the structure of the cell state distribution is not universal but context dependent. In our work, we utilized the calcium response of cells to extracellular ATP as a model to investigate cell-to-cell variability. Extracellular ATP is a damage-associated molecular pattern (DAMP) that is released in response to wounds to activate neighboring cells (Handly *et al*, 2015). Response to ATP is one of the first steps in a very complex process of wound response and healing. It is plausible that during wound healing, there are different functional requirements for epithelial cells and therefore the observed variability is possibly a result of the need to generate functional diversity to enable multiple distinct roles (Sonnemann & Bement, 2011). Future work is needed to determine the degree of functional heterogeneity of epithelial cells during wound healing and whether any physiological functional diversity indeed corresponds to the identified clusters.

The functional diversity we identified can be mechanistically explained by changes in a few parameters within the calcium signaling pathway. IP3R channels on the ER play a key role in calcium signaling and have been shown to be responsible for the large diversity of calcium dynamic signaling responses to large array of ligands. Furthermore, IP3R channels are subjected to positive and negative feedback regulation by calcium (Lemon *et al*, 2003). In our model fitting results, the most notable variation in parameters among the three clusters of cell states is those in the parameters that govern the affinity of IP3R channels to both IP3 and the calcium ion. This key prediction was validated experimentally to show that indeed response to IP3 is the key difference between cells

and by itself can explain 70% of the observed variability. Multiple mechanisms can cause such parameter differences including differences in expression of different subtypes of the IP3R (Wojcikiewicz, 1995) or in the structure of the ER that have an important role in the spatial distribution of IP3R channels (Meldolesi & Pozzan, 1998). What causes different cells to have different IP3R responses is still unclear. Like many cases of “symmetry breaking”, it is possible that separation of cells into the three clusters of IP3R activity starts with stochastic events that get stabilized to create the emergence of three distinct populations. In addition to differences in IP3R regulation, a few other parameters showed differences between the three clusters such as receptor activity level (Fig EV2.4A). The “negative feedback” cluster had the highest level of receptor activity. However, in this cluster, due to the strong negative feedback, the low response was “robust”, that is, insensitive to receptor levels. Therefore, the parameter levels of the receptor have minor effects on cellular response. This result exemplifies the need to analyze parameter sensitivity at the single-cell level because the sensitivity of receptor activity parameters varies in different cells based on their cluster identity.

We used an inferred kinetic parameter signature to define cell state and the subsequent analysis of these signature distributions indicated that the cellular population is composed of a mixture of three distinct cell states. Other signatures for cell state have been used in the past including single-cell genomics (Trapnell, 2015), single-cell cytokine secretion (Lu *et al*, 2015) and mass cytometry (Spitzer *et al*, 2015). The identification of distinct subpopulations is common across all these different definitions of cellular signature. Each one of these cellular signatures only captures some aspect of true cell state, and the recurrence of a complex population structure indicates that the high level of cellular variability is a result of multiple cell states that coexist in a population. In addition, our work and that of others suggest that these subpopulations are not static but dynamic

states among which cells can potentially transition (Bendall *et al*, 2014; Marco *et al*, 2014; Durruthy-Durruthy & Heller, 2015; Setty *et al*, 2016). It will be interesting to apply a similar trajectory analysis to other kinetic parameter signature datasets to track how cells respond to ligand changes over time and to what degree these changes are reflected by transitions between the identified distinct cell states.

Prior to this work, the predominant method for fitting models to signaling pathways has been to fit only the population average and come up with one best fit parameter vector. The prevailing assumption of this traditional method is that the population average data are a good representation of the overall behavior of the cell population, and therefore, the best fit parameter vector is also the best representation of the cell population at the mechanistic level. While experimental work has consistently demonstrated the limitations of population averaging, mathematical modeling has lagged behind by treating of cells in the population equally and modeling only the average behavior. Difficulties in parameter inference due to model non-identifiability (Raue *et al*, 2011) or parameter “sloppiness” (Gutenkunst *et al*, 2007) further limit the ability to perform such parameter fitting at the single-cell level. Our work demonstrates that fitting parameters at the single-cell level are useful because it prevents information loss due to population averaging and enables the discovery of distinct subpopulations. The improved performance is despite the fact that, at least in some of the parameters, the confidence interval of the identified parameter values is quite large and therefore the fit could be considered “sloppy”. Because the presented framework is independent of the pathway in consideration, it can be applied to other signaling pathways which also show considerable cell-to-cell variability in many other signaling networks. It will be interesting to see whether the structure distribution and existence of multiple distinct cell states are universal and occur in other signaling networks.

Chapter 2.5 Materials and Methods

Chapter 2.5.1 Cell plating

MCF10A cells were cultured in F12/DMEM media supplemented with horse bovine serum, insulin, cholera toxin, and hydrocortisol according to (Debnath *et al*, 2003). Before the ligand perturbation experiment, cells were cultured overnight in a 96-well plate in assay media as described in Debnath *et al* (2003). On the day of the experiment, cells were incubated with a solution of 4 μ M Fluo-4 (a calcium indicator) and 1 μ M Hoechst (a nucleic acid dye) for 20 min. After these 20 min, we exchanged the media with extracellular Hepes buffer supplemented with probenecid and glucose.

Chapter 2.5.2 Preparation of calibration buffer

To calibrate the concentration of cytosolic calcium from the raw intensity, we prepared calibration buffer solutions according to an established protocol (Bao *et al*, 2010). The minimum calibration buffer was comprised of 50 μ M ionomycin, 5 μ M thapsigargin, and 12.5 mM EGTA; the maximum calibration buffer was comprised of 50 μ M ionomycin, 5 μ M thapsigargin, and 36 mM CaCl_2 . Both buffers were in EC buffer supplemented with glucose and probenecid.

Chapter 2.5.3 Measurements of single-cell calcium response to an increase in extracellular ATP perturbation

Image acquisition of the 96-well plate was conducted on a Nikon Ti microscope using a 10 $\times$ objective (Selimkhanov *et al*, 2014). The microscope was automated using micro-manager (www.micro-manager.org) through its MATLAB scripting interface. After acquiring baseline calcium levels for 5 min, 10 μ M ATP was added and acquisition

continued for another 15 min. To allow single-cell calibration of the fluorescent response, the extracellular buffer was washed three times and replaced with minimum calibration buffer. Acquisition continued until equilibration of calcium levels followed by additional washes and the addition of maximum calcium buffer. The images were segmented to locate the Hoechst signals as the positions of the nuclei. The calcium signals were mapped to the Hoechst images taken at different time points to establish single-cell trajectories. The procedure is identical to that described in Selimkhanov *et al* (2014). Concentration of calcium was calibrated following Kao *et al* (1989). Briefly, for each single-cell, the $F_{\min}$ and $F_{\max}$ were estimated following treatment with the minimal and maximal calibration buffers, respectively. Calcium levels were calculated as: $[Ca^{2+}]_{\text{free}} = K_d (F - F_{\min}) / (F_{\max} - F)$ where F is the intensity measurement and K_d is the dissociation constant between Fluo-4 and calcium.

