# Phenotypic Plasticity of Cells in Epithelial-Mesenchymal Transition: An Experimental and Mathematical Study

*A thesis submitted for the degree of*

DOCTOR OF PHILOSOPHY

BY

VIMALATHITHAN D

Department of Biosciences and Bioengineering,
Indian Institute of Technology Guwahati, India

July 2020

Indian Institute of Technology
Guwahati, India

# DECLARATION

The thesis titled **'Phenotypic Plasticity of Cells in Epithelial-Mesenchymal Transition: An Experimental and Mathematical Study'** is a presentation of my original research work carried out in the Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, India, under the supervision of Dr. Biplab Bose. This work has not been submitted elsewhere for the award of any degree, diploma or equivalent.

Significant works from other researchers and other resources used in this thesis have been cited in the reference section. Those who have provided suggestions and technical help have been duly acknowledged.

July 2020

Vimalathithan D,
Roll no. 146106011,
Department of Biosciences and Bioengineering,
Indian Institute of Technology Guwahati, India.

# Indian Institute of Technology Guwahati, India

## CERTIFICATE

This is to certify that the work presented in the thesis titled, **'Phenotypic Plasticity of Cells in Epithelial-Mesenchymal Transition: An Experimental and Mathematical Study,'** by Vimalathithan D (Roll no. 146106011) for the award of the degree of Doctor of Philosophy is an authentic record of the research work carried out under my supervision in the Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, India.

This thesis or any part thereof has not been submitted elsewhere for the award of any other degree or diploma.

July 2020

Dr. Biplab Bose,
Associate Professor,
Department of Biosciences and Bioengineering,
Indian Institute of Technology Guwahati, India.

# Acknowledgments

I would like to express my sincere gratitude to all who have helped me in completing this dissertation.

I'm deeply indebted to my thesis advisor, Dr. Biplab Bose. He introduced me to the field of Systems Biology, and with his support now, I could submit a dissertation on mathematical modeling of biological systems. His immense knowledge, strong determination, and willingness to learn new things always motivated me. Most importantly, he has never fixed any working hours to the lab but only deadlines. He has given full freedom to work whenever I wish. I hope I have made use of it.

His most uttered statement, " So what?" during regular discussions, was instrumental in taking up challenging experiments. He was with me during failures and motivated me to troubleshoot and overcome those failures. He arranged funding for my research work, travel grants to attend several conferences, my monthly fellowship, and always provided a pleasant ambiance to carry out research. Apart from research, his interests in the bird diversity of the campus motivated me to become a birder. Thanks a lot, sir!

I would like to thank the funding agencies, Department of Biotechnology (DBT), Indian Council of Medical Research (ICMR), and Indian Institute of Technology Guwahati (IITG) for their support towards the smooth functioning of my research work. I wish to thank the Ministry of Human Resource Development (MHRD) and DBT for my fellowship.

I would like to extend my sincere thanks to the Doctoral Committee, Prof. Siddhartha S. Ghosh, Prof. Bhubaneswar Mandal, Prof. Ranjan Tamuli for their critical

suggestions in my every annual progress seminar and their consistent support in completing this dissertation.

I am incredibly grateful to Prof. Senthilkumar S. for his constant encouragement and moral support during my stay at IITG.

I cannot begin to express my thanks to the serene, beautiful green campus of IITG that provided the right ambiance to complete my research work. The dense hills, lakes, and wetlands within the campus and the mighty river bed near the KV gate are my stress busters.

Many thanks to Yashavanth P. R., Kamleswar C., Dr. Sharmila N., Kaushalesh Gupta, Ganesh N., Dixcy Jaba Sheeba J. M., Dr. Sajitha S., and Dr. Namami Goswami for their enormous assistance in my research work. Short discussions with them during coffee breaks were pivotal. I appreciate the timely help from Dr. Anil P. Bidkar for assisting me in operating the Confocal microscope.

I would like to acknowledge the help of my present and past lab members and all the DBT support facility members in maintaining the lab, inventory, and bill settlements. Special thanks to Dr. Mohitosh Dey, the scientific officer of the DBT support facility.

I would also like to extend my deepest gratitude to my family for their love and moral support.

I am supremely grateful to my friends at IITG, Ganesh N., Dixcy Jaba Sheeba J. M., Dr. Sajitha S., Rajeev Anupoju, A. Vamsi Krishna Reddy, Yashavanth P. R., Kamleswar C., Manojkumar M., Srirupa Bhattacharyya. This thesis would not have been possible without the support of them. Thank you all!

# Abstract

Phenotypic plasticity is the potential of cells to switch between different phenotypes of the cell. The transition from one phenotype to the other is often referred to as phenotypic state transition of cells and is observed during early embryogenesis, wound healing, and cancer metastasis. Epithelial-Mesenchymal Transition (EMT), one of the possible mechanisms of cancer metastasis involves phenotypic state transition. In the current work, we investigated the phenotypic plasticity in EMT at the phenomenon-level using EGF-induced EMT of MDA-MB-468 cells as an experimental system. Though most of the studies on EMT define phenotypes in terms of molecular markers, we defined phenotypic states in terms of cell morphology. Based on morphology, we defined three phenotypic states for these cells- Cobble, Spindle, and Circular. We used a quantitative imaging analysis to capture the dynamics of state transition and developed a mathematical method to estimate state transition trajectories from that data. Our analysis showed that in this experimental system, the dominant, reversible state transition path is Cobble → Circular → Spindle → Cobble. We also showed that an ultrasensitive on/off switch involving phospho-EGFR controls the state transition dynamics in these cells.

Further, we used the same experimental system to investigate how background chatter or interference by another signaling molecule can affect the cellular state transition. We introduced background noise by using suboptimal doses of TGF-β1. TGF-β1 alone did not induce any state transition in these cells. We used statistical analysis and information theory-based methods to understand the effects of such low doses of TGF-β1 on EGF-induced state transition. Our analysis showed that TGF-β1

exerted a positive synergistic influence to push cells towards Spindle and Cobble states but, at the same time, increased the noise in the process.

In this work, we classified cells based on morphology, and we did not have molecular information of cells in different phenotypic states. Quantitative PCR (qPCR) is a convenient method to investigate changes in gene expression. However, qPCR generates ensemble-averaged data. Moreover, we cannot physically segregate the subpopulations based on shape. In our current work, we developed a mathematical tool to estimate cell-type-specific gene expression by deconvoluting the qPCR data of a mixed population of cells, using the population distribution data obtained from quantitative image analysis. Our algorithm is generic and can be used to deconvolute any population-level data (like qPCR, Western Blot) when the distribution of subpopulations is known.

In this thesis, we used the EGF-induced EMT of MDA-MB-468 cells as an experimental system. However, the physical concepts and mathematical approaches developed in this work are generic and can be used for any other cellular phenomenon involving phenotypic state transition.

# Table of Contents

*List of Tables* *xv*

*List of Figures* *xvii*

*Table of Abbreviations and Acronyms* *xxi*

*Flow of the Thesis* *xxv*

**CHAPTER 1**

***Introduction*** ***1***

**CHAPTER 2**

***Review of Literature*** ***5***

2.1. Phenotypic state transition 5
2.2. Epithelial-Mesenchymal Transition 6
2.3. How are the cell states defined? 8
2.4. The potential landscape of cell state transition 9
2.5. Discrete cell state transition 15
2.6. Signal transduction in cells 20
2.7. Background noise in signal transduction 22
2.8. An information-theoretic approach to study signal transduction 24
2.9. Deconvolution of cell-type-specific gene expression from ensemble data 32
2.10. Existing deconvolution methods 33
2.11. Limitations of the existing deconvolution methods 35
2.12. Bayesian method of parameter estimation 36
2.13. Objectives 38

**CHAPTER 3**

***Materials and Methods*** ***39***

3.1. Cell culture 39
3.2. Phalloidin-FITC staining 41
3.3. Immunofluorescence 41

3.4. Isolation and quantification of RNA 42
3.5. Synthesis and quality check of cDNA 44
3.6. Quantitative PCR 46
3.7. Quantitative image analysis 47
3.8. Migration assay 49
3.9. Western blotting 50
3.10. Sandwich-ELISA to measure EGF in the media 53
3.11. Cell death estimation by flow cytometry 53
3.12. phospho-EGFR measurement through flow cytometry 54
3.13. Microplate assay to estimate live and dead cells 55
3.14. Cell viability assay 56
3.15. Data analysis 57

**CHAPTER 4**

***The Signaling and Dynamics of EGF-induced Epithelial-Mesenchymal Transition 59***

4.1. Introduction 59
4.2. EGF-induced EMT 60
4.3. Phenotypic states of MDA-MB-468 cells 61
4.4. Functional characterization of cell states 64
4.5. Population distribution of MDA-MB-468 cells 66
4.6. Dose-dependent temporal dynamics of the phenotypic states of MDA-MB-468 cells 67
4.7. Effect of EGF on cell proliferation and cell death 69
4.8. The cell state transition model 71
4.9. Trajectories of cell state transition 81
4.10. The null model 85
4.11. The Dynamics of phospho-EGFR drives the state transition 88
4.12. Adhesion signaling in cell state transition 93
4.13. Ultrasensitive switch-like response in cell state transition 94
4.14. Discussion 97

**CHAPTER 5**

***The Effect of Background Noise on EGF-induced Epithelial-Mesenchymal Transition 105***

5.1. Introduction 105

5.2. TGF-β1 modulated the EGF-induced state transition 106
5.3. TGF-β1 exerts a positive synergistic effect on Spindle and Cobble cells 108
5.4. Information-theoretic analysis 110
5.5. EGF-induced cell state transition is noisy 112
5.6. TGF-β1 amplifies the noise in EGF-induced state transition 118
5.7. Discussion 119

**CHAPTER 6**

***DEBay: A tool for estimation of cell-type-specific gene expression from quantitative PCR of ensemble of cells*** **123**

6.1. Introduction 123
6.2. The deconvolution algorithm 124
6.3. The graphical user interface of DEBay 132
6.4. Evaluation of DEBay with synthetic data 134
6.5. Evaluating DEBay with real biological data 144
6.6. Discussion 148

**CHAPTER 7**

***Conclusion*** **151**

***Bibliography*** **155**

***Publications*** **171**

***Presentations*** **171**

***Appendix A*** **173**

Section A-1: Estimation of percentage dead cell population 173

***Appendix B*** **185**

Section B-1: Reagents used in cell culture 185
Section B-2: Reagents used in imaging 186
Section B-3: Reagents used in agarose gel electrophoresis 186
Section B-4: Reagents used in western blotting 187
Section B-5: Reagents used in cell cycle analysis 188

# List of Tables


Table 3.1. Treatment conditions followed in experiments. 40

Table 3.2. DNase digestion reaction setup. 43

Table 3.3. Components of cDNA master mix. 45

Table 3.4. PCR reaction preparation. 45

Table 3.5. PCR reaction conditions. 46

Table 3.6. Components of the quantitative PCR reaction. 47

Table 3.7. Quantitative PCR conditions. 47

Table 5.1. Contingency table to calculate mutual information. 112

Table 5.2. Contingency table for 24 h of EGF treated cells. 113

Table 5.3. The joint probability of the input signal and the response for 24 h of EGF treated cells. 114

Table 5.4. Marginal probability of the input signal for 24 h of EGF treated cells. 114

Table 5.5. Marginal probability of the response for 24 h of EGF treated cells. 114

Table 6.1. Expression of target mRNAs in each cell type as a function of time – SET 1. 139

Table 6.2. Expression of target mRNAs in each cell type as a function of time – SET 2. 140

Table 6.3. Expression of target mRNAs in each cell type as a function of time – SET 3. 140

Table A-1: The fractional cell division values estimated from the state transition model. 175

Table A-2: The fractional state transition values of 5 ng/mL EGF treated samples estimated from the state transition model. 176

Table A-3: The fractional state transition values of 10 ng/mL EGF treated samples estimated from the state transition model. 177

Table A-4: The fractional state transition values of 25 ng/mL EGF treated samples estimated from the state transition model. 178

Table B-1: List of antibodies used in western blotting. 189

Table B-2: List of antibodies used in Immunofluorescence. 189

Table B-3: List of antibodies used in flow cytometry experiments. 190

Table B-4: List of primers used in PCR. 190

Table B-5: The composition of different percentages of resolving gel. 191

Table B-6: The composition of stacking gel. 193

# List of Figures

Figure 2.1: Cell state transition during EMT and MET. 6

Figure 2.2: Bifurcation plot. 11

Figure 2.3: The potential landscape of the cell. 13

Figure 2.4: Discrete Markov model for cell state transition. 16

Figure 2.5: Different ways of interconnections in signaling pathways of the cell. 23

Figure 2.6: Signal discrimination in cells. 25

Figure 2.7: Entropy plot of a binary discrete random variable. 26

Figure 2.8: Contingency table to elucidate mutual information 29

Figure 2.9: Cell-type-specific gene expression is obscured in ensemble measurement. 33

Figure 4.1: EGF-induced cytoskeletal changes. 61

Figure 4.2: EGF-induced change in the gene expression of EMT markers. 62

Figure 4.3: Distinct morphologies of MDA-MB-468 cells. 63

Figure 4.4: EGF-induced change in the population distribution of MDA-MB-468 cells. 63

Figure 4.5: Migratory potential of MDA-MB-468 cells. 65

Figure 4.6: Scattering potential of MDA-MB-468 cells. 66

Figure 4.7: Steady-state population distribution of MDA-MB-468 cells. 67

Figure 4.8: EGF-induced temporal dynamics of the population distribution of MDA-MB-468 cells. 68

Figure 4.9: EGF-induced reversible change in the population distribution of MDA-MB-468 cells. 69

Figure 4.10: Dose-dependent effect of EGF on total cell number. 70

Figure 4.11: Dose-dependent effect of EGF on cell death. 71

Figure 4.12: State transition model. 73

Figure 4.13: Pareto front analysis of 10 ng/mL EGF treated experiment. 79

Figure 4.14: Fraction of each cell state estimated from the state transition model. 80

Figure 4.15: State transition trajectories of cells treated with 5 ng/mL EGF. 81

Figure 4.16: State transition trajectories of cells treated with 10 ng/mL EGF. 82

Figure 4.17: EGF-induced change in the population distribution of cells observed at a shorter time interval. 83

Figure 4.18: State transition trajectories of cells treated with 10 ng/mL EGF for a period of 0-12 h. 83

Figure 4.19: State transition trajectories of cells treated with 10 ng/mL EGF for a period of 24-36 h. 84

Figure 4.20: State transition trajectories of cells treated with 25 ng/mL EGF. 84

Figure 4.21: Null model. 86

Figure 4.22: Null model validation. 88

Figure 4.23: Time-dependent EGF availability to the cells. 89

Figure 4.24: The dose-dependent temporal dynamics of phospho-EGFR. 90

Figure 4.25: Sustained EGF signaling favors the Circular cell state. 91

Figure 4.26: Single-cell level measurements of phospho-EGFR and total-EGFR of 10 ng/mL EGF treated cells. 92

Figure 4.27: Single-cell level measurements of phospho-EGFR and total-EGFR of 25 ng/mL EGF treated cells. 92

Figure 4.28: Dose-dependent temporal dynamics of phospho-FAK. 93

Figure 4.29: Ultrasensitive switch-like response during EGF-induced state transition. 95

Figure 4.30: Effect of gefitinib on cell viability. 95

Figure 4.31: Blockade of EGFR, turns OFF the ultrasensitive switch. 96

Figure 4.32: Gefitinib does not affect the EGF-induced state transition. 97

Figure 4.33: Distribution of molecular markers of EMT in each cell type. 99

Figure 4.34: The discrete energy-based diagram of cell state transition. 102

Figure 4.35: A possible hypothesis of the cell fate decision during the EGF-induced state transition. 104

Figure 5.1: TGF-β1 did not induce any change in the population distribution of MDA-MB-468 cells. 107

Figure 5.2: TGF-β1 modulated the EGF-induced cell state transition. 107

Figure 5.3: TGF-β1 promotes the evolution of Spindle and Cobble cells. 109

Figure 5.4: EGF-induced apoptosis in MDA-MB-468 cells. 110

Figure 5.5: Entropy of the input signal and the response of the EGF-induced state transition. 115

Figure 5.6: The mutual information of the EGF-induced state transition. 115

Figure 5.7: The channel capacity of EGF-induced state transition. 118

Figure 5.8: Mutual information and channel capacity of EGF-induced state transition in the presence of TGF-β1. 119

Figure 6.1: Graphical representation of the hierarchical model 131

Figure 6.2: Workflow of DEBay. 133

Figure 6.3: The GUI of DEBay. 134

Figure 6.4: Effect of noise in mRNA level on population-level fold change in gene expression. 136

Figure 6.5: Evaluation of the performance of DEBay for synthetic data sets. 137

Figure 6.6: Evaluating DEBay with time-independent synthetic data. 138

Figure 6.7: Evaluating DEBay with time-dependent synthetic data set-1. 141

Figure 6.8: Evaluating DEBay with time-dependent synthetic data set-2. 142

Figure 6.9: Evaluating DEBay with time-dependent synthetic data set-3. 143

Figure 6.10: Accuracy of the deconvoluted parameters of time-dependent synthetic data sets. 144

Figure 6.11: Deconvolution of real qPCR data using DEBay. 145

Figure 6.12: Evaluating DEBay with real time-dependent data. 147

Figure A-1: Standard curve of dead cell number estimation assay. 179

Figure A-2: Quality check of the dead cell number estimation assay. 180

Figure A-3: Standard curve of total cell number estimation assay. 181

Figure A-4: Optimal number of training objects per cell type. 182

Figure A-5: Quality check of the image classifier training. 183

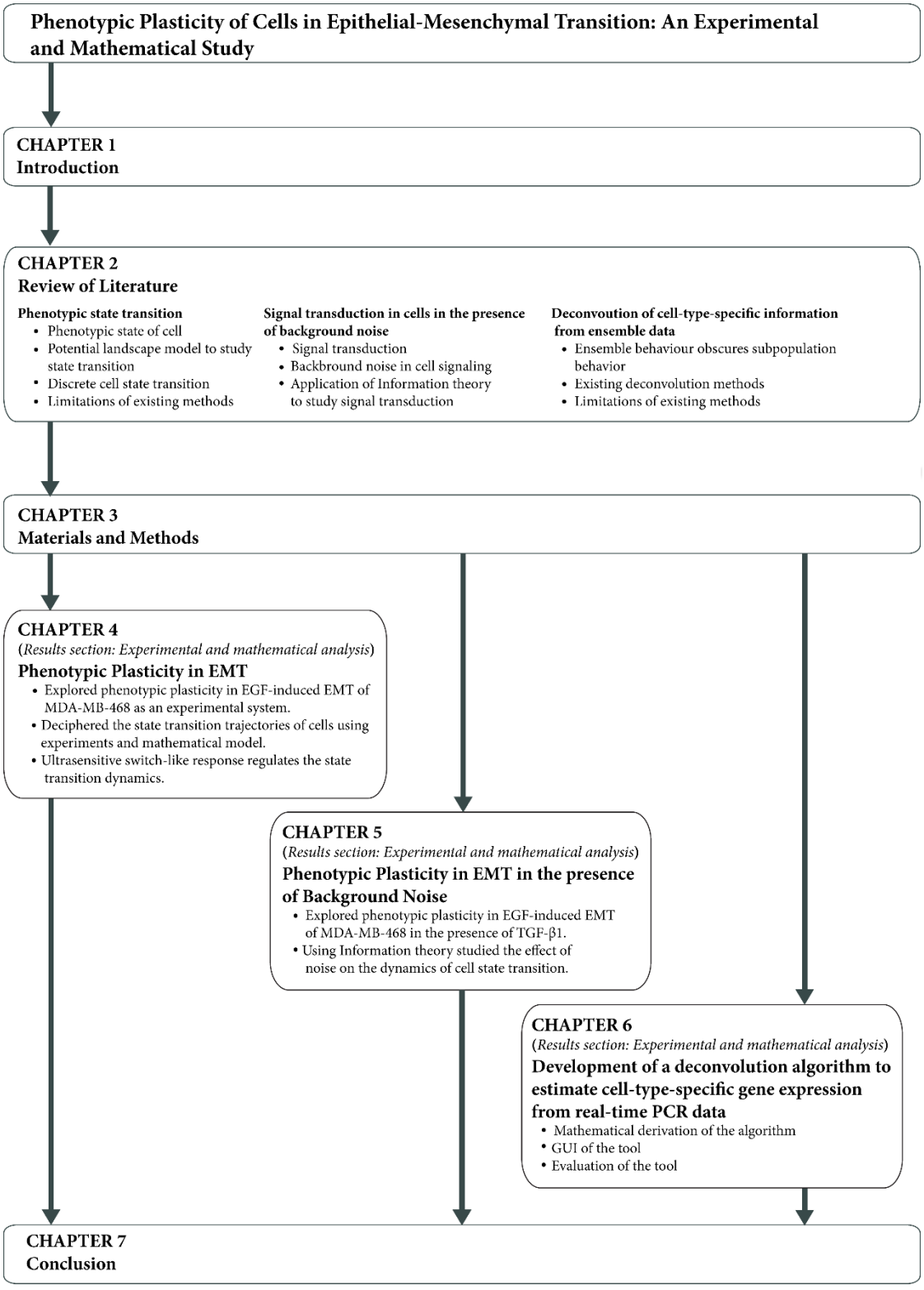
# Table of Abbreviations and Acronyms

| | |
|---|---|
| ANOVA | Analysis of variance |
| APS | Ammonium persulfate |
| BIC | Bayesian information criterion |
| BSA | Bovine serum albumin |
| C | Channel capacity |
| CB | Cobble cell state |
| CD24 | Cluster of differentiation 24 |
| CD44 | Cluster of differentiation 44 |
| cDNA | Complementary deoxyribonucleic acid |
| CR | Circular cell state |
| DAPI | 4′, 6-diamidino-2-phenylindole |
| DD | Dead cell state |
| DEBay | Deconvolution of ensemble through Bayesian approach |
| DEPC | Diethyl pyrocarbonate |
| DMEM | Dulbecco's modified eagle's medium |
| DMSO | Dimethyl sulfoxide |
| DNA | Deoxyribonucleic acid |
| DNase | Deoxyribonuclease |
| dNTP | Deoxynucleoside triphosphate |
| E-cadherin | Epithelial cadherin |
| ECM | Extracellular matrix |
| EDTA | Ethylenediaminetetraacetic acid |
| EGF | Epidermal growth factor |

| | |
|---|---|
| EGFR | Epidermal growth factor receptor |
| ELISA | Enzyme-linked immunosorbent assay |
| EMT | Epithelial-Mesenchymal transition |
| ERK | Extracellular signal-regulated kinase |
| ES | Spindle cell state |
| FAK | Focal adhesion kinase |
| FBS | Fetal bovine serum |
| FITC | Fluorescein isothiocyanate |
| GCC | GNU compiler collection |
| GRB2 | Growth factor receptor-bound protein 2 |
| GSK3β | Glycogen synthase kinase 3 beta |
| GUI | Graphical user interface |
| HCS | High-content screening |
| HRP | Horseradish peroxidase |
| $IC_{50}$ | Half maximal inhibitory concentration |
| JAK | Janus kinase |
| JNK | c-Jun N-terminal kinase |
| MAPK | Mitogen-activated protein kinase |
| MCMC | Markov chain Monte Carlo |
| MEK | Mitogen-activated protein kinase kinase |
| MET | Mesenchymal-Epithelial transition |
| MI | Mutual information |
| MinGW | Minimalist GNU for windows |
| MMP | Matrix metalloproteinase |
| mRNA | Messenger RNA |
| MTT | 3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide |
| N-cadherin | Neural cadherin |
| NF-κβ | Nuclear Factor kappa-light-chain-enhancer of activated B cells |
| NGEC | Normalized gene expression coefficient |

| | |
|---|---|
| NGF | Nerve growth factor |
| NUT sampler | No-U-Turn sampler |
| ODE | Ordinary differential equation |
| PAGE | Polyacrylamide gel electrophoresis |
| PBS | Phosphate-buffered saline |
| PCR | Polymerase chain reaction |
| PDF | Probability density function |
| PI | Propidium iodide |
| PI3K | Phosphoinositide 3-kinase |
| PMSF | Phenylmethylsulfonyl fluoride |
| PVDF | Polyvinylidene fluoride |
| qPCR | Quantitative polymerase chain reaction |
| RIPA | Radioimmunoprecipitation assay |
| RNA | Ribonucleic acid |
| RNase | Ribonuclease |
| rRNA | Ribosomal ribonucleic acid |
| SDS | Sodium dodecyl sulfate |
| TAE | Tris-acetate-EDTA |
| TBS | Tris-buffered saline |
| TEMED | Tetramethylethylenediamine |
| TGF-β1 | Transforming growth factor-beta 1 |
| TNF | Tumor necrosis factor |
| UTR | Untranslated region |

# Flow of the Thesis



CHAPTER 1

# Introduction

The phenotype of a cell refers to any observable physical or functional features of the cell. The change from one phenotype to the other is referred to as the phenotypic cell state transition or phenotypic plasticity. The cell state transition is observed in several biological processes like stem cell differentiation, B-cell differentiation, wound healing, and cancer metastasis (*1-4*). Aims of studying the cell state transition are constructing the evolutionary trajectories of cells and understanding the driving force of the cell state transition.

Several approaches are used to estimate the trajectories of the cell state transition from experimental data (*5-10*). However, there are some lacunae in the existing methods:

1. Several state transition processes in biology are reversible. Most of the available methods address only unidirectional irreversible state transitions.

2. Existing methods do not consider cell birth and cell death.

3. The rate of transition from one state to the other is often considered constant. However, these parameters vary with time and strength of the stimulus.

Much of the existing works have defined the state of the cell based on the molecular processes. In this work, we studied the dynamics of state transition at the phenomenon-level using EGF-induced Epithelial-Mesenchymal Transition (EMT) of MDA-MB-468 cells as an experimental system. EMT is a process during which the cells lose contact between neighboring cells and acquire migratory potential (*11*, *12*). EMT involves transitions between multiple phenotypic states and can be easily studied in the laboratory.

In this study, we defined the phenotypic states based on the morphology of cells. We observed three distinct morphologies in MDA-MB-468 cells. We call them as Cobble, Spindle, and Circular. We treated MDA-MB-468 cells with EGF and measured the population distribution of three cell types using quantitative image analysis. We developed a mathematical method to decipher the lineage trajectories of cells from the image analysis data. Our analysis showed that the cells followed a reversible path, and the dominant transition path of the cells is Cobble → Circular → Spindle → Cobble. We also investigated the process at the molecular level. The dynamics of the state transition were dependent on the dose of the EGF and the phosphorylation of EGFR. We observed an ultrasensitive switch involving phospho-EGFR, that controls the transition of cells in and out of the circular state.

In this work, we developed a discrete-time state transition method that uses population distribution data collected at discrete time points to decipher the dynamical trajectories of cells. Our model is generic and can be applied to any experimental system where the cells are categorized discretely. For example, cells can be categorized based on functional features like the migratory potential of cells,

scattering potential of cells. These phenotypic states can be experimentally measured, and the same modeling approach can be used to trace the transition path of the cells.

Cells are constantly exposed to a diverse signaling molecules in the external environment. Cells perform various cellular functions like cell division, growth, differentiation, and cell death based on the instructions from the external cues. The signal transduction in cell does not occur in isolation, rather in the presence of several other molecules. Multiple signals activate several signaling pathways within cells. These signaling pathways are interconnected to form a web of signaling networks. Therefore, there is always a possibility of signal interference across several pathways, creating a background chatter of signals.

Using the same experimental system of EGF-induced state transition, we investigated the effect of background noise in the cell state transition. We introduced background noise by adding suboptimal doses of TGF-β1. TGF-β1 at suboptimal dose did not induce any morphological change in cells, whereas when supplemented with EGF, it modulated the EGF-induced state transition. Using the concepts of information theory and statistical analysis, we showed that TGF-β1 increases the noise in the system and synergistically favors the transition towards Spindle and Cobble states.

In this work, we categorized cells based on the morphology, and we do not have any information on the gene expression pattern of these cellular states. Quantitative PCR is often used to study gene expression changes in cells. These different cell types could be sorted using flow cytometry and further analyzed in qPCR if the cell types were classified based on molecular markers. However, we cannot use this approach as we defined cell types in terms of morphology.

This issue is not specific to our experimental system. The same problem exists whenever there is a mixed population of cells as qPCR measures ensemble-average. The population-level measurement obscures the gene expression of sub-population

present in the population. These issues can be sorted with the help of computational deconvolution algorithms (*13-22*). All the existing methods consider the population-level gene expression as a linear combination of the gene expression of each subpopulation present in the population. However, there are lacunae in the existing methods:

1. Gene expressions are dynamic and are dependent on the treatment conditions like various doses of a drug, different duration of treatment. The available methods do not consider this scenario.

2. Most of the methods employ frequentist-approach that provides only the point estimate of the parameters and does not give an estimate of the probability distributions of the estimated parameters.

3. Existing methods do not give any information on the physical meaning of the deconvoluted parameters.

In the present work, we developed a computational tool, DEBay, to estimate the cell-type-specific gene expression data from the qPCR data of the mixed population. Our algorithm considers both experimental condition-dependent and experimental-condition independent gene expression cases. We used a Bayesian method of parameter estimation and estimated the probability distribution of parameters. Our method can be used in any experimental setup where the population distribution of cells is known.

The outcomes of this work are based on EGF-induced EMT of MDA-MB-468 cells. However, the experimental and the modeling strategy used in this study are comprehensive and can be employed to any cellular systems involving phenotypic state transition.

CHAPTER 2

# Review of Literature

## 2.1. Phenotypic state transition

Phenotype refers to the observable trait or attribute of a cell. It includes physical, functional features of the cell as a result of the interaction between the gene and the environment (*23*). Depending on the environment or the external cue, the cell switch from one phenotype to the other. This transformation is referred to as the phenotypic state transition or phenotypic plasticity (*24, 25*). The switching of cells to other phenotypes can also happen spontaneously in the absence of an external cue (*26*). Phenotypic plasticity does not involve a change in the genotype. Instead, it involves changes in gene expression possibly influenced by environmental factors or epigenetic regulation (*27, 28*).

Phenotypic plasticity of cells is observed during the differentiation of embryonic stem cells to other specialized cell types (*29*), the emergence of drug-resistant cancer cells (*30, 31*), and the development of cancer stem cells (*32, 33*). One of the typical examples of phenotypic plasticity is Epithelial-Mesenchymal Transition of cells.

## 2.2. Epithelial-Mesenchymal Transition

Epithelial-Mesenchymal Transition (EMT) is a state transition phenomenon that involves multiple phenotypic states (*4, 34, 35*). During EMT, the epithelial cells lose cell-to-cell contact and become more migratory-mesenchymal cells (Figure 2.1). The phenotypic changes involve the loss of apical-basal polarity, loss of cell-cell adhesion, reorganization of the cytoskeleton, and the acquisition of cellular motility (*36, 37*). The reverse process of EMT is known as Mesenchymal-Epithelial Transition (MET). Both EMT and MET are observed during early embryogenesis, wound healing, and cancer metastasis (*38, 39*).

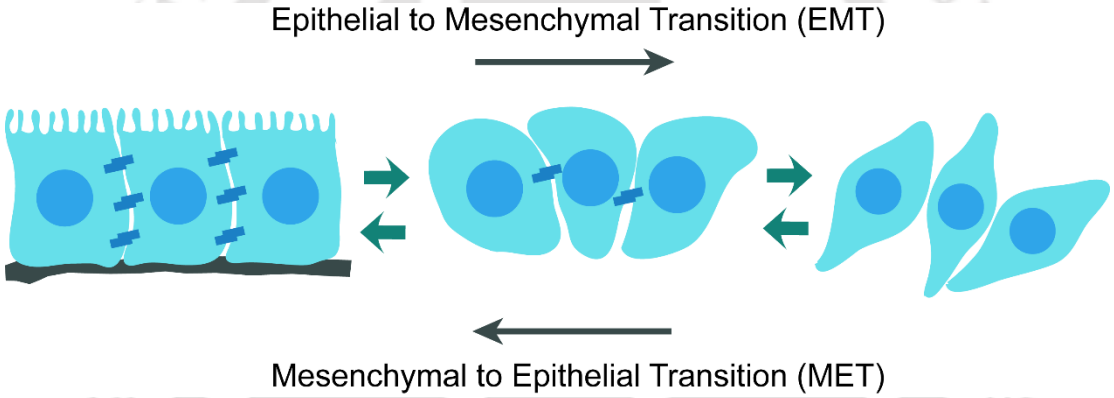



**Figure 2.1: Cell state transition during EMT and MET.**

### 2.2.1. Crucial players of EMT

The key players of EMT are the transcription factors SNAIL1 and ZEB1. These proteins bind to the promoter region of E-cadherin and block the transcription of E-cadherin. At the same time, they activate the transcription of mesenchymal markers like N-cadherin, Vimentin, and several other matrix metalloproteases (MMPs) (*11, 40*). This cadherin switching from E-cadherin to N-cadherin is the hallmark of EMT. Increased expression of the mesenchymal markers and MMPs facilitate the cells to degrade the extracellular matrix (ECM). Through this, the cells become more migratory, and they start invading the surrounding tissues.

Several microRNAs were reported to regulate EMT through the transcription factors SNAIL1 and ZEB1. Generally, microRNA binds to the 3' UTR of their target mRNA and either blocks the translation or promotes the degradation of the target mRNA (*41*). For example, miR-29b and miR-30a inhibit SNAIL1 expression (*42, 43*), and miR-205 inhibits ZEB1 expression (*44*). Thus, these microRNAs promote MET by repressing the EMT transcription factors. There exist several double-negative feedback loops between EMT transcription factors and the microRNAs, where the expression of one represses the other. The well-studied double-negative feedback loops are miR-34/SNAIL1 and miR-200/ZEB1, which are the master regulators of EMT (*4, 45, 46*). These key controllers of EMT are regulated by several signaling pathways activated by signaling molecules like TGF-β1 (*4, 40, 47*) and EGF (*48-50*).

### 2.2.2. Signaling pathways of EMT

EGF cascade its signal by binding to EGF receptors (EGFR) on the cell surface. The binding of EGF phosphorylates EGFR and activates several cardinal-signaling pathways of the cell. One such is the PI3K-Akt pathway. Activation of Akt inhibits GSK3β, which is a negative regulator of SNAIL1 (*51, 52*). Other signaling cascade activated by EGF is the MAPK. Phosphorylated EGFR activates SHCA, which recruits GRB2-SOS. This complex activates a series of kinases, RAS → RAF → MEK → ERK (*11*). Both these pathways promote EMT through the activation of the EMT transcription factors.

TGF-β1 induces EMT through Smad-dependent and -independent pathways. TGF-β1 binds to TGF-β type II receptors on the cell surface and phosphorylates TGF-β type I receptor (*53*). This event activates Smad2/3, which in turn binds to Smad4 and forms a complex. This complex translocates to the nucleus and activates the expression of EMT transcription factors (*40*). The Smad-independent arm of TGF-β1 signals through PI3K-Akt and MAPK pathway. Other signaling pathways like JNK, p38 MAPK, Notch signaling, NF-κβ were also reported to regulate TGF-β1 induced-EMT (*11, 40*).