Chapter 2.5.4 Longitudinal dual-stimulation experiment

Dual-stimulation experiments were performed by first stimulated with ATP and then with intracellular IP3 by uncaging a previously loaded caged IP3 in MCF10A cells that stably express genetically encoded calcium activity reporter GCaMP5 (Akerboom *et al*, 2012) fused to mCherry (Su *et al*, 2013). The mCherry was used to normalize for expression level by dividing the GCaMP signal for each cell by the average mCherry signal over the same segmentation mask. Prior to experiments, cells were first incubated in 1:2,000 dilution of 10 mg/ml Hoechst and then incubated in 1.5 μM membrane permeable caged IP3 for 1 h (Enzo Life Sciences, Yan *et al*, 2015). During the experiment, the cells were first stimulated with 10 μM ATP and imaged for the subsequent 15 min. After the 15-min period, cells were allowed 35 min to recover from the ATP stimulated while still imaged to enable cell tracking. After the recovery period, cells were exposed to 365-nm LED light

(Thorlabs) for 5 s to uncage IP3. The subsequent calcium response from the cells was imaged for 5 min.

Chapter 2.5.5 Model of calcium signaling

Our calcium model (Fig 2.2) is based on a compilation of models from Lemon *et al* (2003) and Li and Rinzel (1994) with few simplifications to improve parameter estimation through the increase of model identifiability. We simplified the receptor dynamics to an ATP-dependent step function followed by first-order decay. This was implemented by multiplying the activation of PLC by receptor and ligand concentration L and an exponential term which accounts for the desensitization of the surface receptor. The equation that describes the rate of change of IP3 is a Hill function which describes the production rate of IP3 in terms of PLC. The degradation rate of IP3 is a linear rate term proportional to IP3 concentration. The description of calcium dynamics was modified from Lemon *et al* (2003) to simplify and have only one term for the calcium buffer. The equation for the fraction of activated IP3 channels was taken from the simplified model of Li and Rinzel (1994). The model equations are presented in Fig 2.2, and parameters are in Table EV1.

Chapter 2.5.6 Parameter estimation: pre-processing of calcium response

Cellular calcium responses contained stochastic elements, and the experimental noise was not captured by the mathematical model described above. To improve parameter estimation, we pre-processed the single-cell calcium response data to remove aspects of the responses that were not captured by our model. We used a low-pass filtering scheme. Because the initial calcium response to ATP was very sharp, we truncated the baseline measurement to include only frames following ATP addition and

utilized a “reflective boundary condition”, that is, we concatenated to the signal a mirror image of itself prior to filtering. The time points before the addition of ATP are not included in the model fitting and are used only to calculate the basal level point, which was designated as the first time point of each trajectory. As a result of this pre-processing, we removed all the high-frequency elements in calcium response that were not captured by our model without dampening the initial sharp response.

Chapter 2.5.7 Parameter estimation: calculating goodness of fit

The goodness of fit between the simulated data and real data was determined by performing a sum square of error on the comparison between a simulation and the smoothed experimental data. We then compared the first derivative of the simulation and the first derivative of the smoothed experimental data. We found that including information on the first derivative improved the fitting as it increased the emphasis on the shape of the curve compared with traditional sum square error. We used the difference between the raw (unprocessed) and the processed data as an acceptance threshold to be considered a “good fit”. In other words, if the simulated data agree with raw data with the same or better goodness of fit than simple smoothing, then the parameters corresponding to the simulated data will be accepted as a good fit.

Chapter 2.5.8 Parameter estimation: rejection sampling

We utilized a SMC-ABC parameter estimation following Liepe *et al* (2014). At the core of the SMC-ABC algorithm is a rejection sampling estimator that estimates posterior distribution based on a prior ($\pi(\Theta)$), a goodness-of-fit score ($\text{score}(D', D)$), and an acceptance threshold (ϵ). A parameter vector is defined by the collection of parameters that are needed to fully specify the model shown in Fig 2.2. To sample an N parameter

vectors from the posterior distribution, the algorithm randomly samples parameters from the prior and simulates them and scores their agreement with the data; it then accepts only the parameters that fit to the data within the threshold score. The rejection sampling pseudocode is shown in Scheme 1.

Chapter 2.5.9 Parameter estimation: approximate Bayesian computation

sequential Monte Carlo

The challenge with the rejection sampling described above is that if the prior is very different from the posterior, the efficiency of sampling can be too low to be practical. The ABC-SMC algorithm circumvents this problem by repeating the rejection sampling iteratively where in each iteration, the prior distribution and the acceptance threshold are updated. The algorithm performs the first iteration of rejection sampling with the prior distribution $\pi_0(\Theta)$ provided by the user. In the actual implementation, $\pi_0(\Theta)$ is a log10 uniform distribution in the range $[-1, 1]$ of the parameter value fold difference from the reference values of the parameters listed in Table EV1. The first iteration terminates when the sampling procedure collects a parameter vector population of size $N_{\text{iteration}}$ which meets the preliminary threshold requirement ϵ_{prelim} . The threshold score for each of the intermediate iterations of rejection sampling is determined relatively. For a parameter vector to be accepted, it has to produce simulated data that score better than the bottom 90% fraction of the population collected in the previous iteration. The entire algorithm terminates when it reaches the desired number of samples, all with goodness-of-fit scores lower than the threshold ϵ_{final} . The threshold ϵ_{final} is determined by scoring the raw data against the processed data. Therefore, the threshold ϵ_{final} is determined for each single cell based on the data quality for that cell.

Chapter 2.5.10 Parameter estimation: convergence test

A key challenge in parameter fitting is the possibility that the fit will identify a “local minimum” that does not represent the full distribution of possible parameters. To test that this is not the case, we utilized a strict test for convergence. For each cell for which we attempted to estimate its parameters, the entire ABC-SMC algorithm described above was performed twice in two completely independent runs (run #1 and run #2) to produce two samples of parameter vectors. To determine whether these two samples were statistically indistinguishable, we performed the following test: For each of the single-cell data, we combined the parameters from both runs and for each parameter identified the run identity (e.g., #1 or #2) of the nearest neighbor to calculate the probability that a nearest neighbor has the same run identity as the tested parameter. We then compared this measured distribution with a “null” model where there was equal (0.5) probability that a sample nearest neighbor was from the same run. If the chi-square statistic failed to reject the null hypothesis with significance level 0.05, we determine that the two populations are convergent.

Chapter 2.5.11 Clustering of the parameter ensemble

To cluster cells by calcium response, we first identified a suitable distance measure. Our estimation for each cell state was a sample from the posterior parameter distribution given the model and the data from that cell. Therefore, we have not only information on the degree of overall difference between two parameter sets, but also degree of variability of parameters within this set. We designed a distance measure that takes this parameter uncertainty into account. Specifically, for every two cells i and j , we first construct a distribution that quantify the probability that a specific parameter vector for cell i (j) has a nearest neighbor parameter vector that belong to the same cell i (j). We

then calculated the Kullback–Leibler (KL) divergence of the measured probability distribution from a null model where identities i and j for all parameter samples are randomly permuted. The permutation reflects a null model where there is no difference between cell i and j in their parameter distributions. The KL divergence from a random model was used as the distance measure for cluster analysis. The pairwise KL divergence was then calculated for all the ensembles (Fig 2.3). We then used hierarchical clustering with average linkage to cluster the ensembles based on the KL divergence measure. The hierarchical tree was subsequently thresholded to generate clusters based on Calinski–Harabasz optimal criterion that used medoids of individual parameter ensembles associated with single cells as the center of those ensembles. In the same way, the medoids of clusters of parameter ensembles were used to represent the cluster centers (Caliński & Harabasz, 1974).

Chapter 2.5.12 Data availability

The mathematical model from this publication has been submitted to BioModels Database and assigned the identifier MODEL1611150001. The same SBML file is also attached as Computer Code EV1 along with this paper.

Chapter 2.6 Author contributions

JY and RW designed the study and performed analysis. JY performed the experiments. AP made key reagent. JY and RW wrote the manuscript.

Chapter 2.7 Conflict of interest

The authors declare that they have no conflict of interest.

Chapter 2.8 Acknowledgements

We thank Noa Pinter-Wollman for critical reading of the manuscript. The work was supported by GM111404 and EY024960 from NIH.

Chapter 2, in full, is a reprint of material as it appears in “Distinct cellular states determine calcium signaling response,” *Molecular Systems Biology*, 2016. Jason Yao, Anna Pilko, Roy Wollman. The dissertation author was the primary investigator of this paper.

Chapter 2.9: References

- Akerboom J, Chen TW, Wardill TJ, Tian L, Marvin JS, Mutlu S, Calderón NC, Esposti F, Borghuis BG, Sun XR, Gordus A (2012) Optimization of a GCaMP calcium indicator for neural activity imaging. *J Neurosci* 32: 13819 – 13840
- Bao XR, Fraser IDC, Wall EA, Quake SR, Simon MI (2010) Variability in Gprotein- coupled signaling studied with microfluidic devices. *Biophys J* 99: 2414 – 2422
- Bendall SC, Davis KL, Amir E-AD, Tadmor MD, Simonds EF, Chen TJ, Shenfeld DK, Nolan GP, Pe'er D (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* 157: 714 – 725
- Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* 33: 155 – 160
- Cali_nski T, Harabasz J (1974) A dendrite method for cluster analysis. *Commun Stat Theory Methods* 3: 1 – 27
- Cohen-Saidon C, Cohen AA, Sigal A, Liron Y, Alon U (2009) Dynamics and variability of ERK2 response to EGF in individual living cells. *Mol Cell* 36: 885 – 893
- Debnath J, Muthuswamy SK, Brugge JS (2003) Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in threedimensional basement membrane cultures. *Methods* 30: 256 – 268
- Durruthy-Durruthy R, Heller S (2015) Applications for single cell trajectory analysis in inner ear development and regeneration. *Cell Tissue Res* 361: 49 – 57
- Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. *Science* 297: 1183 – 1186
- Eydgahi H, Chen WW, Muhlich JL, Vitkup D, Tsitsiklis JN, Sorger PK (2013) Properties of cell death models calibrated and compared using Bayesian approaches. *Mol Syst Biol* 9: 644
- Furusawa C, Kaneko K (2012) A dynamical-systems view of stem cell biology. *Science* 338: 215 – 217
- Geva-Zatorsky N, Rosenfeld N, Itzkovitz S, Milo R, Sigal A, Dekel E, Yarnitzky
- T, Liron Y, Polak P, Lahav G, Alon U (2006) Oscillations and variability in the p53 system. *Mol Syst Biol* 2: 2006.0033
- Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP (2007) Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol* 3: 1871 – 1878
- Handly LN, Pilko A, Wollman R (2015) Paracrine communication maximizes cellular response fidelity in wound signaling. *Elife* 4: e09652
- Kao JP, Harootunian AT, Tsien RY (1989) Photochemically generated cytosolic calcium pulses and their detection by fluo-3. *J Biol Chem* 264: 8179 – 8184
- Ladbury JE, Arold ST (2012) Noise in cellular signaling pathways: causes and effects. *Trends Biochem Sci* 37: 173 – 178

- Lee TK, Denny EM, Sanghvi JC, Gaston JE, Maynard ND, Hughey JJ, Covert MW (2009) A noisy paracrine signal determines the cellular NF-kappaB response to lipopolysaccharide. *Sci Signal* 2: ra65
- Lemon G, Gibson WG, Bennett MR (2003) Metabotropic receptor activation, desensitization and sequestration-I: modelling calcium and inositol 1,4,5- trisphosphate dynamics following receptor activation. *J Theor Biol* 223: 93 – 111
- Li Y-X, Rinzel J (1994) Equations for InsP 3 receptor-mediated $[Ca^{2+}]_i$ oscillations derived from a detailed kinetic model: a Hodgkin-Huxley like formalism. *J Theor Biol* 166: 461 – 473
- Liepe J, Kirk P, Filippi S, Toni T, Barnes CP, Stumpf MPH (2014) A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nat Protoc* 9: 439 – 456
- Lu Y, Xue Q, Eisele MR, Sulistijo ES, Brower K, Han L, Amir E-AD, Pe'er D, Miller-Jensen K, Fan R (2015) Highly multiplexed profiling of single-cell effector functions reveals deep functional heterogeneity in response to pathogenic ligands. *Proc Natl Acad Sci USA* 112: E607 – E615
- Marco E, Karp RL, Guo G, Robson P, Hart AH, Trippa L, Yuan G-C (2014) Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc Natl Acad Sci USA* 111: E5643 – E5650
- Meldolesi J, Pozzan T (1998) The heterogeneity of ER Ca^{2+} stores has a key role in nonmuscle cell signaling and function. *J Cell Biol* 142: 1395 – 1398
- Oates AC (2011) What's all the noise about developmental stochasticity? *Development* 138: 601 – 607
- Raue A, Kreutz C, Maiwald T, Klingmuller U, Timmer J (2011) Addressing parameter identifiability by model-based experimentation. *IET Syst Biol* 5: 120 – 130
- Rhee A, Cheong R, Levchenko A (2014) Noise decomposition of intracellular biochemical signaling networks using nonequivalent reporters. *Proc Natl Acad Sci USA* 111: 17330 – 17335
- 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
- Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathali P, Choi K, Bendall S, Friedman N, Pe'er D (2016) Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol* 34: 637 – 645
- Sheftel H, Shoval O, Mayo A, Alon U (2013) The geometry of the Pareto front in biological phenotype space. *Ecol Evol* 3: 1471 – 1483
- Shibata T, Fujimoto K (2005) Noisy signal amplification in ultrasensitive signal transduction. *Proc Natl Acad Sci USA* 102: 331 – 336

Snijder B, Sacher R, Rämö P, Damm E-M, Liberali P, Pelkmans L (2009) Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature* 461: 520 – 523

Sonnemann KJ, Bement WM (2011) Wound repair: toward understanding and integration of single-cell and multicellular wound responses. *Annu Rev Cell Dev Biol* 27: 237 – 263

Spitzer MH, Gherardini PF, Fragiadakis GK, Bhattacharya N, Yuan RT, Hotson AN, Finck R, Carmi Y, Zunder ER, Fantl WJ, Bendall SC, Engleman EG, Nolan GP (2015) IMMUNOLOGY. An interactive reference framework for modeling a dynamic immune system. *Science* 349: 1259425

Su S, Phua SC, DeRose R, Chiba S, Narita K, Kalugin PN, Katada T, Kontani K, Takeda S, Inoue T (2013) Genetically encoded calcium indicator illuminates calcium dynamics in primary cilia. *Nat Methods* 10: 1105 – 1107

Tay S, Hughey JJ, Lee TK, Lipniacki T, Quake SR, Covert MW (2010) Single-cell NF-kappaB dynamics reveal digital activation and analogue information processing. *Nature* 466: 267 – 271

Toettcher JE, Weiner OD, Lim WA (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

Wojcikiewicz RJ (1995) Type I, II, and III inositol 1,4,5-trisphosphate receptors are unequally susceptible to down-regulation and are expressed in markedly different proportions in different cell types. *J Biol Chem* 270: 11678 – 11683

Yan J, Schmid E, Hosseinzadeh Z, Honisch S, Shumilina E, Fuchs J, Lang F (2015) Impact of janus kinase 3 on cellular Ca²⁺ release, store operated Ca²⁺ entry and Na⁺/Ca²⁺ exchanger activity in dendritic cells. *Cell Physiol Biochem* 36: 2287 – 2298

Chapter 2.10 Figures

Figure 2.1 Calcium Figure 2.1. Calcium signal response heterogeneity

A. matrix of single-cell calcium response to ATP perturbation. Each row represents the calcium response of a single cell, and columns are time points. The heatmap figure contains data from a total of 1,904 cells.