These signaling pathways give us an idea about the signaling events happening at the molecular-level during EMT. To study the cell state transition, firstly, we need to define the cell state. Then we can use the knowledge from the signaling pathways to perturb the process and get more insights about the dynamics of cell state transition.

## 2.3. How are the cell states defined?

The phenotypic state of cells is often defined based on the relative expression of specific molecular markers (*6*, *31-33*, *54*). For example, in the case of EMT, the cells are categorized into multiple phenotypes based on the relative expression of E-cadherin and Vimentin (*4*). In cancer stem cells, the phenotypic states were defined based on the relative expression of CD44 and CD24 (*55*). In these cases, the cells were classified into specific phenotypes through experiments like flow cytometry and microscopic imaging. With the advent of high throughput experiments like microarray and RNAseq, the cells were also classified based on genome-wide gene expression data (*8*, *56-58*).

The other alternate way is to classify cells based on physical or functional features like change in morphology, change in the migratory potential of cells. These physical or functional properties are regulated by changes in the expression of several genes. The gene expression measurements are proxies to the phenotype of the cells, while the functional features are the direct measure of the cell's phenotype. Therefore, it will be pertinent to categorize cells directly based on functional features through techniques like microscopic image analysis (*59-61*). For example, Kimmel et al. (*62*) have quantitatively measured the motility features of cells from time-lapse images.

Irrespective of the method used to define the cell states, the main goal is to understand the time evolution of different phenotypic states from the experimental data. Several mathematical approaches, like the potential landscape model, Markov's discrete state

transition model, have been developed to understand the dynamics of cell state transition.

## 2.4. The potential landscape of cell state transition

### 2.4.1. Waddington's epigenetic landscape

In 1957, Waddington (*63*) proposed a metaphor of the epigenetic landscape that depicts the evolution of cells during embryonic development. Waddington visualized the landscape as a continuous series of hills and valleys. Here, the hills represent the higher potential state, while the valleys represent the lower potential states. Each valley corresponds to a phenotypic state of the cell, and the branching valleys resemble the various developmental changes in the cell. The ridges between the valleys prevent the cells from moving into other valleys, thus giving directionality to the cell differentiation. During the developmental process, the pluripotent cells move from a higher potential to the lower potential, and the fate of the cell depends on the valley it lands.

The essential aspects of Waddington's landscape are: 1) the time evolution of cell from a higher potential to a lower potential and 2) alternate paths leading to different lower potential states. This landscape metaphor can be created using dynamical models for molecular circuits of cells that regulate the state transition.

### 2.4.2. Bifurcation in the cell state transition

A dynamical model of a cell's regulatory circuit is a system of Ordinary Differential Equations (ODEs). Each ODE describes the dynamics of the entities involved in the molecular circuit. From these ODEs, the trajectories of the system can be drawn. The trajectories in the phase plane show the directional time evolution of the system, and the trajectories converge into the stable steady states of the system. Each stable steady-state is a phenotypic state of the cell. In cell biology, each axis of the phase plane denotes the concentration of a molecule in the cell signaling circuit. Whenever the cell

is perturbed from its initial state, the cell moves towards the stable steady-state by following the trajectories.

The change in the qualitative behavior of the phase plane is referred to as bifurcation (*64*). For example, depending on a critical parameter of the dynamical model, the number of possible steady states of the system may change. Usually, the external signal influences the critical parameter.

Let us consider an isogenic population of cells. The concentration of a molecule $x$, determines the phenotype of the cell, and an external signal $S$, regulates $x$. The dynamical model of this system contains only one entity, $x$. Let us assume that the bifurcation plot of this system resembles Figure 2.2. Here, the horizontal axis is the concentration of $S$ and the vertical axis is the steady-state of $x$. The basal level expression of $x$ in each cell represents the initial position of the cells in the phase plane. Though the population of cells is isogenic, the basal level expression of $x$ in each cell varies because of the inherent noise in the gene expression. Therefore, all the cells in the phase plane will be in a different initial position.

When $S = S_1$, the system is monostable, and the trajectories of all the cells will travel towards the steady-state, and all the cells will be of phenotype A. Similarly, when $S = S_3$, the system is monostable with a different higher steady-state. Therefore, the trajectories of all the cells will move towards the higher steady-state and all the cells will be of phenotype C. However, at $S = S_2$ the system is tristable, and the cells can exist in any of the three phenotypes. Depending on the initial position of cells in the phase plane, the cells move towards any one of the three steady-states A, B, or C and acquire the corresponding phenotype.

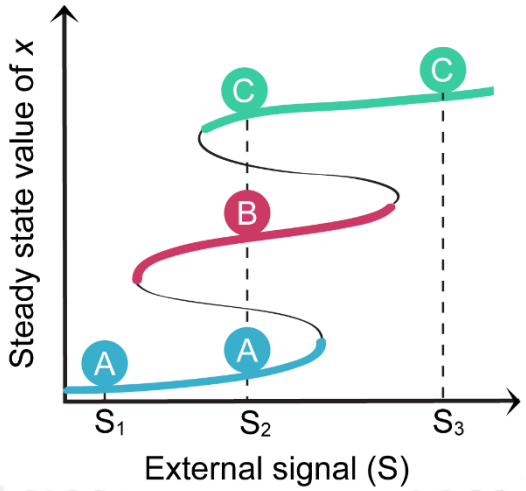



**Figure 2.2: Bifurcation plot.**
$S_1, S_2$, and $S_3$ are the strength of the external signal. The vertical axis is the steady-state value of $x$. The thick colored lines represent stable steady-states, and the thin black line represents the unstable steady-state. A, B, and C are the different phenotypes of the cell.

Bifurcation is often used to explain the existence of multiple stable phenotypes in EMT. Tian et al. (*34*) have constructed an ODE-based mathematical model for the core regulatory circuit of EMT involving two double-negative feedback loops, the SNAIL1/miR-34 and ZEB1/miR-200. Through bifurcation analysis, Tian et al. (*34*) have shown that the cells can exist in three stable phenotypic states: Epithelial, partial-EMT/Hybrid EMT, and Mesenchymal based on the strength and duration of the input signal. Each of the double negative feedback loops functions as a bistable switch. The SNAIL1/miR-34 loop controls the initial transition from epithelial to partial-EMT, while the ZEB1/miR-200 loop regulates the transition from partial-EMT to mesenchymal state (*34*). The same behavior has been experimentally validated by Zhang et al. (*4*) in MCF-10A cells. Several others have used bifurcation analysis to study cell-to-cell heterogeneity during EMT (*65-67*).

### 2.4.3. The potential landscape of the cell

Waddington's landscape metaphor of cell differentiation was later created from the dynamical model of the cell signaling circuit. The potential landscape is a multi-

dimensional function of the concentrations of several molecules that defines the state of the cell. This landscape will be useful in understanding the lineages of the cell during state transition.

Let us consider a cell signaling circuit with one variable for the sake of simplicity. Let us assume that the concentration of the molecule determines the phenotype of the cell, and we will call this molecule as the phenotypic determinant. Therefore, the potential landscape will be a function of the phenotypic determinant and can be visualized in a two-dimensional plot (Figure 2.3). Consider the example in Figure 2.3a. Here, there is only one stable steady-state, which is the lowest potential in the landscape. The cells with different initial values gradually flow towards this stable steady-state, and all the cells will be of the same phenotype. Now, if the system is perturbed with some external signal, then the potential landscape changes, leading to a new stable steady-state with a higher value of the phenotypic determinant. All these cells now flow towards this new steady-state as this is the new lowest potential and acquire different phenotypes.

There can be a situation where the external signal, changes the potential landscape such that there emerge two new minima (Figure 2.3b). These minima are stable steady-states that correspond to two different phenotypic states. This is an example of bifurcation, where the qualitative behavior of the system changes depending on the external signal. Initially, the system was monostable, and all the cells were of the same phenotype. Though the phenotypic states of the cells were the same, there exists a considerable amount of cell-to-cell variation at the molecular level because of the stochastic nature of biochemical reactions. Now, the perturbation from the external signal converted the monostable system to a bistable system. Depending on the noise-level, the cell will move to either of the new steady states.

Cells can move across different local minima, even in the absence of an external signal (*6*). Figure 2.3c shows a system with two local minima separated by a small energy barrier. Here, the potential landscape does not change, but the inherent biochemical noise allows cells to switch between the two local minima.

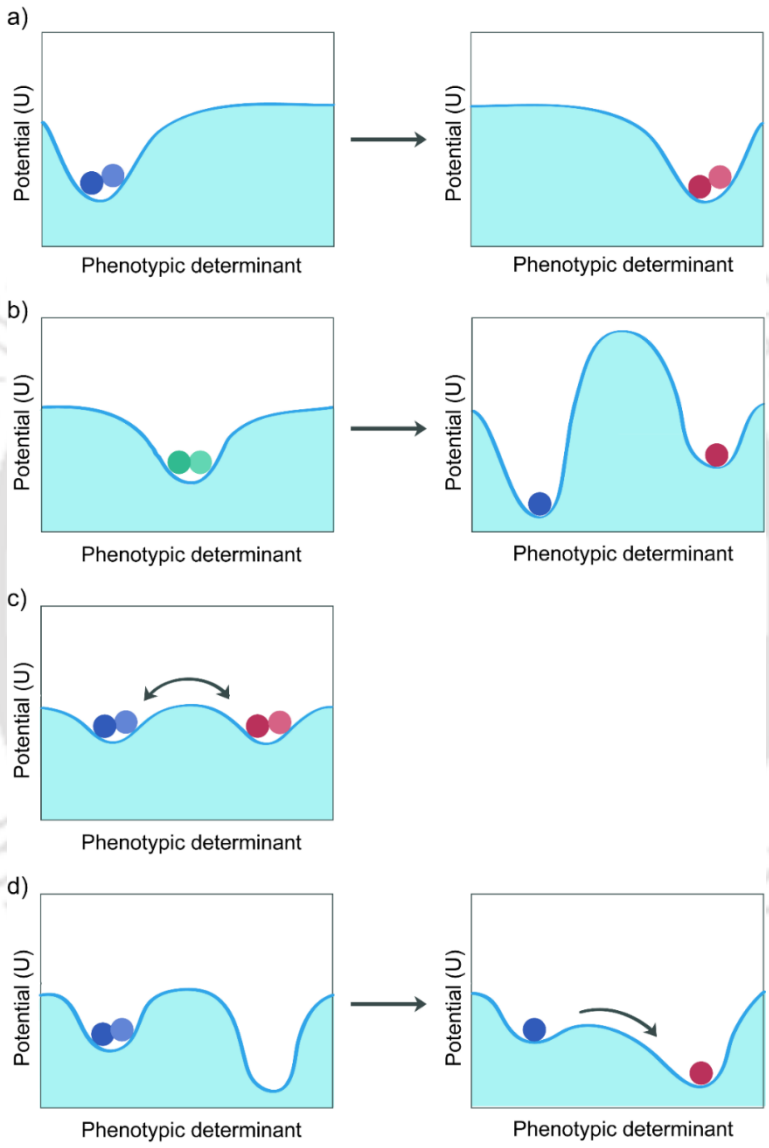



**Figure 2.3: The potential landscape of the cell.**
The phenotypic state is defined in terms of the concentration of a molecule called the phenotypic determinant. The local and global minima are different phenotypic states of the cell. Colored circles represent cells. a-b) An external signal changes the potential landscape of the cell. c) Cells switch between the two states due to the stochastic fluctuations in the molecular processes. d) The external signal changes the potential landscape, thereby favoring the noise-induced cell state transition.

The external signal, as well as the noise, can together drive the transition of cells between multiple phenotypic states. Consider the example in Figure 2.3d. Here, the system has two local minima with all the cells in one of the local minima. The energy barrier between the two local minima is enormous such that the noise in the system could not favor the stochastic transition to the other phenotypic state. However, when the external signal is given, it reduces the energy barrier between the two minima. Now, cells can flow to the other state.

### 2.4.4. Construction of the potential landscape

In several studies, the phenotypic states of the cells are defined based on the expression of molecular markers. Here, ODEs are used to define the changes in the expression of molecular markers, and the potential landscape is a function of these markers. From the concepts of dynamical systems theory, the potential landscape of the cells is calculated from the following relation (*64*),

$$\dot{\boldsymbol{x}} = -\nabla U(\boldsymbol{x})$$

where $\dot{\boldsymbol{x}}$ is the vector of ODEs and $U(\boldsymbol{x})$ is a continuously differentiable single valued scalar function of $x$. This potential can be derived only for gradient systems, where the energy is conserved. However, most of the biological systems follow non-equilibrium thermodynamics, and the energy is not conserved.

The other approach is to construct the pseudo-potential landscape using the Boltzmann distribution. Consider a system with $n$ different phenotypic states that represent different energy levels. If the average energy remains constant, then using Boltzmann distribution, the energy of the $i^{th}$ state is,

$$U_i = -\ln(p_i)$$

here $p_i$ is the steady-state probability of the system in $i^{th}$ state. This relation emphasizes that the occupancy of lower energy states will be greater than the higher

energy states. The above equation can be used even for a non-equilibrium system at steady-state. Therefore, from the probability distribution of different phenotypic states in the steady-state, we can compute the pseudo-potential landscape. Here, $U$ is dimensionless and is different from the potential in equilibrium thermodynamics and, therefore, referred to as pseudo-potential.

Several studies on cell state transition have used this approach to construct the potential landscape (*68-71*). Li et al. (*72*) have used this approach and constructed the potential landscape for the core EMT circuit involving miR-200, SNAIL1, and ZEB1. Using this landscape, Li et al. (*72*) have analyzed the energy barriers around steady-states that pose difficulty for the transition of cells from one state to the other. Using a similar approach Biswas et al. (*73*) have calculated the residence time of cells in each state during EMT.

The potential landscape approach is useful when we have an idea about the molecular processes governing the state transition. The expression of molecular markers is usually measured through experiments like flow cytometry, microarray. Usually, the phenotypic states are defined based on the range of expression of these markers rather than a single value. However, in the potential landscape model, the phenotypic states correspond to a specific value of the expression of molecular markers. These issues can be sorted if the cell states are defined discretely, and the dynamics of cell state transition can be studied using the appropriate mathematical models.

## 2.5. Discrete cell state transition

The discrete state transition is often considered as a Markov's process. In Markov's process, the cells can switch across different states stochastically, and the transition from one state to another state depends only on the current state of the system (*74*). Let us consider a three-state Markov's process, as shown in Figure 2.4a. At a given time point, a cell can be in any of the three states, A, B, and C. In the next time interval,

the cell can either stay in the same state or stochastically jump to the other two states with some probability. The transitions can be sequential, as shown in Figure 2.4b. There can also be an absorbing state in a Markov process (Figure 2.4c). The cells cannot jump to other states on reaching the absorbing state.

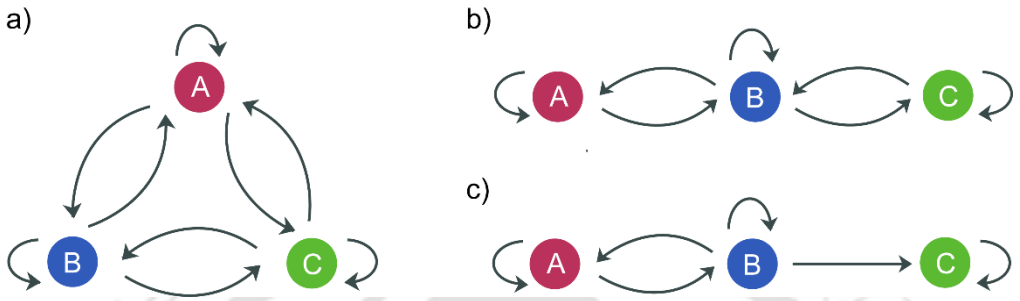



**Figure 2.4: Discrete Markov model for cell state transition.**
A three-state Markov model where a cell can move from one state to the other state (a) directly, (b) sequentially. (c) State transition with an absorbing state.

If there are $n$ different cell states $S_1, S_2, \ldots S_n$, then the probabilistic state transition in the time interval $t$ to $t + \Delta t$ can be written as,

$$\begin{pmatrix} Q_1 \\ Q_2 \\ . \\ Q_n \end{pmatrix}_{t+\Delta t} = \begin{pmatrix} P_{11} & P_{21} & . & P_{n1} \\ P_{12} & P_{22} & . & P_{n2} \\ . & . & . & . \\ P_{1n} & P_{2n} & P_{3n} & P_{nn} \end{pmatrix} \times \begin{pmatrix} Q_1 \\ Q_2 \\ . \\ Q_n \end{pmatrix}_t \tag{2.1}$$

where $[Q_1, Q_2, \ldots Q_n]$ are the probability of being in the state $S_1, S_2, \ldots S_n$, respectively. $P_{ij}$ is the probability of transition from $i^{th}$ state to the $j^{th}$ state. $[P_{ij}]_{i,j}^{n}$ is a $n \times n$ probability matrix, such that $0 \leq P_{ij} \leq 1$ and $\sum_j^n P_{ij} = 1$.

In a Markov process, the system is conserved; i.e., the total number of cells remains constant throughout the process. $P_{ii}$ represents the probability of staying in the same state and does not represent cell division. However, cell death can be incorporated by considering cell death as an absorbing state.