B. The average calcium response of the population. The solid green line is the population average, and the light green shade is the standard deviation.

C. Examples of single-cell calcium data.

D. Example of single-cell calcium data with representative features of the basal value (F_o), time to reach half maximum (T_{50U}), time to reach maximum (T_m), maximum value (F_m), time to decay to half maximum (T_{50D}), and steady-state final value (F_t).

E. Boxplots of representative time-series features of calcium signals in the cell population, normalized to the average value of the particular feature. The red lines of the boxplots represent the median of the normalized values, and the ranges of the whiskers are defined to be 1.5 times the interquartile range (difference between 75th and 25th percentiles).

F., G. Spatial distributions within one representative well of the maximum signal (F_m) and the time to decay to half maximum (T_{50D}).

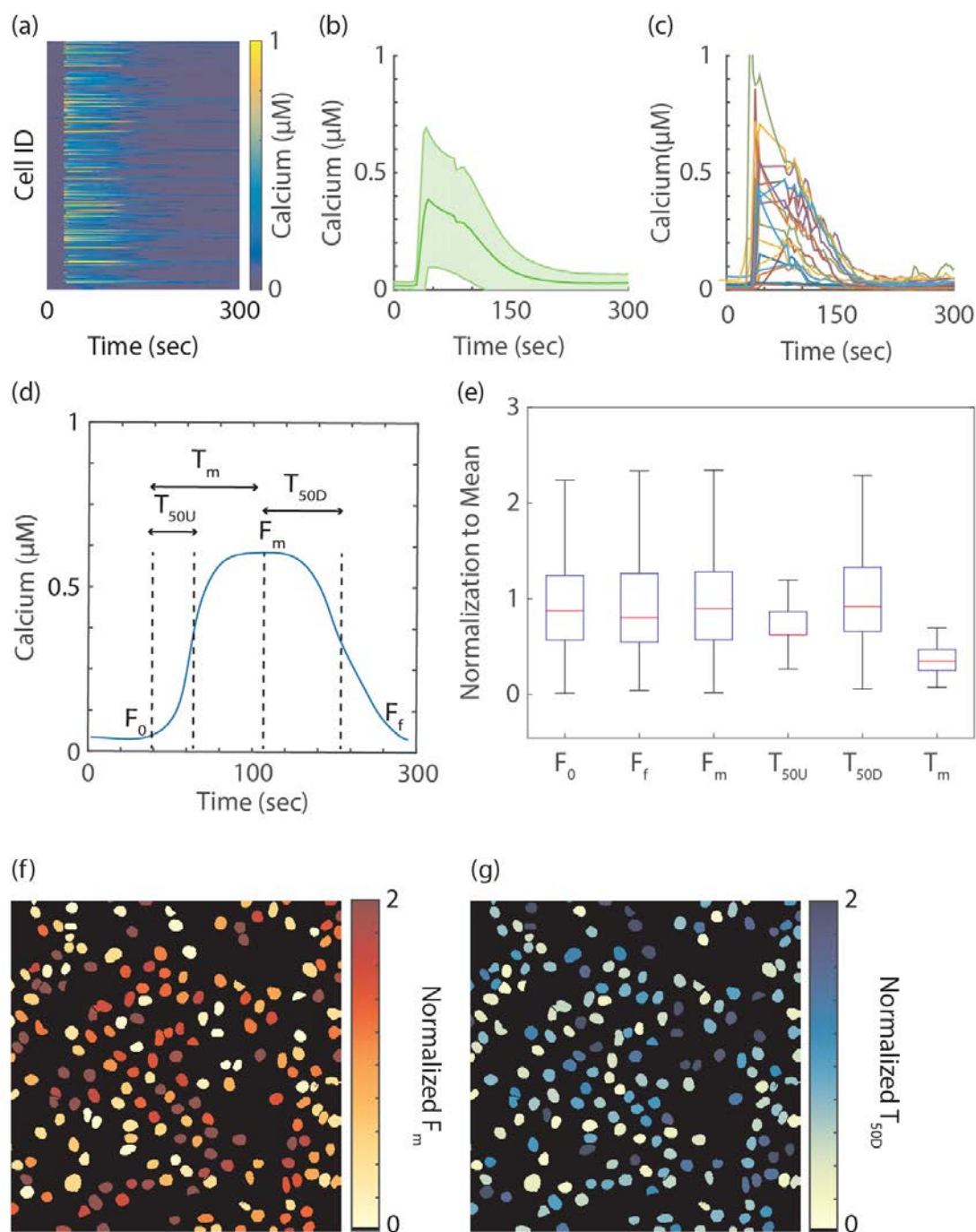


Figure EV2.1. Clustering results for cells based on only time-course data

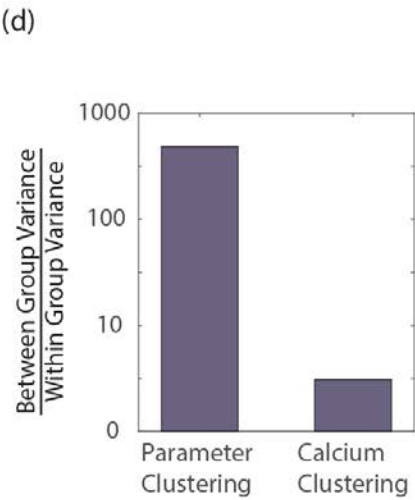
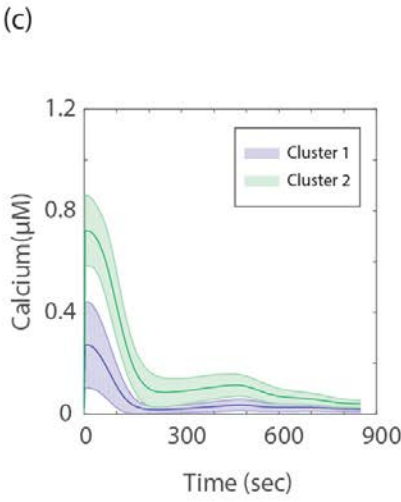
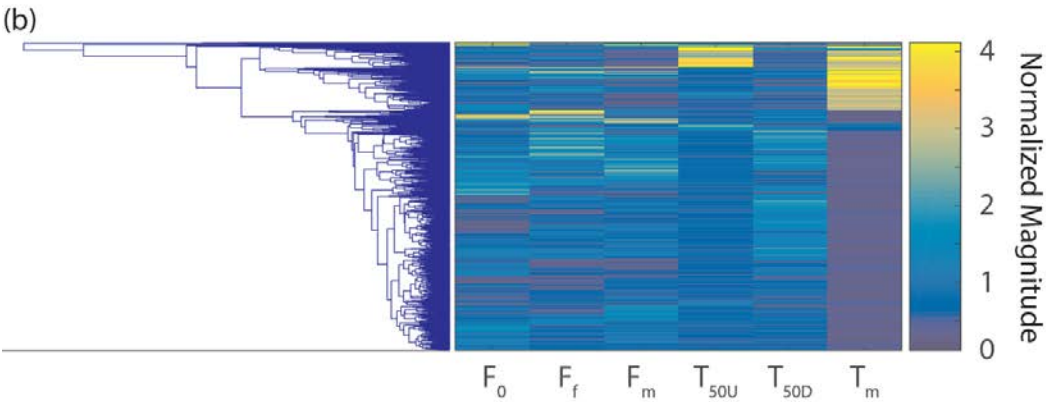
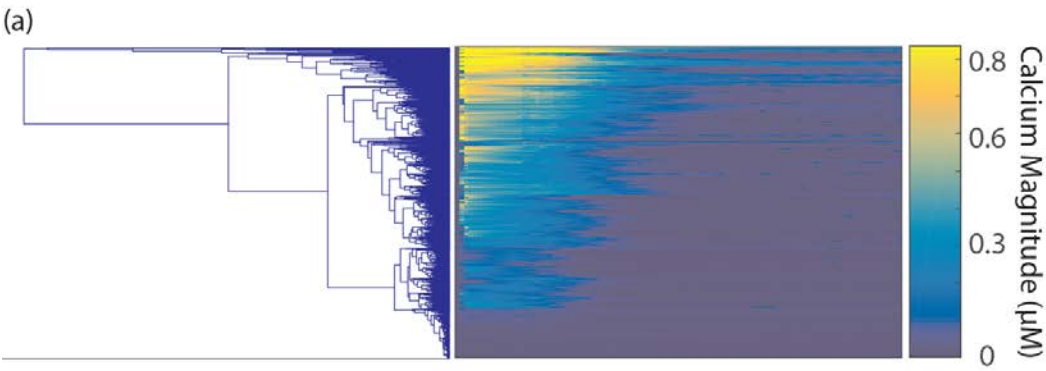
A. Clustering based on calcium data. Results of cluster analysis based on the Euclidean distances between the whole time series of calcium data. The left side of the panel shows the dendrogram, and the right side of the panel shows the corresponding matrix of all the time-course calcium data, with each row being from a single cell and the columns being the time points.

B. Clustering based on time-series features. The left side of the panel shows dendrogram of the clustering based on time-series features of the data, and the right side of the panel shows the corresponding matrix of the time-series features of the data. Each of the rows represents a single cell; the columns are elements of the time-series features for the cell. The features from left to right are basal level before stimulation, time taken to reach half maximal level, time taken to reach maximum level, maximum level, time from maximal level to half maximal, and steady state after stimulation. The individual features are normalized with respect to the average values.

C. The plot of the average and standard deviation of calcium data of the two clusters of the clustering that was based on calcium response data where the hierarchical tree (panel A) was thresholded based on the Calinski–Harabasz criterion.

D. The bar graph of the ratio of between group variance to within group variance from both the parameter clustering and calcium data clustering. The results show that the kinetic parameter clustering scheme has higher separation between the individual clusters compared to that of calcium data.

E. The confusion table of cells that belong in the three most significant identified clusters in parameters and the two clusters with respect to calcium data clusters.



(e)

Parameter Cluster	Calcium Cluster	
	1	2
1	391	73
2	125	0
3	68	0

$$\frac{d[PLC]}{dt} = K_{on,ATP} * [ATP] * e^{-K_{ATP}t} - K_{off,PLC}[PLC]$$

$$\frac{d[IP3]}{dt} = V_{PLC} \frac{[PLC]^2}{K_{IP3}^2 + [PLC]^2} - K_{off,IP3}[IP3]$$

$$\frac{dh}{dt} = a ([Ca^{2+}] + d_{inh}) \left(\frac{d_{inh}}{[Ca^{2+}] + d_{inh}} - h \right)$$

$$\frac{d[Ca^{2+}]}{dt} = \beta \left(\varepsilon(\eta_1 \cdot m_{\infty}^3 \cdot h^3 + \eta_2)(c_0 - (1 + \varepsilon)[Ca^{2+}]) - \eta_3 \frac{[Ca^{2+}]^2}{k_3^2 + [Ca^{2+}]^2} \right)$$

$$\beta = \left(1 + \frac{K_e[B_e]}{(K_e + [Ca^{2+}])^2} \right)^{-1} \quad m_{\infty} = \left(\frac{[IP3]}{d_1 + [IP3]} \right) \left(\frac{[Ca^{2+}]}{d_5 + [Ca^{2+}]} \right)$$

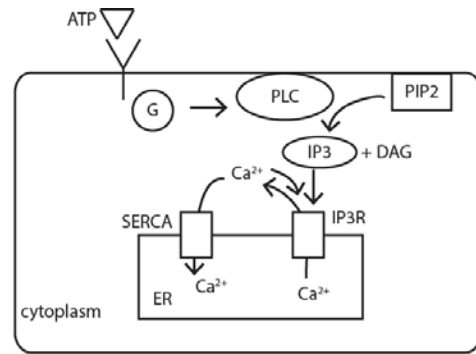


Figure 2.2 Calcium model

A. A system of ordinary differential equations of calcium signaling. The system contains four state variables of IP3, PLC, calcium, and IP3-activated receptor fraction.

B. Schematic of calcium signaling pathway. Through ATP stimulation, GPCR is recruited to activate enzyme PLC, which in turn cleaves PIP2 to form IP3 and DAG; IP3 then activates the IP3R channel to release calcium into the cytoplasm. The calcium is recycled back into the ER membrane through the SERCA pump.

Figure 2.3 Process of single-cell parameter fitting

A. Workflow of the single-cell data fitting process. The raw data are pre-processed to remove the technical noise and high-frequency biological responses. The model is fitted to the processed data in two independent runs of the algorithm to generate two posterior parameter samples. A statistical test based on similarities of the two independent runs was used to determine if the algorithm successfully converges to same posterior distribution. The result of the entire workflow is a population of parameter vectors that represent the single-cell fitting.

B. The progression of parameter fitting for a single trajectory. The parameter vectors through the iterations were projected using multidimensional scaling onto a 2D plane to illustrate the convergence of the parameters. The simulated data for each of the selected iterations are plotted below.

C. Comparison between a convergent runs and non-convergent runs. The crosses and the filled circles represent independent runs of the same single-cell data. The parameter posterior is projected onto the 2D plane through multidimensional scaling. There is a wider separation of parameter groups in the non-convergent fitting as opposed to the convergent run.

D. The violin plot of all kinetic parameters from the fit results to a single-cell calcium trajectory. The plot values are in terms of the log of 10 of the parameter values normalized according to the reference values determined either through literature or values within reasonable biological ranges. The meanings of the kinetic parameter symbols and their reference values are listed in Table EV1.