The vector notation of equation (2.1) is,

$$\boldsymbol{Q}(t+\Delta t)=\boldsymbol{P}\times\boldsymbol{Q}(t)$$

Usually, in experiments, we measure the fraction of cells in each state. If the sample size is large, then the fraction of cells in a specific state can be approximated to the probability of being in the state. Therefore, equation (2.1) can be re-written as follows,

$$\boldsymbol{F}(t+\Delta t)=\boldsymbol{P}\times\boldsymbol{F}(t) \tag{2.2}$$

where $\boldsymbol{F}=[F_1\ F_2\ \ldots\ F_n]$. $F_i$ is the fraction of cells in the $i^{th}$ state.

The main goal is to estimate $\boldsymbol{P}$, from which the most probable transition of cells can be traced out. If we have an estimate of the transition of cells from one state to another state at a single cell-level, then the estimation of $\boldsymbol{P}$ becomes much more straightforward. In experiments like live cell imaging, the cells are continuously monitored, and the transition to other states is collected at single-cell resolution. The probability of transition from $i^{th}$ state to $j^{th}$ state is, $P_{ij}=N_{ij}\Big/\sum_{j=1}^{n}N_{ij}$; where $N_{ij}$ is the number of cells moved to $j^{th}$ state from $i^{th}$ state. However, many of the experiments, like flow cytometry and fixed cell microscopic imaging, produces aggregate data. The population distribution of cells in different states are measured at discrete time points. In a time homogenous Markov process, $\boldsymbol{P}$ remains constant over time. If the data are collected at discrete time points, $T=t_1,t_2,\ldots t_k$, then equation (2.2) becomes,

$$\boldsymbol{F}(t_k)=\boldsymbol{P}^k\times\boldsymbol{F}(t_1)$$

$$\boldsymbol{P}=\left(\boldsymbol{F}(t_k)\,\boldsymbol{F}(t_1)^{-1}\right)^{1/k}$$

The computation of $\boldsymbol{P}$ from the above equation requires computing the $k^{th}$ root of a matrix provided $\boldsymbol{F}(t_1)$ is invertible. If we start with the pure population of different

cell state then $\boldsymbol{F}(t_1)$ will be an identity matrix and invertible. Though a $\boldsymbol{F}(t_1)$ with a mixed population can be invertible, but achieving such a titrated sample in an experiment is not realistic. Moreover, $\boldsymbol{P}$ should be a stochastic matrix, satisfying the conditions, $0 \leq P_{ij} \leq 1$ and $\sum_j^n P_{ij} = 1$. However, the $k^{th}$ root calculated from the above equation need not necessarily be a stochastic matrix (*75*). To overcome this issue, Buder et al. (*5*) have regularized the root matrix to a stochastic matrix using the quasi-optimization algorithm.

The other method to compute $\boldsymbol{P}$ is through optimization. The true population distribution of cells and the observed population distribution of cells deviates by an error margin, $\boldsymbol{e}(t + \Delta t)$. Therefore, following the equation (2.2), the transition of cells between time points $t$ and $t + \Delta t$ is,

$$\boldsymbol{F}(t+\Delta t) - \boldsymbol{P} \times \boldsymbol{F}(t) = \boldsymbol{e}(t+\Delta t)$$

The probability matrix $\boldsymbol{P}$, remains constant for a time homogenous process. Therefore, $\boldsymbol{P}$ can be estimated through simultaneous optimization across all time points,

$$\min_{\boldsymbol{P}} \sum_{t \in T} \sum \left[ \boldsymbol{e}_{t+\Delta t}^T \, \boldsymbol{e}_{t+\Delta t} \right]$$

where $T = t_1, t_2, \ldots t_k$.

Through this method, $\boldsymbol{P}$ can be estimated from the aggregate data. However, like any other regression-based methods, this method is also constrained by the size of the data. The number of observed data points should be high enough to avoid the overfitting of data. Also, long time-series data may suffer from the multi-collinearity problem (*76*).

These approaches use fractions of cell populations in each state to estimate the probability matrix. Farahat and Asada (*9*) have proposed a Bayesian method to

estimate $\boldsymbol{P}$ from the number of cells in each state. The probability matrix is estimated based on the flow of cells from one state to the other state. For a given $\boldsymbol{P}$ for $t$ to $t + \Delta t$, there can be multiple different flow matrix that could explain the observed number of cells at $t + \Delta t$.

Let us consider that the number of cells in $n$ different cell states, $\boldsymbol{N} = [N_1, N_2, \dots N_n]$. Let $F_{ij}$ be the number of cells moving from $i^{th}$ state to $j^{th}$ state in a time interval $t$ to $t + \Delta t$. Similarly, the flow of cells from $\boldsymbol{N}(t)$ to $\boldsymbol{N}(t + \Delta t)$ can be represented by a square matrix, $\left[F_{ij}\right]_{i,j}^{n}$ such that $\sum_{j=1}^{n} F_{ij} = N_i(t)$ and $\sum_{i=1}^{n} F_{ij} = N_j(t + \Delta t)$. For a given $\boldsymbol{N}(t)$ and $\boldsymbol{N}(t + \Delta t)$, multiple possible flow matrices are possible. Let the all possible flow matrix set be $\boldsymbol{\varphi}(t)$. Now, the probabilistic state transition given multiple possible flow matrix can be represented by the multinomial distribution,

$$\Pr(\boldsymbol{N}(t+\Delta t)\,|\,\boldsymbol{N}(t), \boldsymbol{P}) = \sum_{\boldsymbol{F}(t)\in\boldsymbol{\varphi}(t)} \left( \prod_{i=1}^{n} N_i(t)! \prod_{j=1}^{n} \frac{P_{ij}^{F_{ij}(t)}}{F_{ij}(t)!} \right)$$

The above equation is the multinomial likelihood function. $P_{ij}$ is estimated by maximizing the likelihood function for a given $\boldsymbol{N}(t)$ and $\boldsymbol{N}(t + \Delta t)$. However, computing $\boldsymbol{\varphi}(t)$ is computationally intensive and it becomes a challenging task when $\sum_{i=1}^{n} N_i$ is very high. Farahat and Asada (*9*) have used a Gibbs sampler to estimate the likelihood function, where the multinomial distribution is approximated to the constrained multivariate gaussian distribution.

The state transition trajectories can also be measured through ODE based model from the discrete-time population distribution data. The cell states can be represented as fractions of cells flowing from one state to the other state. The state transition diagram shown in Figure 2.4a can be modeled as follows.

$$\frac{df_A}{dt} = k_{BA} \times f_B + k_{CA} \times f_C - k_{AB} \times f_A - k_{AC} \times f_A$$
$$\frac{df_B}{dt} = k_{AB} \times f_A + k_{CB} \times f_C - k_{BA} \times f_B - k_{BC} \times f_B$$
$$\frac{df_C}{dt} = k_{AC} \times f_A + k_{BC} \times f_B - k_{CA} \times f_C - k_{CB} \times f_C$$

here, $k_{ij}$ is the rate constant for the transition from $i^{th}$ state to $j^{th}$ state; $i, j = A, B,$ and $C$.

For a system with $n$ number of cell states, the ODEs can be represented as follows.

$$\dot{\boldsymbol{f}}(t) = \boldsymbol{K} \times \boldsymbol{f}(t)$$

where $\dot{\boldsymbol{f}}(t)$ is the vector of derivatives of $\boldsymbol{f}$. $\boldsymbol{K}$ is the square matrix with the rate constants for the transition of cells from $i^{th}$ state to $j^{th}$ state. The rate constants can be estimated by fitting the observed data to the model. There are several parameter estimation tools for ODE based systems (*77, 78*). Zhou et al. (*79*) have used this approach to understand the dynamics of state transition in cancer cells.

All these different approaches have certain limitations and were widely used to study the cell state transition phenomenon. Briefly, the phenotypic state transition happens through the intervention of external signal or inherent biochemical noise or both the external signal and the biochemical noise. Majority of the studies use a single external signaling molecule to induce state transition of cells. However, the cells are regularly exposed to a dynamic environment with multiple different signaling molecules in the external environment. In such cases, how does the cell respond to such diverse input signals?

## 2.6. Signal transduction in cells

The information flow from the external environment starts with receptor-ligand interaction that triggers a series of downstream signaling, eventually leading to specific cellular responses. Multiple signals in the external environment, activate

several such signaling cascades, and these signaling pathways are interconnected to form a web of signaling networks (*80, 81*). The signal transduction from a specific input signal does not occur in isolation, rather in the presence of several other molecules. Therefore, there is always a possibility of signal interference across several pathways (*82, 83*).

### 2.6.1. How does the cell distinguish diverse input signals?

The different input signals converge to a few molecules that form the node or hub of the cell's signal processing machinery. From the dynamics of the activation status of these molecules, the cells decode the information in the input signal and respond accordingly. This gives rise to the hourglass model of signal transduction, where multiple input signals activate the same node molecule but generate a different response (*84, 85*). For example, Akt is one such hub molecule involved in diverse responses like cell survival, cell cycle regulation, protein synthesis, and cellular metabolism (*86, 87*). Murray (*88*) has shown that forty different cytokine receptors signal through just four combinations of Janus Kinase (JAK).

The input message is often encoded in the identity, concentration, and the temporal dynamics of the signaling molecules (*89*). For example, EGF and NGF activate the same ERK signaling pathway but generates a distinct response. EGF induces transient activation of ERK and promotes cell proliferation, whereas NGF produces sustained ERK activation and induces cell differentiation. The message from EGF and NGF is encoded in the temporal dynamics of ERK (*90*).

The cells employ several conserved network motifs to decode the message in the dynamics of the signaling molecules. Well studied and explored motifs are feedforward, positive feedback, negative feedback, and incoherent feedforward (*91, 92*).

Signal transduction in the presence of multiple input stimuli has been experimentally studied (*93-95*). Natarajan et al. (*83*) have used twenty-two receptor-specific ligands in a combinatorial approach and showed that the response is non-additive for many combinations of the input signal. As discussed earlier, the response depends on the dynamics of the signaling molecules and the network architecture of the pathway. In an extensive interconnected signaling network, the signal transduction of specific input stimuli depends on the availability of other signaling molecules and the status of the dynamics of other signaling pathways (*89*), i.e., the network state of the cell (*96*). For example, JNK is reported to be anti-apoptotic (*97*), pro-apoptotic (*98*), and in some cases, it is reported to be not involved in apoptosis (*99*). Using a large scale experimental study, Janes et al. (*100*) have shown that the fate of the cell does not only depend on the dynamics of JNK, instead the network state of the cell.

## 2.7. Background noise in signal transduction

The signaling pathways of the cell are interconnected, leading to potential signal interference across signaling pathways (Figure 2.5). Extracellular media often contains several growth factors, hormones, and nutrients. Because of the higher degree of interconnected pathways, the presence of these molecules generates a continuous chatter of signals or background noise.

Several studies on the crosstalk of signaling pathways use different input signals that trigger specific responses and compare the outcome of a specific signal with the combination of other signals. Here, the doses of the molecules used were usually optimum to high (*83, 101*). However, cells usually receive a specific signal to perform a distinct task in the presence of suboptimal amounts of other signals. Therefore, the signaling of each input signal generating a distinct response should be studied with the background noise.

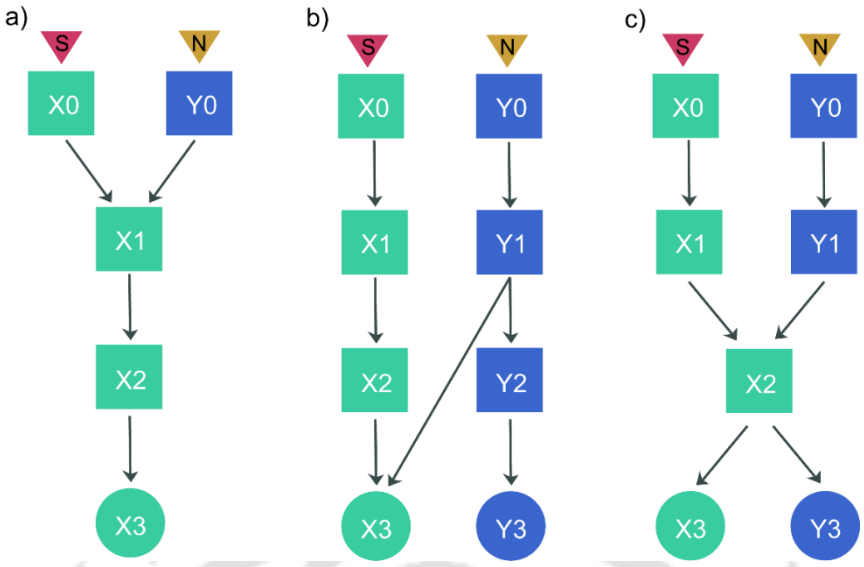



**Figure 2.5: Different ways of interconnections in signaling pathways of the cell.** Here, S is the primary signal, and N is another signaling molecule that activates signaling pathways that have interconnections with S. Continuous presence of N will create background noise to S.

The cells have evolved several mechanisms to overcome background noise. Cells fix a threshold level of the signal based on the noise and respond only when the signal goes beyond the threshold. This mechanism is usually achieved by oligomerization of signaling molecules that trigger signaling beyond the threshold, while the background noise cannot (*102*). Cells employ multi-component signaling, where different signaling molecules assemble and promote signal transduction. Therefore, binding of other inappropriate signaling molecules cannot perpetuate the signal (*102*). The background noise also aids cells in signal transduction. Using a Boolean network model, Domedel-Puig et al. (*103*) have shown that specific levels of background chatter give rise to different responses given the same input signal. Thus, the background chatter channelizes the information flow through the cell signaling network.

## 2.8. An information-theoretic approach to study signal transduction

The cells receive a myriad of cues from the environment and process those cues to perform various cellular activities. In cell signaling, a signaling molecule (or ligand) binds to its cognate receptor and triggers a cascade of reactions through a pathway. The signaling triggered by the ligand depends on its dose or concentration, and a cell can discriminate different levels of signaling through the dynamical processes in the pathway. However, noise in molecular processes affects signal transduction and impedes the ability of cells to distinguish different signals.

Consider a specific cell signaling pathway as an input-output system. Here, the concentration or dose of the ligand molecule is the input, and change in expression of a gene or change in the level of phosphorylation of a molecule is the output. In the absence of any noise, each input (S) would generate a unique response (R). For example, in Figure 2.6, the red line represents the input-output relation. When the input is $s_1$ the output of the pathway is $r_1$. However, due to noise in the signaling pathway, there will be variability in the input-output relation (the shaded region around the red line).

The response of the signal $s_1$, would vary from $r_{1,L}$ to $r_{1,H}$ around $r_1$. Similarly, for another input signal $s_2$, the response would vary from $r_{2,L}$ to $r_{2,H}$ around $r_2$. Though the system has noise, the cells can generate distinct responses to the two input signals. The response zones of these input signals do not overlap (Figure 2.6a). However, when the noise is high, as in Figure 2.6b, the response zones of these signals may overlap. This would reduce the cells ability to differentiate two inputs, $s_1$ and $s_2$.

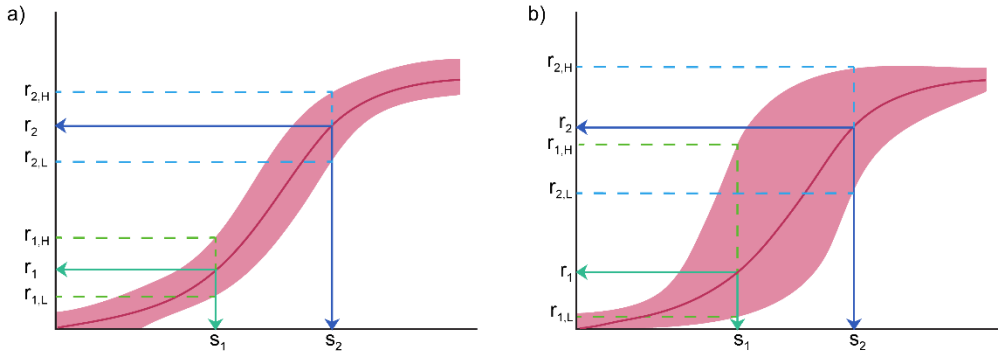



**Figure 2.6: Signal discrimination in cells.**
*S* and *R* are the concentrations of input signal and response of the cell, respectively. The solid red line is the relation between the input and response, and the red shaded region is the noise. a) Each concentration of the response corresponds to a unique concentration of the input signal. b) Each concentration of the response does not always correspond to a unique concentration of the input signal. The overlapping region of the response maps to both the input signal.

With the rapid advancement in single-cell experiments, we are now being to probe and understand stochasticity in molecular processes and its effect on cell-to-cell variability. Experiments using flow cytometry (*104*), immunofluorescence, fluorescent biosensors for kinase pathways (*105-108*) are being used to understand molecular cell signaling in individual cells. These techniques allow us to investigate the effect of noise on cell signaling and how a cell handles such noise. Such investigations need mathematical formulations to understand cellular communications quantitatively. Information theory developed for the communication system is now widely used for such purposes. Here, we briefly discuss the vital mathematical concepts of information theory and their uses in cell signaling.

Consider, $X$ as a discrete random variable that can take any one of the possible values $(x_1, x_2, \dots x_m)$, with a probability $p(x_i)$. For example, the concentration of a ligand activating a pathway can be considered as a random variable $X$, and in an experiment, the ligand can be used at different concentrations, $x_1, x_2, \dots x_m$.

Shannon defined entropy of a random variable as (*109*),

$$H(X) = -\sum_{x \in X} p(x) \times \log_n p(x) \tag{2.3}$$

Usually, in digital communication, $n = 2$ is used in this equation, and the unit of Shannon's entropy is bits. Hereafter, Shannon's entropy will be called entropy. $H(X)$ is a concave function and is maximum when all the states of $X$ are equiprobable. For example, suppose $X$ can take two values $H$ and $T$, just like in a coin toss. When $p(H) = 1$, or $p(T) = 1$, $H(X)$ is minimum at $H(X) = 0$. $H(X)$ has its maximum value, $H(X) = 1$, when $p(H) = p(T) = 0.5$ (Figure 2.7). Therefore, entropy is a measure of uncertainty of a random variable. The uncertainty is maximum when all the values or states of the random variables are equiprobable.

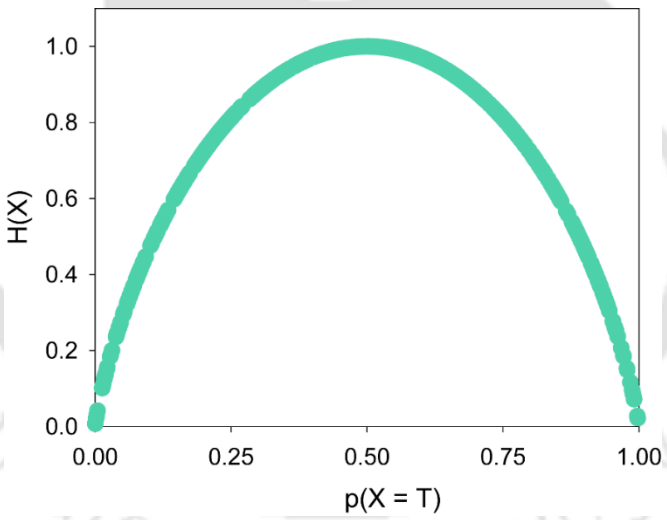



**Figure 2.7: Entropy plot of a binary discrete random variable.**
$X$ is a binary discrete random variable that can take either H or T. In this figure, the entropy of $X$ is shown as a function of probability of $X = T$. $H(X)$ is maximum when both the states of $X$ are equiprobable.

For a continuous random variable $X$, entropy is defined as,

$$h(X) = -\int_X f(x) \log f(x) dx$$

Here, $f(x)$ is the probability density function (PDF) of $X$.

For two discrete random variables, $X$ and $Y$, the conditional entropy is defined as,

$$H(X|Y) = -\sum_{i,j} p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(y_j)}$$

Here, $p(x_i, y_j)$ is the joint probability of $X = x_i$ and $Y = y_j$. $p(y_j)$ is the marginal probability of $Y = y_j$.

The statistical dependency between these two random variables can be measured by mutual information (MI) that is defined as (*110*),

$$I(X;Y) = H(X) - H(X|Y) \tag{2.4}$$

This can be expanded as,

$$I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y) \tag{2.5}$$

Following the definition of entropy and conditional entropy, we can rewrite equation (2.4) as,

$$I(X;Y) = \sum_{i,j} p(x_i, y_j) \times \log_2 \frac{p(x_i, y_j)}{p(x_i) \times p(y_j)} \tag{2.6}$$

Suppose $X$ and $Y$ are input and output of a communication system, respectively. Then following equation (2.4), we can say that MI is the reduction in uncertainty about $X$ given $Y$. Equivalently, the MI measures how accurately the value of $X$ can be determined based upon the value of $Y$.

Note that MI is symmetric in a sense that we can also write it as

$$I(X;Y) = H(Y) - H(Y \mid X) \tag{2.7}$$

Therefore, $Y$ gives as much information about $X$, as $X$ gives about $Y$.

Further, it can be shown that

Minimum of MI, $\min(I(X;Y)) = 0$

Maximum of MI, $\max(I(X;Y)) = \min(H(X), H(Y))$

We can use the concept of mutual information to analyze cell signaling quantitively. Note that mutual information is a statistical measure that depends only on the input and output measurements and does not depend upon the underlying details of the communication system. This makes MI very useful in quantitative analysis of cell signaling as we often have limited knowledge of the details of the pathway.

Suppose $S$ is the input signal to a cell, and $R$ is the corresponding response. As cell signaling is stochastic, both $S$ and $R$ are random variables. When $S$ is independent of $R$, then $H(S|R) = H(S)$. So mutual information, $I(S;R) = 0$ (Figure 2.8a). That means there is no correlation between $R$ and $S$. On the other hand, when there is one unique realization of $R$ for each realization of $S$, then $I(S;R) = H(S) = H(R)$ (Figure 2.8b). So, when there is complete dependence between input and output, the uncertainty in the output variable is the same as that of the input variable.

However, most input-output systems lie somewhere in between these two extremes. There exists a dependency given by the conditional probability, $p(S = s|R = r)$. In such circumstances, mutual information gives us an estimate of statistical dependency between $R$ and $S$. In such cases, the mutual information $I(S;R) \leq min(H(S), H(R))$ (Figure 2.8c).

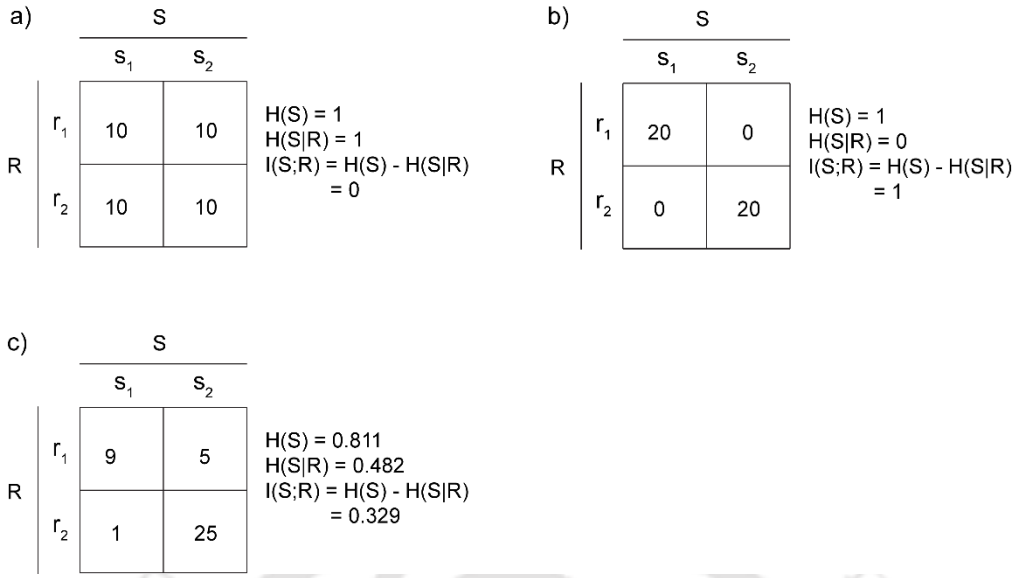



**Figure 2.8: Contingency table to elucidate mutual information**
*S* and *R* are signal and response, respectively. Assume that 40 cells were treated with two doses of the signal $s_1$and $s_2$ and the corresponding responses were $r_1$ and $r_2$. The mutual information is calculated by the contingency table method, and the calculations are shown beside each table. a) No correlation between *S* and *R*. Therefore, MI is zero. b) A one-to-one correspondence between *S* and *R* and therefore, MI is maximum. c) The system is noisy, but still retains some statistical correlation between *S* and *R*.

MI depends upon the distribution of the input. For a given system, MI varies if the probability distribution of the input changes. Therefore, one can manipulate it to maximize MI. Further, as MI depends upon the input distribution, we cannot compare two systems using MI only. In a biological experiment, where the dose or concentration of a ligand is used as an input signal, knowing the real distribution of the input signal is near to impossible. The channel capacity of a communication system circumvents these issues. The channel capacity of the system is defined as (*110*),

$$C = \sup_{p(x)} I(X;Y) \tag{2.8}$$

where the supremum is taken over all possible choices of $p(x)$. Estimation of $C$ can be computationally intensive, particularly for higher-dimensional data. In experimental

biology, often, a limited number of common continuous PDFs are used to estimate the $C$ from experimental data (*111*, *112*).

Estimation of MI and channel capacity of a pathway requires measurements of dose-response behavior in individual cells in an ensemble. Usually, cells are treated with different doses of a signaling molecule, and then the cellular response is measured by imaging or flow cytometry (*111-114*). When both input ($S$) and output ($R$) are discrete one can calculate MI using the data in a contingency table as shown in Figure 2.8 (*111*). However, one can argue that the concentration of a molecule is a continuous variable. This makes the estimation of MI non-trivial. Usually, binning-based methods are used to convert continuous variables to discrete, and then MI is estimated (*115*, *116*). Such binning-based methods have limitations and are suitable when a large number of states of the variable is measured in the experiment (*111*).

In a classic paper, Cheong et al. (*112*) used an information-theoretic approach to study TNF signaling. They formulated the computational aspects for reliable estimation of MI and C from experimental data. They had shown that TNF signaling in individual cells carries enough messages to make binary decisions. They further showed that network architecture could create a bottleneck in information flow, and a negative feedback can reduce such bottleneck. Similarly, Voliotis et al. (*117*) had also shown that negative feedback increases information transfer in a protein kinase pathway.

Suderman et al. (*111*) studied TRAIL-induced apoptosis using information theory. They measured information transfer at two levels. In one, they estimated MI and channel capacity considering cleaved caspase-3 or cleaved PARP in individual cells as the outputs. On the same experimental system, they also calculated MI and C, considering the percentage of apoptotic cells as an output. This measure is a population-level measure. Interestingly, they observed that the signal transmission at

the population-level had high channel capacity when compared to very noisy signaling at the single-cell level.

Information theory had been used in several other signaling systems in mammalian cells (*118*), Dictyostelium (*119*), Yeast (*120*), and bacteria (*121, 122*). Dubuis et al. (*123*) used information theory to explain how cells in a developing embryo precisely decide their position based on the expression level of a handful of genes.

Mutual information, as defined earlier, measures the ability of a communication system to transmit information through a noisy channel reliably. Therefore, MI = 1 bit indicates that the receiver will be able to discriminate only two ($2^1$) distinct input signal. Similarly, channel capacity sets the upper bound of how accurate the data can be transmitted through a channel. Interestingly, in many biological experiments, estimated MI and C were found to be low, often lower than 1 bit (*111, 112, 118, 124*). There is no doubt that molecular cell signaling is noisy, but such low MI or C is unexpected as a cell needs to differentiate different levels of an input signal and act.

Several authors have cautioned on interpreting the MI and C estimated from biological experiments (*111, 125, 126*). Estimation of MI is always data-intensive and may cause underestimation with limited data (*115, 125*). Levchenko and Nemenman have argued that the lack of precision in measurements in conventional cell biology experiments could explain the lower estimates of MI and channel capacity (*125*). Suderman and Deeds (*127*) have shown that the MI estimate is affected by the range of input signal sampled in the experiment and the density of sampling. They proposed a thumb rule to optimize the sampling in case of sigmoidal dose-response.

Beyond its physical meaning in a communication system, MI is also a measure of statistical dependencies between two variables (*125, 128, 129*). Linfoot (*130*) has defined an information coefficient of correlation $r = \sqrt{(1 - e^{-2MI})}$ and has shown that

$r$ is equivalent to the classical correlation coefficient when $p(x, y)$ is normal. MI is now widely used to quantify the correlation between different variables in large data sets, including time-series data, and proven very useful for nonlinear systems (*131*). In our work, we have used MI primarily as a measure of statistical correlation between the input signal and the response.

## 2.9. Deconvolution of cell-type-specific gene expression from ensemble data

So far, we discussed the phenotypic state transition and the various approaches used to study the dynamics of cell state transition. Let us consider, we have constructed the different lineages of cells during cancer metastasis and found that a cell type is more migratory than others, and it is the critical player in cancer metastasis. The next step would be to intervene in the system and target that cell type. To target a cell type, we need to know the molecular level information like gene expression of that cell.

The first step is to isolate the different cell types present in the mixture of cells, followed by gene expression analysis. Through techniques like flow cytometry, the different cell types can be sorted based on the cell-type-specific markers. This approach can be used if the information about the cell-type-specific markers is available. However, if the phenotypic states were defined in terms of functional features like change in morphology of cells, differential migratory potential of cells, then the cells cannot be separated. Moreover, if we do not have any prior knowledge of the molecular markers of the cells, then we cannot isolate the different cell types.

### 2.9.1. Ensemble behavior obscures subpopulation behavior

Quantitative PCR (qPCR) is the simplest and the most economical method to study gene expression quantitatively. In qPCR, we measure gene expression from an ensemble of cells. The cell population might contain different cell types or subpopulation of cells. However, as we measure gene expressions from an ensemble

of cells, the gene expression of different cell types present in the population is obscured (Figure 2.9). For example, Tumor-infiltrating lymphocytes (*132*) and endothelial cells (*133, 134*) are often seen in tumor tissues. Therefore, when tumor tissues are subject to gene expression analysis, there will be a high chance of contamination of non-tumor cells. Thus, leading to a biased gene expression measurement of the target tissue. These issues paved the way for the development of several computational deconvolution methods to nullify the gene expression of undesired cell types from the ensemble measurement.

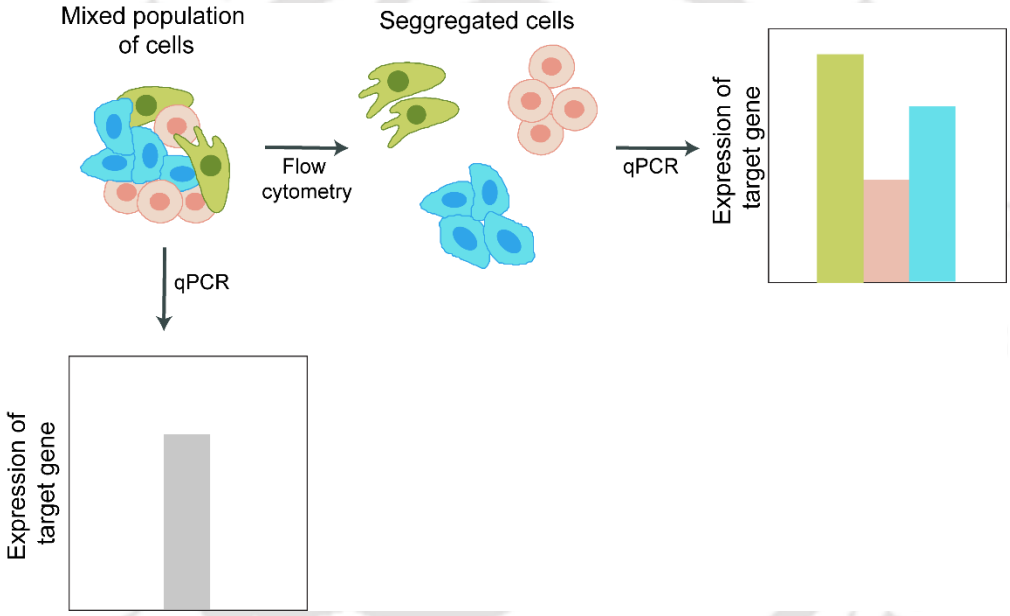



**Figure 2.9: Cell-type-specific gene expression is obscured in ensemble measurement.**

## 2.10. Existing deconvolution methods

Any deconvolution method requires the following information:

1. Population-level measurement of the target gene expression.

2. The proportion of various cell types in the population or the cell-type-specific expression of the target gene.

The gene expression of an ensemble of cells can be expressed as a linear combination of the gene expression of the individual cell types or subpopulation present in the population. All the existing deconvolution methods are based on this relation. If there are $n$ different cell types in a population of cells ($N_T$), then the total number of target mRNA is,

$$G = \sum_{i=1}^{n} g_i \times N_i$$

where $g_i$ is the number of target mRNAs in the $i^{th}$ cell type and $N_i$ is the number of cells of $i^{th}$ type. The above relation can be expressed in terms of fractions of cells as,

$$\hat{G} = \sum_{i=1}^{n} g_i \times f_i$$

where $\hat{G} = {}^{G}/_{N_T}$ ; $f_i$ is the fraction of the $i^{th}$ cell type.

The vector notation of the above equation is,

$$\hat{G} = \boldsymbol{g} \cdot \boldsymbol{f} \tag{2.9}$$

The existing methods deconvolute either $\boldsymbol{g}$ given $\boldsymbol{f}, \hat{G}$ (*21*) or $\boldsymbol{f}$ given $\boldsymbol{g}, \hat{G}$ (*15, 17, 19, 20, 135, 136*) or both $\boldsymbol{g}, \boldsymbol{f}$ given $\hat{G}$ (*16, 18, 22, 137, 138*). Abbas et al. (*13*) have deconvoluted the microarray of blood tissue to estimate the proportion of various constituents of the blood. Shenn-Orr et al. (*21*) have measured the cell-type-specific gene expression of various cell types from human renal transplant microarray. Dimitrakopoulou et al. (*14*) have used unsupervised machine learning techniques to estimate both the proportion of various cell types as well as the gene expression of each cell type from population-level gene expression data.

## 2.11. Limitations of the existing deconvolution methods

The deconvolution methods were widely used in microarray data analysis. Microarray data are subject to a series of data processing like background subtraction, log transformation, normalization (*139*). Depending on the type of processed data used in the deconvolution, the physical meaning of $\boldsymbol{g}$ in the equation (2.9) varies. Existing methods do not provide the physical interpretation of the deconvoluted parameter. For example, Lu et al. (*136*) and Shen-Orr et al. (*21*) have used log-transformed microarray data in their method. Log transformed data cannot be used in the linear equation. Zhong et al. (*140*) have shown that log transformation results in an underestimation of deconvoluted parameters. Moreover, existing methods are customized for high throughput experiments like microarray and RNA-seq.