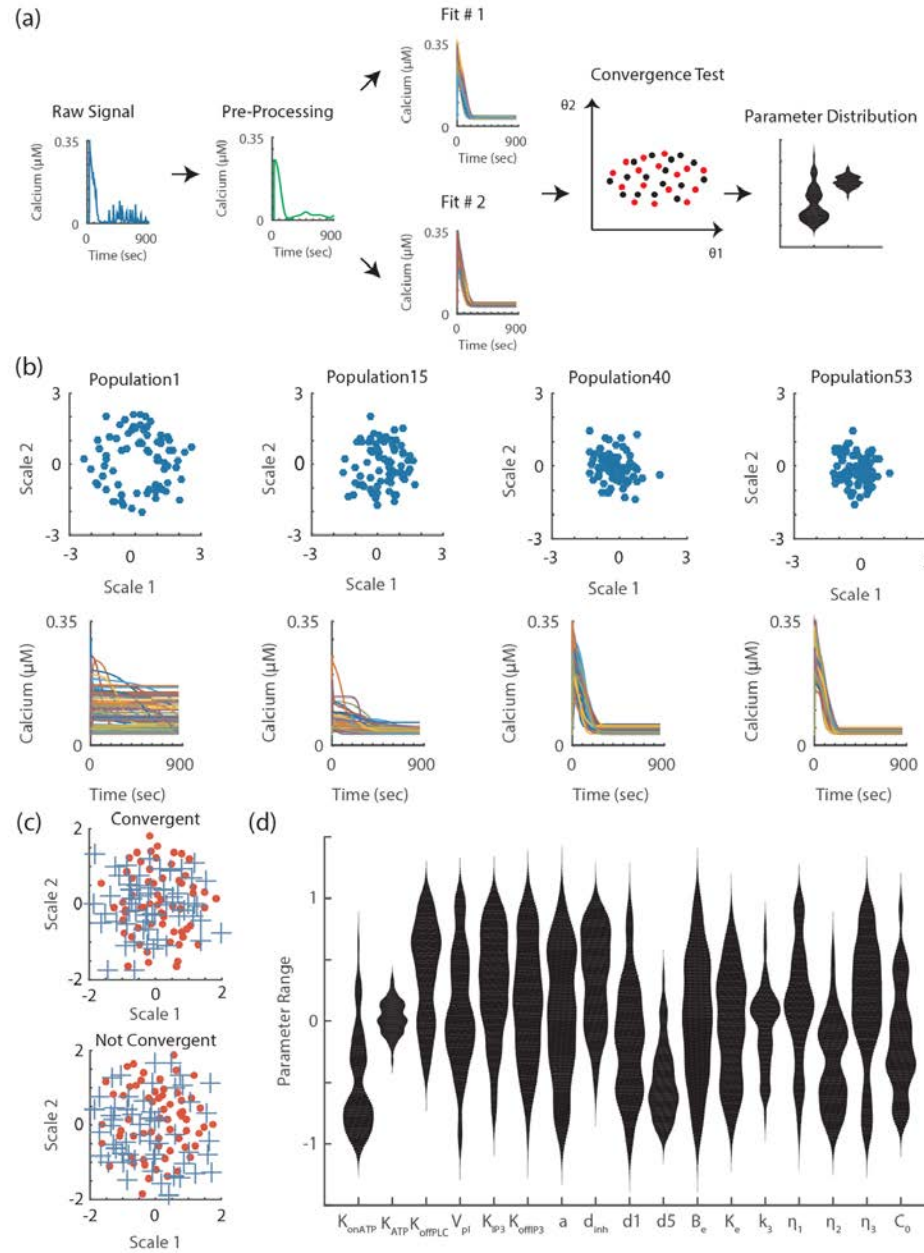


Figure EV 2.2. Quality of convergent single-cell model fitting

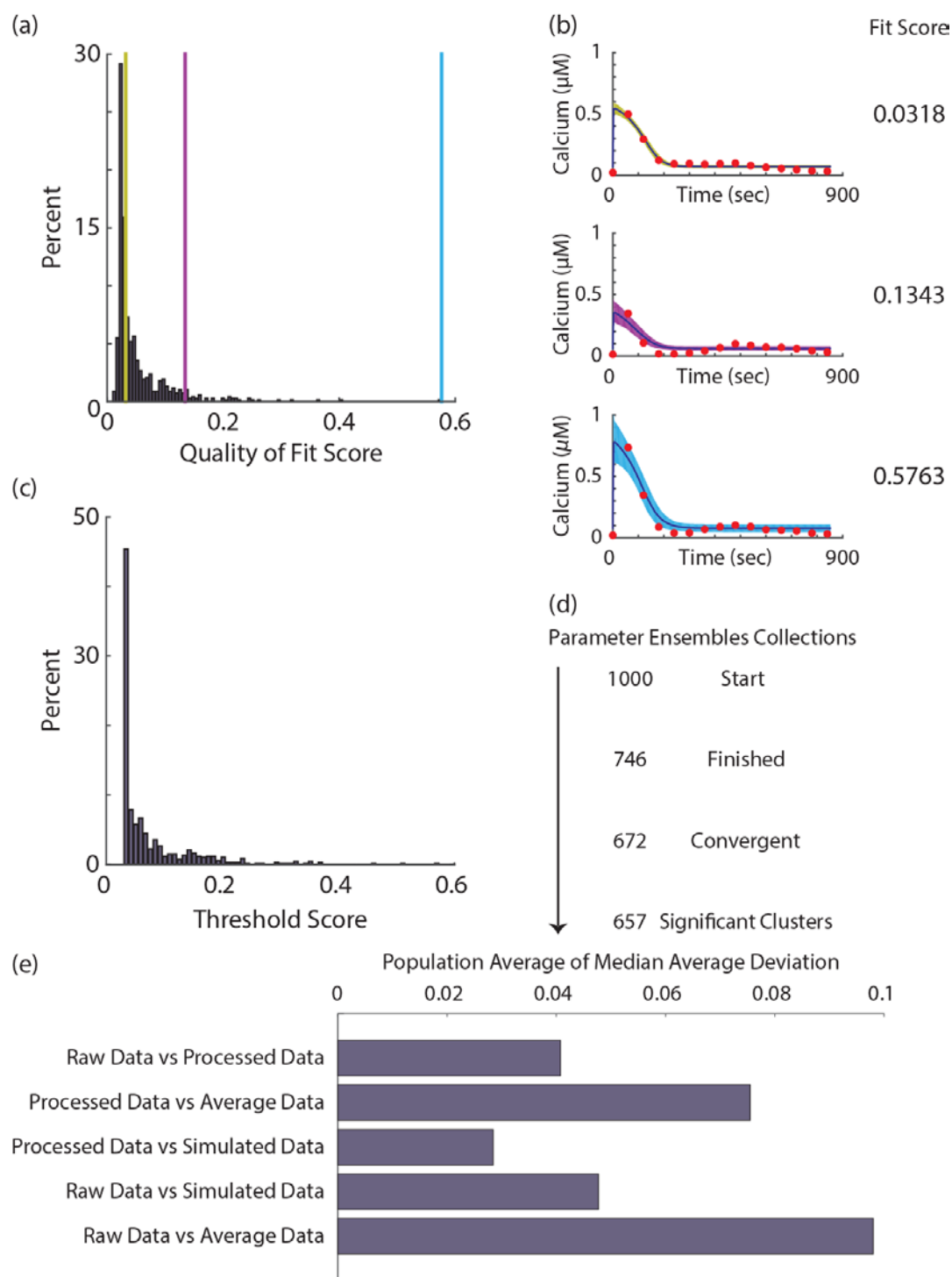
A. Histogram of average fit scores of convergent parameters. Each of the single-cell data is fitted to the differential equations model to produce a posterior parameter vector population and the average of the scores of the simulated data is calculated. The histogram shows the distribution of the average goodness-of-fit scores for all cells. The three vertical color bars show examples of three fits shown in panel (B).

B. Example Bayesian fit. The three instances of Bayesian fits show the fitted data as red dots, the average of the simulated data in dark blue trajectories, and confidence intervals whose colors correspond to their respective location in the histogram in panel. The fit ensembles average score is shown on the right of each plot. The three examples demonstrate visually the spectrum of quality of fits from the algorithm.

C. In our fit, each cell has its own final acceptance threshold based on experimental noise for that cell. The histogram shows the distribution of these acceptance threshold scores. All scores in panel (A) are below the corresponding individualized acceptance value shown in (C).

D. The progression number of cells in the fitting process. The entire workflow started out with 1,000 fits. Due to technical issues, 746 instances of fitting successfully finished. After the test of convergence, 672 fits were found to be convergent; out of the convergent instances of fits, 657 were found to belong to one of the three major clusters of parameter ensembles.

E. The bar plot of the median absolute deviation comparing raw data, processed data, and simulation for each single cell.