The parameters $\boldsymbol{g}$ and $\boldsymbol{f}$ in equation (2.9) should always be positive. In some of the deconvolution methods, the non-negativity constraint is not maintained. Shen-Orr et al. (*21*) have used zero in place of negative gene expression coefficients, which is mathematically incorrect. Abbas et al. (*13*) have iterated the optimization until all the estimated parameters were non-zero. This approach will not yield the optimal result since the non-negativity constraints were not imposed during the parameter estimation.

Gene expressions are usually studied across various experimental conditions like cells treated for different durations, cells treated with different doses of a drug. Gene expression is a dynamic process, and it changes depending on the experimental conditions. Existing deconvolution methods assume that the gene expression remains constant across experimental conditions or time. Therefore, these methods cannot be used to deconvolute the experimental condition-dependent or time-dependent gene expression profile of each cell type present in the population.

The available deconvolution methods estimate the unknown parameter through optimization techniques like linear least square method (*136*), quadratic programming (*17, 22*), expectation-maximization (*138*). All these mathematical optimizations are based on the frequentist-approach of parameter estimation. The frequentist approach reports the point estimate of the parameter of interest and does not give any information on the probability distribution of the estimated parameters.

## 2.12. Bayesian method of parameter estimation

The other alternative method is the Bayesian approach of parameter estimation or the Bayesian inference. In this approach, the probability densities of the unknown parameters are estimated from the observed data. Bayesian inference involves three steps: 1) formulation of a mathematical model that describes the observed data, 2) defining the prior probability distribution of the parameter to be estimated, and 3) computation of the posterior probability distribution of the parameter.

Let us consider a simple example of plotting a standard curve, which is widely performed in protein estimation assays. In this experiment, the absorbance of different known concentrations of a standard protein is measured. As it is known that the absorbance of the solution increases linearly with the increasing concentrations, we can represent the experimental observation in a simple linear equation, $\boldsymbol{Y} = m \times \boldsymbol{X}$. Here $\boldsymbol{X} = (x_1, x_2, \dots, x_n)$ are the different concentrations of the protein and $\boldsymbol{Y} = (y_1, y_2, \dots, y_n)$ are the respective absorbance values. Now, we want to estimate the unknown parameter $m$. When we do not have any prior idea about the distribution of $m$, the simplest way is to assume a uniform distribution, $P(m) \sim U(lower\ bound, upper\ bound)$. As per Bayes theorem, the posterior distribution of $m$ is,

$$P(m|\boldsymbol{Y}) = \frac{P(\boldsymbol{Y}|m) \times P(m)}{P(\boldsymbol{Y})}$$

here $P(\boldsymbol{Y}|m)$ is the data likelihood; $P(m)$ is the prior distribution of $m$; $P(\boldsymbol{Y})$ is the marginal probability of $\boldsymbol{Y}$, which is a normalizing constant.

Generally, normal likelihood is used to estimate $P(\boldsymbol{Y}|m)$, considering the data is normally distributed around the true value with $\mu_i = m \times x_i$ and a known variance ($\sigma^2$).

$$P(\boldsymbol{Y}\,|\,m) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\times\pi\times\sigma^2}} \times e^{\frac{-\left(y_i - \left(\sum_{i=1}^{n} m\times x_i\right)\right)^2}{2\times\sigma^2}}$$

The direct sampling of $P(m|\boldsymbol{Y})$ is difficult. Usually, a Markov chain is generated through a series of random walks performed using sampling methods like Gibbs sampling, Metropolis-Hastings. From this Markov chain, $P(m|\boldsymbol{Y})$ is computed. The random walk through Metropolis-Hasting is executed as follows.

1. An initial value of $m$ is picked from $P(m)$.

2. A new step is proposed, $m' \sim P(m)$. The ratio of the new posterior density and the old posterior density is calculated. This ratio is called the acceptance probability, $P(\alpha) = \frac{P(Y|m')\times P(m')}{P(Y|m)\times P(m)}$.

3. If $P(\alpha) > 1$, then the $m'$ is accepted. If $P(\alpha) < 1$, then a random number is generated, $b = U(0,1)$. If $b < P(\alpha)$, then $m'$ is accepted or else $m'$ is rejected.

4. Steps 2-3 are repeated a sufficiently higher number of times so that the random walk reaches the stationary distribution.

The stationary distribution of the random walk is the posterior distribution of $m$. From $P(m|\boldsymbol{Y})$, the credible intervals of the parameter can be computed. Thus, Bayesian inference gives us a measure of the uncertainty in the estimated parameter.

## 2.13. Objectives

Considering the existing limitations in the study of cell state transition, we proposed the following objectives for this thesis work.

1. To study the signaling and dynamics of cells state transition in an in vitro experimental system of EGF-induced EMT.

2. To study the effect of background noise on EGF-induced EMT in an in vitro experimental system.

3. To develop a mathematical tool for the deconvolution of quantitative PCR data of a heterogeneous sample for estimation of cell-type-specific gene expression.

CHAPTER 3

# Materials and Methods

The detailed information on various experimental methods used in this thesis is given in this chapter.

## 3.1. Cell culture

### 3.1.1. Cell lines and culture conditions

Human breast cancer cell lines MDA-MB-468, MCF-7, MDA-MB-231 were procured from National Centre for Cell Sciences, Pune, India, and cultured in complete growth medium at 37 ˚C in a humidified incubator with 5 % $CO_2$. When cells were 80-90 % confluence, they were either subcultured or seeded for experiments. Stocks of cells were made for long term storage. Cells were preserved in 1 mL of cryopreservative (fetal bovine serum (FBS) with 5 % DMSO) and stored in the vapor phase of liquid nitrogen. The details about the cell culture media are given in Section B-1 in the Appendix.

### 3.1.2. Treatment conditions

In experiments involving the treatment of cells with growth factors or any drug, cells were seeded in the appropriate plates and maintained in complete growth medium for 36 h. Cells were washed with phosphate-buffered saline (PBS) and were kept in reduced serum media for 12 h, followed by treatment in reduced serum media (0.5 % FBS). The cell seeding density and the treatment media volume were scaled up such that the ratio of drug molecules to cells is maintained the same for a given concentration of the treatment molecule (Table 3.1). Whenever cells were treated with drugs dissolved in dimethyl sulfoxide (DMSO), the final concentration of DMSO in the treatment media is always kept less than 0.5 %.

**Table 3.1. Treatment conditions followed in experiments.**

For a given concentration of the drug, the treatment volume of media was scaled up according to the cell seeding density to maintain the same number of drug molecule availability per cell.

| *Cell culture plate/dish* | *Approximate seeding density of cells* | *Treatment volume* |
|---|---|---|
| 96 well plate | $8\times10^3$ cells per well | 100 µL |
| Transwell inserts for 24 well plates | $12\times10^3$ cells per insert | 150 µL |
| 12 well glass chamber slide | $14\times10^3$ cells per well | 175 µL |
| 24 well plate | $48\times10^3$ cells per well | 600 µL |
| 6 well plate/35 mm dish | $2.4\times10^5$ cells per well | 3 mL |

### 3.1.3. Cell counting

Cells were counted using the dye exclusion method (*141*). Cells were trypsinized and resuspended in 1 mL of complete growth media. 10 µL of cell suspension was mixed with 10 µL of trypan blue (dilution factor = 2). 10 µL of the mixture was loaded onto

the hemocytometer and visualized under an inverted microscope (CKX41, Olympus). The dead cells with a compromised membrane will take up the dye and become blue while the live cells will stay clear. The number of live cells in all four corner squares were counted manually by visual examination of the cells. The live-cell density in the suspension was calculated using the following formula.

$$Live\ cell\ density = \frac{Number\ of\ live\ cells\ counted}{4} \times dilution\ factor \times 10^4 cells / mL$$

## 3.2. Phalloidin-FITC staining

We performed this experiment to examine the changes in the cytoskeleton of the cells post EGF treatment. Cells were grown in 96 well plates and were treated as per the experimental conditions. After the experiment, the spent media was removed, and the cells were fixed with 4 % paraformaldehyde for 10 minutes. Cells were washed with 100 µL of PBS followed by permeabilization of cell membrane using 0.1 % Triton X-100 in 100 µL of PBS for 10 minutes. Cells were washed with 100 µL of PBS, and the cells were stained with 100 µL of 0.1 µM FITC Phalloidin conjugate in PBS for 1 h. Cells were washed with 100 µL of PBS to remove the unbound phalloidin conjugate and were counterstained with 30 µM DAPI in PBS for 5 minutes. Cells were washed twice with PBS and were imaged in 96 well plates with PBS using an Epi-fluorescence microscope (Nikon Eclipse Ti-U). All the steps were carried out at room temperature. The reagents were prepared as per the details given in Section B-2 in the Appendix.

## 3.3. Immunofluorescence

Cells were grown in 12 well glass chamber slides (Ibidi-81201) and were treated as per the experimental conditions. After the experiment, cells were fixed with ice-cold methanol-acetone mixture (1:1) for 10 minutes at -20 °C. Cells were washed with 200 µL of PBS. Cells were incubated with 200 µL of blocking buffer (1 % BSA, 0.3 M glycine in PBS containing 0.1 % Tween 20) for 30 minutes at room temperature, followed by PBS wash. Cells were incubated with fluorophore-conjugated primary

antibody diluted in 100 µL of antibody dilution buffer (1 % BSA in PBS containing 0.1 % Tween-20) overnight at 4 °C. After incubation, cells were washed twice with PBS. The silicone gasket on the glass chamber was removed, and mounting media was added on the surface of the glass chamber. A coverslip was placed over the mounting medium, and the glass slide was kept at room temperature for 15 minutes. The cells were imaged using a confocal microscope (Zeiss LSM 880). The details about the antibodies are given in Table B-2 in the Appendix.

## 3.4. Isolation and quantification of RNA

### 3.4.1. RNA isolation

Cells were grown in 35 mm cell culture dishes, and the appropriate experiment was performed. After the experiment, the cells were washed with PBS, and 500 µL of TRI reagent (SIGMA) was added to each dish. The cell lysate was homogenized using a 1 mL insulin syringe. To the lysate, 100 µL of chloroform was added and mixed by vigorous shaking. The mixture was kept at room temperature for 10 minutes, followed by centrifugation at 12,000 rpm at 4 °C for 15 minutes. Centrifugation resulted in clear phase separation. The upper aqueous phase containing RNA was transferred to a fresh tube without disturbing the DNA layer at the interface. An equal volume of isopropanol was added to the aqueous phase and mixed by inversion. The mixture was stored at room temperature for 10 minutes, followed by centrifugation at 12,000 rpm at 4 °C for 10 minutes. The supernatant was discarded, and the pellet was washed with 75 % ethanol. Ethanol was removed by centrifugation at 12,000 rpm at 4 °C for 10 minutes. The RNA pellet was air-dried and dissolved in 40 µL of prewarmed double distilled water. The centrifuge tubes, pipette tips, and double-distilled water were treated with 0.1 % diethylpyrocarbonate (DEPC) at 37 °C overnight and autoclaved before RNA isolation.

### 3.4.2. Removal of genomic DNA contamination from RNA

There may be some amount of genomic DNA in the isolated RNA. DNase treatment was performed to remove genomic DNA contamination. 50 µL reaction was set up (Table 3.2), and the reaction mixture was incubated at 37 °C for 45 minutes. The salts in the reaction buffer will hinder the downstream reactions. Therefore, after DNA digestion, the complete RNA isolation steps (section 3.4.1) were repeated to purify the RNA. The final RNA pellet was dissolved in 30 µL of prewarmed DEPC-treated double distilled water.

**Table 3.2. DNase digestion reaction setup.**

The genomic DNA contamination in RNA was removed by DNase digestion. RQ1 RNase-Free DNase was purchased from Promega (Cat. # M6101), and the reaction was performed as per the manufacturer's protocol.

| *Components* | *Volume for one reaction* |
|---|---|
| 10x RNase-free DNase buffer | 5 µL |
| RNase-free DNase (1000 U/mL) | 2 µL |
| Isolated RNA | 30 µL |

The total volume was made to 50 µL using nuclease-free water.

### 3.4.3. Quantification of RNA

RNA was quantified using a UV-Visible spectrophotometer (DU730, Beckman-Coulter). 5 µL of RNA sample was diluted in double distilled water to make up the final volume to 1 mL (dilution factor = 200). The absorbance was measured at 260 nm. 1 unit of absorbance is equivalent to 40 µg/mL of RNA. The concentration of RNA was calculated from the following formula.

$$Concentration\ of\ RNA = Absorbance_{260} \times dilution\ factor \times 40 \mu g / mL$$

### 3.4.4. Quality check of RNA

Quantified RNA was analyzed in a 1.5 % agarose-bleach gel as per the method of Aranda et al. (*142*). The gel was placed in the electrophoresis unit filled with 1x TAE. An equal amount of RNA was loaded into each well along with agarose gel loading dye. The gel was run at 80 V until the dye moved 75 % down the gel. The gel was visualized under UV-Transilluminator (MacroVue UVis-20, Hoefer), and the image was taken using the Gel Documentation System (ChemiDoc XRS+, BioRad). Two crisp bands corresponding to 28S and 18S rRNA confirms a good quality of RNA. The information about the buffers and the gel preparation is given in Section B-3 in the Appendix.

## 3.5. Synthesis and quality check of cDNA

### 3.5.1. Synthesis of cDNA

cDNA was synthesized using the Verso cDNA synthesis kit (Thermo Fisher Scientific). 0.8 µg of total RNA was used as the template in a 20 µL cDNA reaction. The RNA was diluted in nuclease-free water to make a final volume of 10 µL. The RNA sample was heated at 70 °C for 5 minutes and snap frozen at 4 °C. A cDNA master mix was prepared (Table 3.3), and 10 µL of the master mix was added to the 10 µL of the RNA sample. One-step reverse transcription reaction was set up for 60 minutes at 42 °C.

### 3.5.2. Quality check of cDNA

The quality of cDNA was analyzed by semi-quantitative PCR. The expression of housekeeping genes was used as a measure of cDNA quality. Cyclophilin A was used as the housekeeping gene in all the experiments. The PCR reaction was prepared, as given in Table 3.4. and the reaction conditions are given in Table 3.5. An equal amount of cDNA was used in all tubes. After the PCR, the samples were analyzed in 1.5 %

agarose gel. Equal band intensity across all samples assures excellent quality of cDNA.

**Table 3.3. Components of cDNA master mix.**

The cDNA reaction master mix was prepared as given in the table. The cDNA synthesis kit was purchased from Thermo Fisher Scientific (Cat. # AB1453A), and the reaction was performed as per the manufacturer's protocol.

| *Components* | *Volume for one reaction* |
|---|---|
| 5x cDNA synthesis buffer | 4 µL |
| dNTP (5 mM each) | 2 µL |
| Random hexamer (400 ng/µL) | 1 µL |
| Verso enzyme | 1 µL |
| Nuclease free water | 2 µL |

**Table 3.4. PCR reaction preparation.**

The ready-to-use Taq enzyme mix containing Taq DNA polymerase, dNTPs, $MgCl_2$ and other buffers at optimal concentrations was purchased from Himedia (Cat. # MBT061). The sequences of primers used are given in Table B-4 in the Appendix.

| *Components* | *Volume for one reaction* |
|---|---|
| PCR TaqMixture (2x) | 10 µL |
| Forward primer (2 µM) | 2 µL |
| Reverse primer (2 µM) | 2 µL |
| cDNA template | Variable volume (20 ng RNA equivalent) |
| The total volume was made to 20 µL using nuclease-free water. | |

**Table 3.5. PCR reaction conditions.**

| *Reaction* | *Temperature* | *Time* | *Number of cycles* |
|---|---|---|---|
| Initial denaturation | 95 °C | 1.30 min | 1 |
| Denaturation | 95 °C | 30 sec | 21-25 (depending on the amount of template) |
| Annealing | 60 °C | 30 sec | |
| Extension | 72 °C | 30 sec | |
| Fina extension | 72 °C | 10 min | 1 |

## 3.6. Quantitative PCR

The expression of target genes was quantitatively studied using real-time PCR. 15 µL reaction was prepared for each sample (Table 3.6), and the experiment was done in triplicates. Thin-walled 0.1 mL strip tubes (Axygen – PCR-0104-C) were used to set up the reactions. cDNA equivalent to 20 ng of RNA was used as a template per reaction. Cyclophilin A was used as a reference gene in all experiments. Target genes were amplified using the QuantiFast SYBR Green PCR kit (QIAGEN), and the reactions were performed on Roto-Gene-Q (QIAGEN). The reaction conditions are given in Table 3.7. Melt curve analysis was performed at the end of the reaction to check for any non-specific amplification. $C_t$ values of each sample and the efficiency of the reaction were calculated using LinRegPCR (*143*). Fold change in target gene expression was estimated by the $\Delta\Delta C_t$ method (*144*, *145*).

$$Fold\ change = \frac{\eta_x^{ct_{x,c}-ct_{x,s}}}{\eta_r^{ct_{r,c}-ct_{r,s}}}$$

here, $\eta_x \rightarrow$ efficiency of amplification of the target gene.

$\eta_r \rightarrow$ efficiency of amplification of the reference gene.

$ct_{x,c} \rightarrow$ threshold cycle of the target gene in the control sample.

$ct_{x,s} \rightarrow$ threshold cycle of the target gene in the test sample.

$ct_{r,c} \rightarrow$ threshold cycle of the reference gene in the control sample.

$ct_{r,s} \rightarrow$ threshold cycle of the reference gene in the test sample.