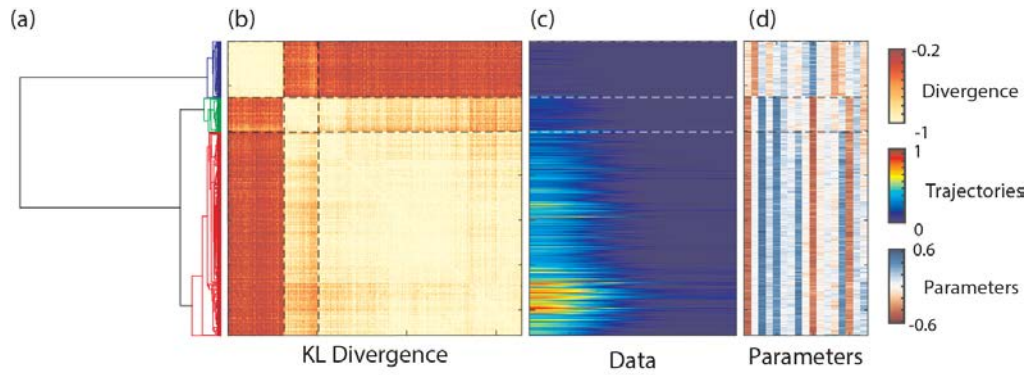


Figure 2.4 Clustering parameter distributions

A–D The clustering results of parameter distributions based on the calculated Kullback–Leibler divergence measure. From the left to the right are the (A) dendrogram of the clustering results, (B) distance matrix between parameter groups of single-cell data, (C) matrix of corresponding single-cell data, and (D) matrix of corresponding average parameter vectors for the single-cell data. Throughout the entire figure, each row represents the same cell.

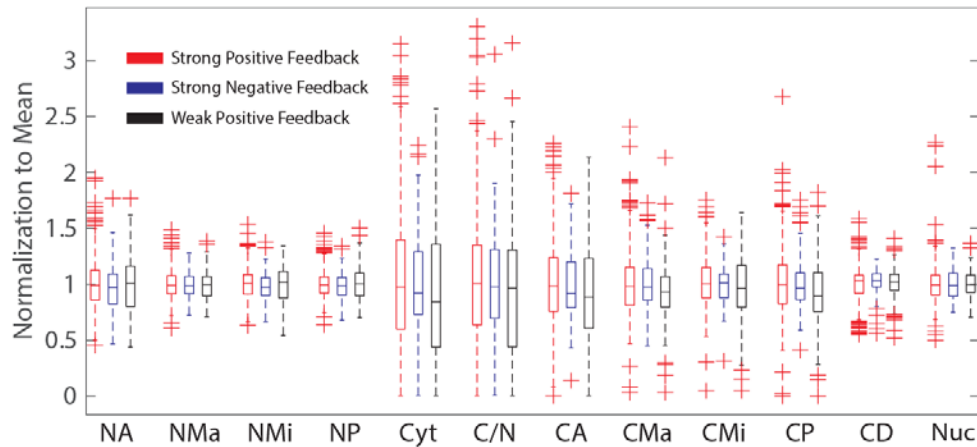


Figure EV 2.3. Distribution of cell physical features among clusters

The boxplots of the different cell physical features according to the three clusters, normalized with respect to the mean values. The acronyms of the cell physical features are the nuclear area (NA), major axis of nuclei (NMa), minor axis of nuclei (NMi), perimeter of the nuclei (NP), cytoplasm area (Cyt), ratio of cytoplasm to nuclei area (C/N), cytoplasm area (CA), cytoplasm area major axis (CMa), cytoplasm area minor axis (CMi), cell perimeter (CP), effective local cell density (CD), Hoechst intensities (Nuc). The red lines of the boxplots represent the median of the distributions, and the ranges of the whiskers are defined to be 1.5 times the interquartile range (difference between 75th and 25th percentiles). Crosses are data outside of the interquartile range.

Figure 2.5 Mechanistic difference between clustered data

A. Calcium signaling schematic diagram showing the difference between parameters of the clusters as reflected in different components of the signaling pathway. The intensity of the arrows in the diagram corresponds to the degree of difference in the associated parameters between the three identified clusters.

B. A boxplot of the parameters of IP3R channels of the three major clusters corresponding to the equilibrium constant of the inactivating subunit of calcium channel (dinh) and equilibrium constant of activating subunit of IP3R (d1 and d5). SP, strong positive; SN, strong negative; WP, weak positive. The values of the parameter range values are the log of 10 with respect to the reference parameter values chosen as listed in Table EV1. The red lines of the boxplots represent the median of the distributions, and the ranges of the whiskers are defined to be 1.5 times the interquartile range (difference between 75th and 25th percentiles). Crosses are data outside of the interquartile range.

C. Illustrations of the differences in parameters as manifested in calcium signaling mechanics. The red color indicates positive feedback interaction, while the blue color indicates negative feedback interaction. The thickness of the arrows corresponds to the strength of those interactions.

D. The scatter plot between the maximum calcium responses to both the ATP stimulation and IP3 uncaging ($R^2 = 0.68$).

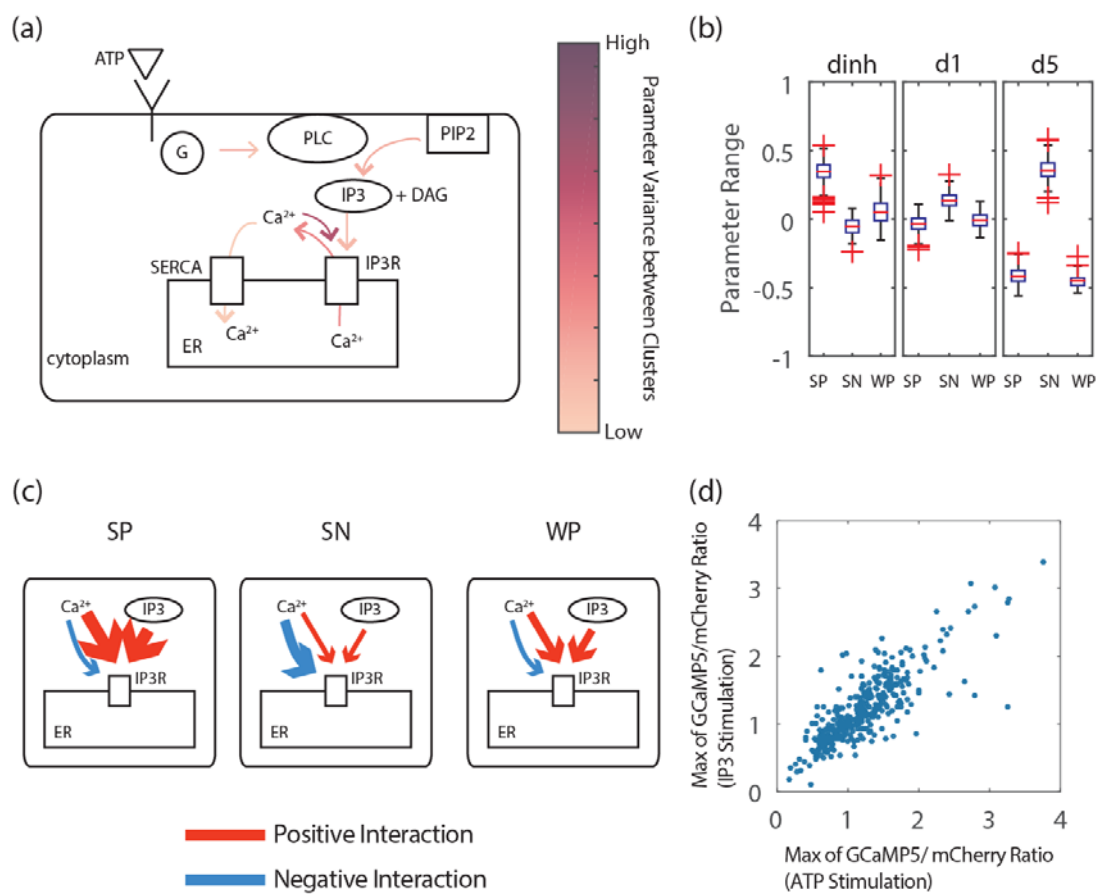
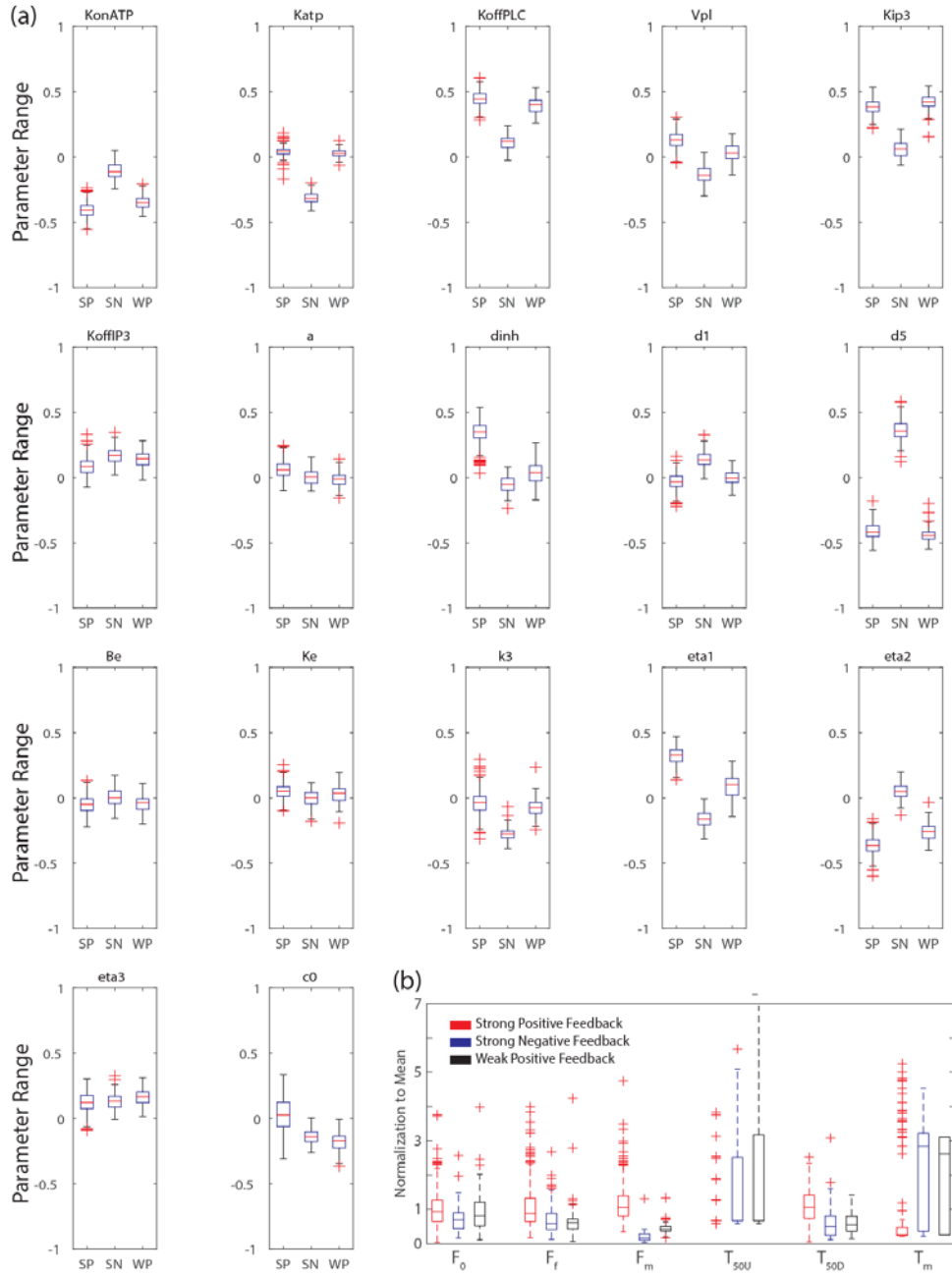


Figure EV 2.4. Boxplot of kinetic parameter distributions in the major clusters

A. The 17 boxplots show the variance of the 17 parameters in the three major clusters: strong positive feedback (SP), strong negative feedback (SN), and weak positive feedback (WP).

B. The boxplots of the cell calcium data temporal features belonging to the three major clusters. The values are normalized according to the mean. The red lines of the boxplots represent the median of the distributions, and the ranges of the whiskers are defined to be 1.5 times the interquartile range (difference between 75th and 25th percentiles).



Chapter 2.11 Schemes

Input:

$\pi(\Theta)$ = The prior probability distribution from which the parameter vector Θ is drawn

N = The size of the parameter vectors needed for the posterior in each iteration of rejection sampling

ϵ = The goodness of fit score needed for the parameter vector to be accepted into the posterior

D = The data to be fitted with the model

$\text{score}(D', D)$ = function to score the goodness of fit of D' against D

Output:

Q = the population of accepted parameter vectors as an approximation of $\pi(\Theta | \text{score}(D_{\text{sim}}, D) < \epsilon)$, the posterior distribution of the parameters

Begin

Initialize Q to be the empty set

While $|Q| < N$

Sample parameter vector Θ' from $\pi(\Theta)$

Carry out simulation with Θ' to produce simulated data D_{sim}

Score D_{sim} against D

If $\text{score}(D_{\text{sim}}, D) < \epsilon$

Add Θ' to Q

EndIf

EndWhile

End

Scheme 2.1. Rejection Sampling

Chapter 2.12 Tables

Table EV 2.1. Model Parameters

Parameter	Meaning	Value for Center of Prior	Reference
$K_{on,ATP} (S^{-1})$	Rate constant for PLC generation through ligand binding	1	N/A
$K_{ATP} (S^{-1})$	Exponential decay constant	8.3e-3	N/A
$K_{off,PLC} (S^{-1})$	Degradation constant for PLC	1.25	N/A
$V_{PLC} (\mu M/s^{-1})$	Maximum velocity for IP3 generation	1	N/A
$K_{IP3} (\mu M)$	Equilibrium constant for IP3 generation through PLC	0.5	N/A
$K_{off,IP3} (S^{-1})$	Degradation rate constant of IP3	1.25	N/A
$a (S^{-1})$	Time constant of IP3 channel	1	Li
$d_{inh} (\mu M)$	Dissociation constant for IP3 channel calcium inhibiting subunit	0.4	Li
$d1 (\mu M)$	Dissociation constant for IP3 channel IP3 activating subunit	0.13	Lemon
$d5(\mu M)$	Dissociation constant For IP3 channel calcium activating subunit	0.0823	Lemon
$B_e (\mu M)$	Concentration of calcium buffer	150	Lemon
$K_e (\mu M)$	Dissociation constant for calcium buffer	10	Lemon
$k3 (\mu M)$	SERCA pump dissociation constant	0.4	Lemon
$\eta_1 (S^{-1})$	IP3 channel permeability constant	575	Lemon
$\eta_2 (S^{-1})$	ER leak permeability constant	5.2	Lemon
$\eta_3 (S^{-1})$	Effective calcium pump permeability	45	Lemon
$c_0 (\mu M)$	Total calcium concentration	2	Lemon