**Table 3.6. Components of the quantitative PCR reaction.**

The SYBR green PCR master mix was purchased from QIAGEN (Cat. # 204054), and the reaction was set up as per the manufacturer's protocol. The sequences of the primers used are given in Table B-4 in the Appendix.

| *Components* | *Volume for one reaction* |
|---|---|
| SYBR green master mix (2x) | 7.5 µL |
| Forward primer (2 µM) | 1.5 µL |
| Reverse primer (2 µM) | 1.5 µL |
| cDNA | Variable volume (equivalent to 20 ng of RNA) |
| The total volume was made to 15 µL using nuclease-free water. | |

**Table 3.7. Quantitative PCR conditions.**

The data was acquired in the SYBR green channel during annealing and extension steps.

| *Reaction* | *Temperature* | *Time* | *Number of cycles* |
|---|---|---|---|
| Initial denaturation | 95 °C | 5 min | 1 |
| Denaturation | 95 °C | 30 sec | 40 |
| Annealing and extension | 60 °C | 45 sec | |

## 3.7. Quantitative image analysis

MDA-MB-468 cells exist in three different morphologies: Cobble, Spindle, and Circular. Depending on the dose and the duration of EGF, the distribution of the three morphological states varied. We used image analysis to quantitatively estimate the

distribution of the three morphological states of MDA-MB-468 cells. We used 96 well plates for all image analysis-based experiments. Here, each well corresponds to a specific experimental condition, like different doses of EGF. To classify the cells based on the shape, we initially trained the algorithm with images containing all three morphologies of MDA-MB-468 cells. For each experiment, we seeded cells in additional wells and treated them with EGF such that there is a mixture of all three cell types. We imaged these cells in the extra wells and used them for training the algorithm. These images were called the training data set, and the experimental samples were called test data set.

### 3.7.1. Sample preparation

The cells were seeded in 96 well plates, and the experiment was carried out. After the experiment, the spent media was removed, and the cells were fixed with 4 % paraformaldehyde for 10 min. Cells were washed with 100 μL of PBS followed by permeabilization of cell membrane using 0.1 % Triton X-100 in 100 μL of PBS for 10 min. The cells were washed with 100 μL of PBS and were stained with 100 μL of 0.001 μg/mL of HCS CellMask Red Stain (Invitrogen) for 30 min. The cells were washed twice in 100 μL of PBS to remove the excess dye. All the experimental conditions were performed at room temperature. The cells were imaged using an Epi-fluorescence microscope (Nikon Eclipse Ti-U) along with the 96 well plates. For each experimental condition, ten non-overlapping fields of view of a single well were taken such that the images represent the population distribution of cells. The images were taken with 100 μL of PBS in each well to prevent the shrinking of cells. The reagents were prepared as per the details given in Section B-2 in Appendix.

### 3.7.2. Training the classifier

The images were analyzed using CellProfiler (stable 2.2.0) (*61*) and CellProfiler Analyst (stable 2.2.1) (*146*). The measurements of the cells like shape features, the granularity of cells, neighbors of the cells were extracted for each cell in the training

data set using CellProfiler. The features were extracted based on the series of instructions given by the user, called pipeline. The data were exported to an SQLite database file as well as a tab-delimited text file. The extracted data were loaded into the CellProfiler Analyst. Here, the cells were manually classified by the user based on their shapes (Cobble, Spindle, or Circular). We trained the classifier with 100 cells in each cell type using the Fast Gentle Boosting algorithm. A set of rules were generated during the training, with which the test data set will be classified. The efficiency of the training was evaluated based on the confusion matrix plot (Figure A-5 in Appendix). The number of cells used for training was chosen based on an optimization experiment (Figure A-4 in Appendix).

### 3.7.3. Classification of the cell types

The test data set were used in the CellProfiler, and the features of each cell were extracted using the same pipeline used in the training data set. The test data set were processed in batches. For example, each experimental condition will have at least ten images (non-overlapping fields of view), and these images were processed together in batch-mode. The entire process was automated using a bash script. The rules generated during the training were used in CellProfiler, and the cells were classified based on the rules. The output was saved to a tab-delimited text file with each cell categorized as either Cobble, Spindle, or Circular. From this data, the percentage of each cell type was calculated.

## 3.8. Migration assay

We categorized MDA-MB-468 cells based on their morphology. Further, to examine how different the cell types were in terms of their functional features, we performed the migration assay. We used transwell insert (Polycarbonate cell culture inserts with 8-micron pore size, Cat. # 140629, Thermo Fisher Scientific) to assess the migratory potential of cells. The cells that can migrate through the transwell insert were considered migratory.

### 3.8.1. Sample preparation

The cells were grown in 24 well plates containing transwell inserts. The cells were seeded in both the insert as well as in the 24 well plates. The cells were treated with EGF as per the experimental conditions. After 24 h of EGF treatment, the media in the insert was discarded, and the insert was transferred to a fresh well containing reduced serum media. 150 µL of reduced serum media (0.5 % FBS) was added to the insert gently, and the setup was incubated for 6 h. After 6 h, the insert was removed, and the spent media was discarded. The cells on the inner membrane of the insert were removed with a cotton swab. The migrated cells to the other side of the membrane were fixed with ice-cold methanol. The membrane was washed with PBS, followed by staining with HCS CellMask Red stain for 30 min. The excess dye was washed with PBS, and the membrane was cut off from the insert and mounted on a glass slide.

### 3.8.2. Imaging of cells

The mounted membrane was imaged using a confocal microscope (Zeiss LSM 880). To image the entire membrane area and to bring all the cells in focus, tile scanning ($2 \times 2\ tiles$) was performed along with Z stacking. All the Z stacked images were projected to a single 2D image using maximum intensity projection. From these images, the morphology of the migrated cells was analyzed visually.

## 3.9. Western blotting

### 3.9.1. Sample preparation

The cells were seeded in 35 mm dishes and were treated with the drug as per the experimental conditions. The spent media was discarded, and the cells were washed with PBS. The cells were lysed with 120 µL of RIPA buffer containing 1 mM PMSF, 1 mM sodium orthovanadate, 50 mM sodium fluoride, and 1 mM EDTA. The cells were incubated on ice for 5 min, and the lysate was pooled down using a cell scraper. The lysate was sonicated at 25 % amplitude and 0.5 s pulse cycle for 10 s, followed by

centrifugation at 12,000 rpm for 10 min at 4 °C. The supernatant was collected and stored at -80 °C in multiple aliquots. The composition of the RIPA buffer, protease, and phosphatase inhibitors are given in Section B-4 in the Appendix.

### 3.9.2. Total protein estimation

The total protein was estimated by Lowry's method (*147*). 10 µL of the protein sample was diluted to 100 µL with distilled water. The diluted sample was mixed with 250 µL of Lowry's reagent (2 % $Na_2CO_3$ in 0.1 N $NaOH$ : 1 % $CuSO_4.5H_2O$ : 2 % potassium tartrate = 100 : 1 : 1). The reaction mixture was incubated for 10 min at room temperature. 25 µL of Folin-Ciocalteu reagent was added to the mixture and incubated in the dark for 30 min at room temperature. 200 µL of the reaction mixture was transferred to a 96 well plate, and the absorbance was measured at 660 nm. Different concentrations of BSA diluted in RIPA buffer was used as a standard. A standard curve was prepared by fitting the absorbance of the different concentrations of BSA to the linear regression equation, $Y = m \times X$. Here, $Y$ is the absorbance; $X$ is the concentration of the solution and $m$ is the slope of the line. The concentrations of the unknown protein sample were estimated by interpolating the absorbance on the standard curve.

### 3.9.3. SDS-PAGE and transfer of resolved proteins to the membrane

The quantified cell lysate samples were resolved by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). The polyacrylamide gel was prepared based on the protocol of Maniatis et al. (*148*). Depending on the size of the target protein to be detected, the appropriate gel was cast. The composition of the different percentages of gels is given in Table B-5 and Table B-6 in the Appendix. An equal amount of protein (10-30 µg) from all samples were diluted in 4x gel loading dye, and the mixture was heated at 95 °C for 5 min. The samples were loaded, and the gel was allowed to electrophorese until the dye front reaches the bottom of the separating gel. In the meantime, a PVDF membrane (0.2 µm, Millipore) and two sets

of filter paper (3 mm Whatman chromatography paper), larger than the dimensions of the gel were cut. The PVDF membrane was activated with methanol for 5 min. The activated PVDF membrane, two sets of filter paper, and the gel were equilibrated with 5x Towbin buffer for 10 min.

The resolved protein samples from the gel were transferred to the PVDF membrane through wet transfer. The transfer stack assembly was prepared as follows. A sandwich was prepared in the following order and placed in the transfer cassette: A foam (fiber), one set of filter paper, polyacrylamide gel, PVDF membrane, one set of filter paper, and a foam. The air bubbles, if formed, were removed using a blotting roller to ensure even transfer. The cassette was placed in the tank with the gel side of the sandwich facing the cathode and the membrane side facing the anode. 5x Towbin buffer was added into the tank, and the transfer was allowed for 3 h at 25 V. The details about the buffers and reagents used are given in Section B-4 in Appendix.

### 3.9.4. Detection of the target protein

After wet transfer, the membrane was removed from the transfer cassette and stained with Ponceau S (SIGMA) to confirm the transfer. From this step, freshly prepared protease and phosphatase inhibitors were used in all the solutions. The membrane was washed with TBS for 5 min, and the membrane was blocked with 3 % BSA in TBST for 2 h at room temperature. After blocking, the membrane was probed with the appropriate primary antibody overnight at 4 °C with gentle rocking. The unbound primary antibodies were removed by washing the membrane with TBST (3 washes × 10 min.), and the membrane was probed with a suitable secondary antibody conjugated with HRP for 1 h at room temperature. The unbound secondary antibodies were removed by washing the membrane with TBST (4 washes × 10 min), and the blot was developed using chemiluminescence (SuperSignal West Dura kit, Thermo Fisher Scientific) and imaged using a gel documentation system (Chemi Doc XRS+, BioRad). The list of antibodies used is given in Table B-1 in the Appendix.

### 3.9.5. Quantification of the target protein

The developed blots were quantified by the densitometry method using ImageJ (*149*). The target proteins were normalized with loading control (beta-actin). In the case of phosphoproteins, the total protein was used as the loading control. Blots developed on different days were compared by considering the densitometric values of the control sample of all the blots were equal.

## 3.10. Sandwich-ELISA to measure EGF in the media

The availability of EGF in the extracellular media was measured through sandwich ELISA using the ready-to-use kit from SIGMA (Cat. # RAB0149). The kit comes with a precoated anti-EGF antibody in a 96 well plate. The detectable range of this kit is 1 pg/mL to 200 pg/mL. The cells were grown in 96 well plates and treated with different doses of EGF for varying time points. The media was collected at the experimental time points and stored at -80 °C. The samples were diluted with the sample dilution buffer such that the concentration of EGF will be within the detectable range of this kit. The dilutions were made based on the amount of EGF at time = 0. The experiment was performed as per the manufacturer's protocol. As a control, we had certain wells without cells, to measure the auto degradation of EGF. The percentage of EGF availability to cells was calculated with respect to the EGF availability at time = 0.

## 3.11. Cell death estimation by flow cytometry

The apoptotic cell population was estimated through the cell cycle analysis experiment. The cells in the sub G0/G1 phase of the cell cycle were considered as apoptotic population. We used a modified protocol of Riccardi and Nicoletti (*150*). The cells were seeded in 35 mm dishes, and the cells were treated as per the experimental conditions. The spent media was centrifuged to collect the dead cells, and the adherent cells were collected by trypsinization. Both the cell pellets (dead + adherent) were resuspended together in 500 μL of PBS. The cells were aliquoted into

5 tubes of 100 µL each, and the cells in each tube were fixed with 900 µL of 70 % ice-cold ethanol at constant vortexing. The fixed cells were passed through 1 mL insulin syringe to break down the cell clumps, and the cells were stored at 4 °C overnight. The cells were centrifuged and were pooled down to a single tube with 1 mL of PBS. To the suspension, 1 mL of DNA extraction buffer was added and incubated for 5 min at room temperature. After incubation, the cells were centrifuged, and the cell pellet was resuspended in 2 mL of DNA staining solution. The cells were incubated with the DNA staining solution for at least 30 min at room temperature. The composition of the buffers used is available in Section B-5 in the Appendix.

The cells were analyzed in CytoFLEX (Beckman Coulter). Twenty-five thousand cells were recorded for each sample, and the data were acquired in the linear-mode in the FL-2 channel. The acquisition settings were adjusted with the unstained cells. Doublet discrimination was performed in the Height-Area plot. The data was analyzed in FCS Express 5 (De Novo Software) using the MultiCycle DNA program.

## 3.12. phospho-EGFR measurement through flow cytometry

The cells were grown in 35 mm dishes and were treated as per the experimental conditions. The cells were trypsinized and were resuspended in 200 µL of ice-cold PBS. The cells were fixed with 800 µL of 100 % ice-cold methanol at constant vortexing. The fixed cells were passed through 1 mL insulin syringe to break down the cell clumps. The cells were stored at -20 °C for at least 15 min. The cells were centrifuged at 3000 rpm for 20 min at 4 °C, and the cell pellet was resuspended in 100 µL of blocking solution (0.5% FBS in PBS) and incubated for 2 h at room temperature. The cells were centrifuged and incubated ($2 \times 10^5$ cells per tube) with the primary antibody in 100 µL of blocking buffer overnight at 4 °C. The excess antibody was removed, and the cells were washed with PBS followed by incubation with the AlexaFluor 488 conjugated secondary antibody in 100 µL of blocking solution. After 1 h incubation at room temperature, the unbound secondary antibody was removed

by centrifugation, and the cells were resuspended in 500 µL of blocking solution. The cells were analyzed in CytoFLEX (Beckman Coulter). The details about the antibodies used are given in Table B-3 in the Appendix.

The initial acquisition settings were adjusted with unstained cells. Twenty-five thousand cells were acquired for each experimental condition, and the data was collected in log-mode in the FL-2 channel. Data analysis was performed using FCS Express 5 (De Novo Software). The positive population was estimated by histogram subtraction. The cells that were stained only with the secondary antibody was used as a control in histogram subtraction.

## 3.13. Microplate assay to estimate live and dead cells

EGF is known to induce cell proliferation as well as cell death (*151-155*). Therefore, in the cell state transition model, we considered both cell birth and death. Using a fluorescence-based plate reader assay, we estimated the fold change in total cell number (live + dead) and dead cell number. We used the method developed by Dengler et al. (*156*) and Wan et al. (*157*). In this experiment, we used propidium iodide (PI), that binds to the double-stranded DNA. The amount of fluorescence from PI is proportional to the DNA content of cells, which in turn corresponds to the number of cells. The fluorescence was measured using a microplate reader at $\lambda_{ex} = 530\ nm$ and $\lambda_{em} = 620\ nm$. To compare the fluorescence reading from different time points (different 96 well plates), few additional wells containing cells were fixed with 100 % methanol in all plates, after 12 h of seeding the cells. The fluorescence reading from these wells was used to normalize fluorescence values across different time points.

### 3.13.1. Estimation of dead cells

Cells were grown in black-walled 96 well plates (Cat. # 137101, Thermo Fisher Scientific) and were treated as per the experimental conditions. PI at a final concentration of 1 µg/mL was added to each well without removing the media. The

plate was incubated at 37 °C for 10 min, followed by fluorescence measurement. The membrane of the dead cells will be compromised, and they will take up the PI. Fold change in dead cell number was estimated with respect to the time = 0 samples. A standard curve was plotted with various cell densities to check the linear regime of the assay (Figure A-1 in Appendix). The fluorescence increased linearly for cell numbers from 0 to 8000 ($R^2 = 0.985$). A positive control experiment was also performed to check the quality of the assay. We treated cells with different doses of etoposide (cytotoxic drug) and measured the percentage of cell death (Figure A-2 in Appendix). The assay showed a dose-dependent increase in cell death.

### 3.13.2. Estimation of the total number of cells

The cells were seeded in black-walled 96 well plates and treated as per the experimental conditions. A staining solution (final concentration: 30 μg/mL PI, 0.01 M EDTA, and 0.5 % Triton X-100) was added to each well without removing the media. Triton X-100 will perforate the cells, and the PI will bind to the DNA of both live as well as the dead cells. The plate was incubated at room temperature for 6 h, and the fluorescence was measured using a microplate reader. Fold change in total cell number was estimated with respect to the time = 0 samples. A standard curve was plotted with different density of cells to check the linearity of the assay (Figure A-3 in Appendix). The assay showed linearity to a wide range of 250 to 30000 cells ($R^2 = 0.977$). The percentage of live and dead cells was calculated from these data (Section A-1 in Appendix).

## 3.14. Cell viability assay

We performed the cell viability assay to check the cytotoxic effect of Gefitinib. MDA-MB-468 cells were seeded in 96 well plates, and the cells were treated with different doses of Gefitinib for different time points. Subsequently, the viability of the cells was measured by 3-(4,5-dimethylthiazol-2yl)-2,5-diphenyltetrazolium bromide (MTT) assay (*158*). DMSO was used as a solvent for Gefitinib. The percentage of cell viability

was calculated relative to cells treated with an equivalent amount of DMSO in media (without Gefitinib).

## 3.15. Data analysis

SigmaPlot, Microsoft Excel, MATLAB, and Python were used for data analysis, plotting graphs, and statistical analysis. Mean of multiple data points are plotted with error bars representing standard deviations. Wherever applicable, suitable statistical tests were performed, and they are represented in the corresponding figure legends.

CHAPTER 4

# The Signaling and Dynamics of EGF-induced Epithelial-Mesenchymal Transition

## 4.1. Introduction

The cell state transition or cellular plasticity is observed in various biological phenomena like embryogenesis, cancer metastasis, and the emergence of drug-resistant cells (*4, 54, 159*). Cell state transitions are driven by an external signal (*160, 161*) or by the noise (*6*) in the cellular system. Understanding the dynamics of cell state transition helps to decipher the time evolution of various cell states.

In this chapter, we explored the dynamics of cell state transition using EGF-induced Epithelial-Mesenchymal Transition (EMT) of MDA-MB-468 cell line as an

experimental system. MDA-MB-468 is a triple-negative breast cancer cell line of basal type A (*162*, *163*). EGF-induced EMT of MDA-MB-468 cells is a well-known system to study EMT (*48-50*, *164-166*).

We used both experimental as well as computational approaches to study the dynamics of cell state transition. We defined the phenotypic state of the cells based on the morphology of the cells. We developed a mathematical method to decipher the evolutionary trajectories of various cell states in EGF-induced EMT of MDA-MB-468 cells. The method is developed in MATLAB and is openly available in Github (https://github.com/biplabbose/StateTransition). We also explored the molecular-level signaling that drives the state transition in MDA-MB-468 cells.

## 4.2. EGF-induced EMT

We performed some preliminary experiments to standardize the experimental conditions for EMT. One of the prominent features of EMT is the cytoskeletal reorganization of cells (*167*). We treated MDA-MB-468 cells with different doses of EGF and stained the cells with the phalloidin-FITC conjugate. Phalloidin binds to F-actin in the cytoskeleton of the cell. The FITC aids in visualizing the change in the cytoskeleton of the cell. The cells were counterstained with a DNA binding dye, DAPI, to distinguish each cell. MDA-MB-468 cells grow as a monolayer and form tight clusters with neighboring cells. On EGF treatment, the cells lost contact between neighboring cells and were scattered. The cells transformed from an initial cobble-like appearance to elongated and circular morphology (Figure 4.1).

The prominent molecular level markers of EMT are the increased expressions of vimentin, fibronectin, SNAIL1, and ZEB1 (*37*). We checked the expressions of a few molecular markers of EMT through quantitative PCR as well as through Immunofluorescence (Figure 4.2). Our results were congruous with the existing

reports on changes in expression of these markers in EGF-induced EMT of MDA-MB-468 cells (*50, 164, 168*).

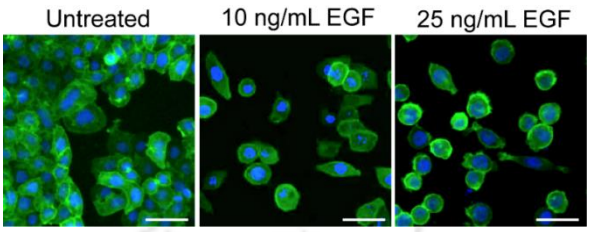



**Figure 4.1: EGF-induced cytoskeletal changes.**
Cells were stained with Phalloidin-FITC and DAPI, post 24 h of EGF treatment. Green color represents the cytoskeleton of the cells, and the blue color represents the nucleus of the cells. The figure shows the representative images of the cell population. Scale bar in images: 50 μm.

## 4.3. Phenotypic states of MDA-MB-468 cells

We defined the phenotypic state based on the morphology of MDA-MB-468 cells. To study the morphology of cells, we stained the cells with HCS CellMask Red Stain. This dye stains the cells completely, both cytoplasm and nucleus, thereby providing a clear distinction between the background and the cells. Through this staining, the intact morphology of the cells can be measured efficiently. We observed three distinct morphologies in MDA-MB-468 cells (Figure 4.3). We call these cells Cobble, Spindle, and Circular. We consider these three morphologies as the morphological or phenotypic states of MDA-MB-468 cells. Cobble cells are polygonal and form tight clusters with the neighboring cells. Spindle and Circular cells are scattered and are loosely adherent. All these cell types grew as a monolayer and were not seen floating over the media.

We treated MDA-MB-468 cells with varying doses of EGF and imaged the morphology pattern of MDA-MB-468 cells (Figure 4.4a). We employed image analysis tools to measure the change in the population distribution of cells in different morphological states of MDA-MB-468 cells quantitatively. We used CellProfiler and

CellProfiler Analyst to classify cells based on morphology. Detailed information about image analysis is given in section 3.7 in chapter 3.

Figure 4.4b shows the quantitative plot of the EGF-induced changes in the population distribution of MDA-MB-468 cells. When there is no EGF stimulation, most of the cells were of Cobble-type. As the dose of EGF increased, a higher number of Circular cells were observed. A considerable amount of Spindle cells was observed at intermediate doses of EGF. Therefore, EGF had a dose-dependent effect on the population distribution of MDA-MB-468 cells.

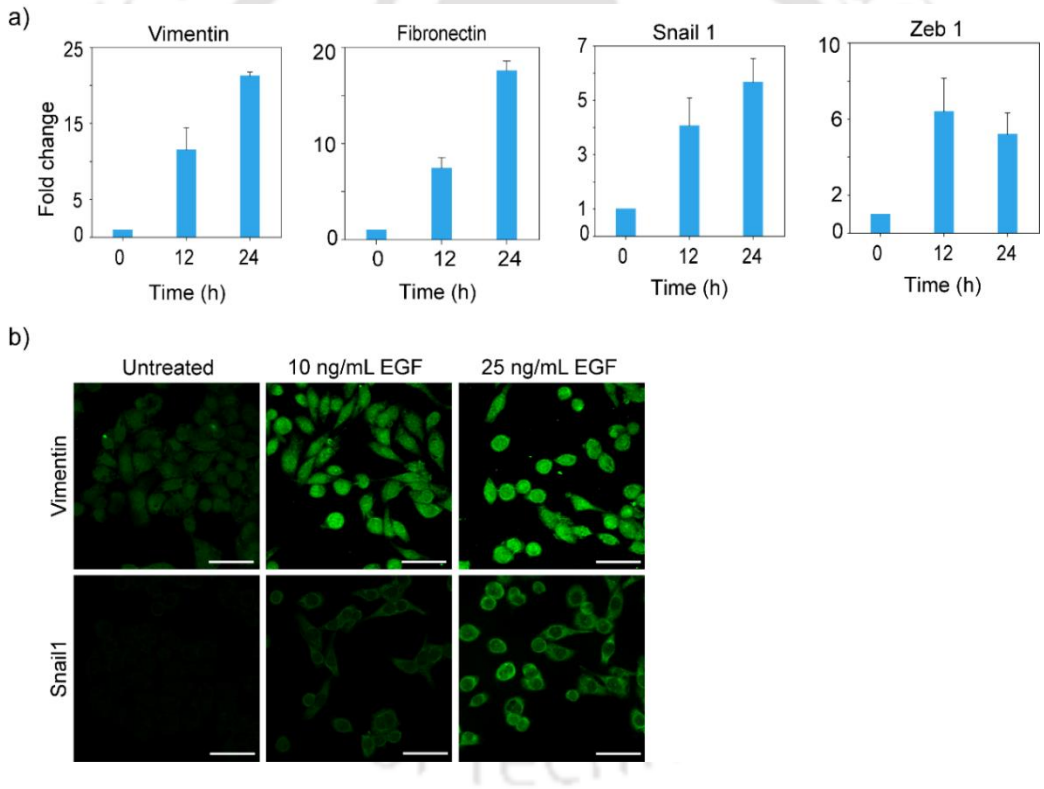



**Figure 4.2: EGF-induced change in the gene expression of EMT markers.**
a) Cells were treated with 10 ng/mL of EGF, and the expression level of EMT markers was measured by qPCR. Fold change in the target gene expression was calculated with respect to untreated cells. Averages of three independent measurements are shown with an error bar representing standard deviation. Observed changes in expression of all the genes were statistically significant (Kruskal-Wallis analysis of variance, $P < 0.01$). Cyclophilin A was used as an internal control. b) Cells were stained with fluorophore-conjugated anti-vimentin and anti-SNAIL1 antibodies, 24 h post EGF treatment. Green color represents the protein-level expression of vimentin (top panel) and SNAIL1 (bottom panel). Scale bar in images: 50 μm.

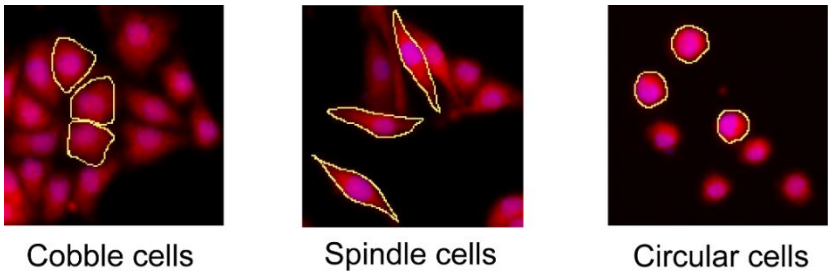



**Figure 4.3: Distinct morphologies of MDA-MB-468 cells.**
The population of MDA-MB-468 cells exists in three distinct morphologies: Cobble, Spindle, and Circular cells. MDA-MB-468 cells were stained with a fluorescent dye and imaged using a fluorescence microscope. The distinct shape of each cell type is highlighted.

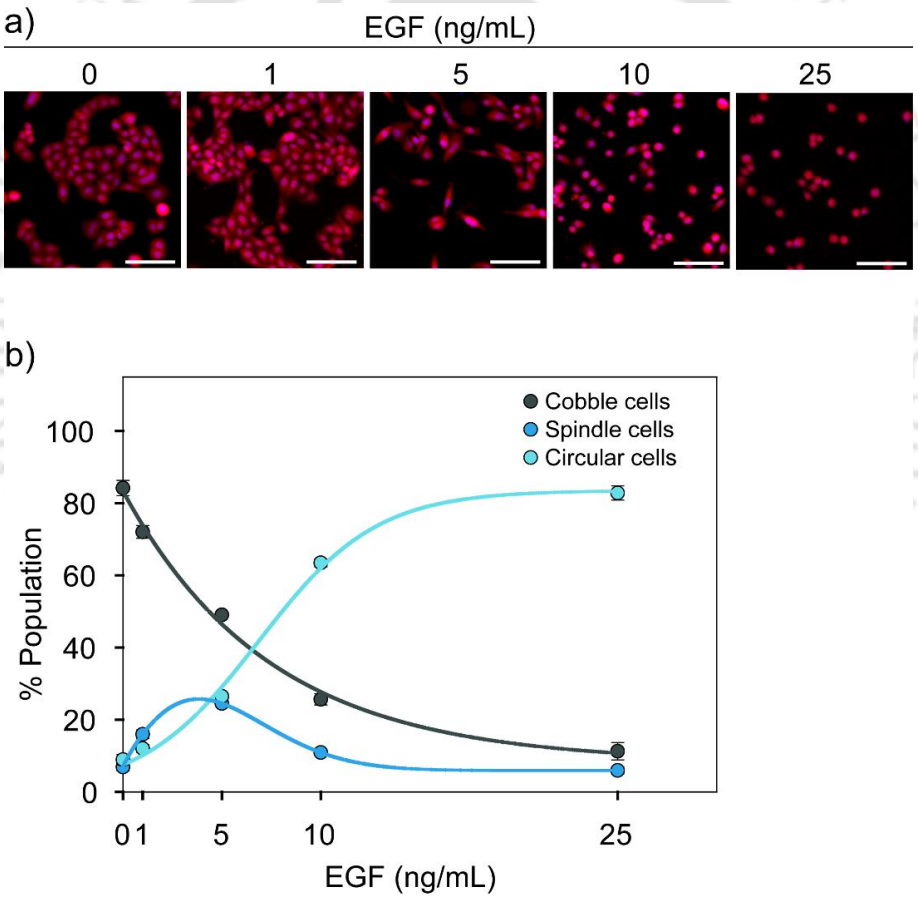



**Figure 4.4: EGF-induced change in the population distribution of MDA-MB-468 cells.**
Cells were treated with different doses of EGF for 24 h. a) Representative images of the population behavior of MDA-MB-468 cells. Scale bar in images: 100 μm. b) The quantitative plot of the EGF-induced morphology changes in MDA-MB-468 cells. Each data represents the mean of three independent experiments, and the error bar represents the standard deviation. The dose-dependent changes in the population distribution of cells in three states were statistically significant (Chi-square test, $P < 0.001$).

## 4.4. Functional characterization of cell states

Our observations show that MDA-MB-468 cells exist in three distinct morphological states. Next, we checked the relevant functional features of these three phenotypic states. The functional features of EMT include the migratory potential of cells, invasion of cells to surrounding tissues, and scattering of cells (*37*).

We performed the Boyden Chamber assay to check the migratory potential of each cell type. The migratory potential of cells is assessed by the capacity of cells to migrate from one side to the other side of a membrane. The complete experimental steps are described in section 3.8 in chapter 3. The migrated cells were stained with HCS CellMask Red Stain and imaged using a confocal microscope (Figure 4.5).

When there was no EGF stimulation, very few cells moved to the other side of the membrane, and they were Spindle and Circular (Figure 4.5b: first row). Therefore, Circular and Spindle are inherently migratory. When cells were treated with varying doses of EGF, a high number of Circular and Spindle cells were observed on the other side of the membrane (Figure 4.5b: last two rows). As observed earlier (Figure 4.4b), EGF treatment induced the formation of Spindle and Circular cells. Since these cells are inherently migratory, they were able to migrate through the membrane. Therefore, the Cobble cells are non-migratory phenotype, while the Spindle and Circular cells are the migratory phenotypes.

We measured the scattering potential of each cell type in the population of MDA-MB-468 cells. Cells were treated with EGF (10 ng/mL) and stained with HCS CellMask Red Stain. We imaged multiple non-overlapping fields of view with a fluorescence microscope. Through image analysis, we measured the number of nearest neighboring cells to each cell. Cells that are present within a radial distance of 5 pixels from the periphery of a cell are classified as neighbors to that specific cell.

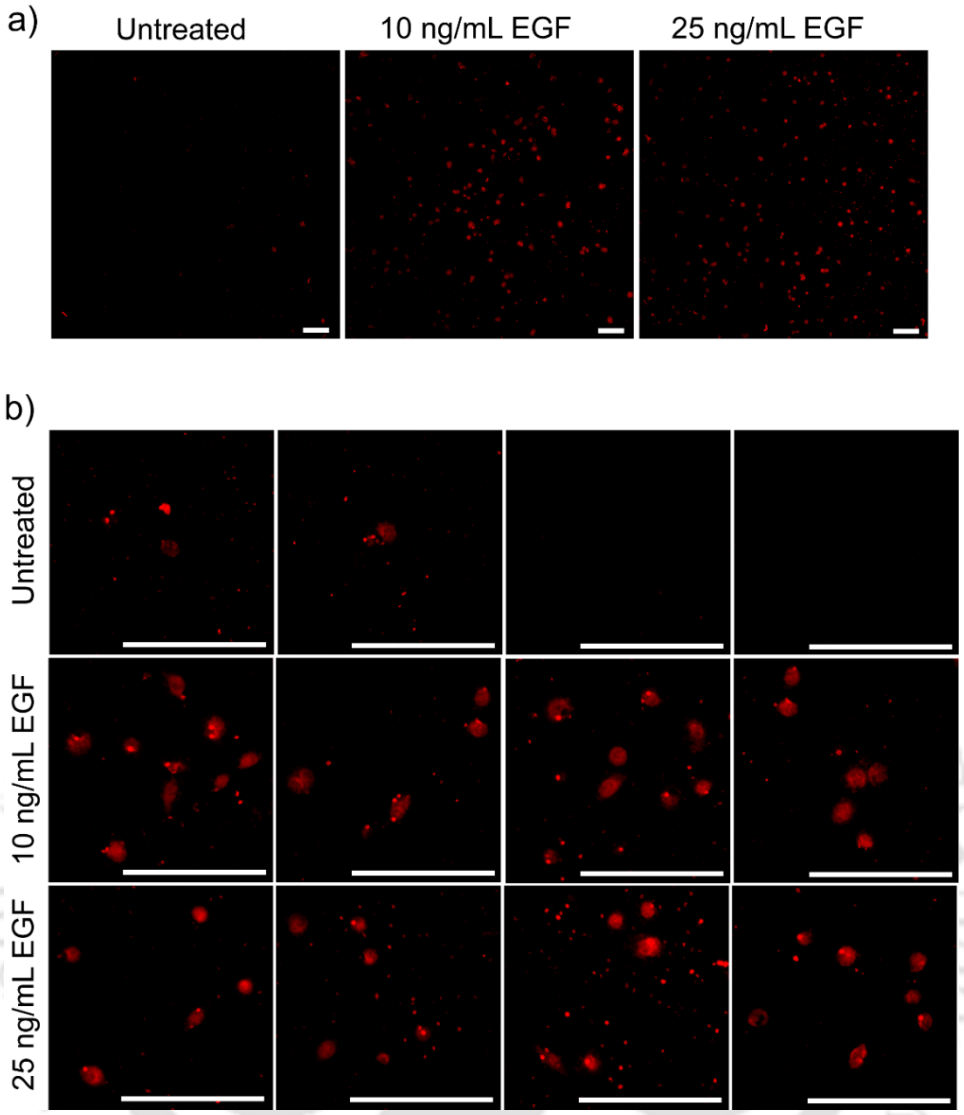



**Figure 4.5: Migratory potential of MDA-MB-468 cells.**
Cells were grown in the presence and absence of EGF, and their migratory potential was assessed through the Boyden Chamber assay. The migrated cells were imaged through a confocal microscope. The entire membrane was imaged using tile scanning ($2 \times 2$ *tiles*), and a representative tile for each dose of EGF is shown in panel-(a). For better visualization of the morphology of cells, the images in panel-(a) were zoomed in and shown in panel-(b). Scale bar in images: 200 μm.

Figure 4.6 shows the distribution of the number of nearest neighbors to each cell type. A higher number of neighbors is a measure of close cell-to-cell contact, and a lower number of neighbors indicate that the cells are scattered more. Cobble cells showed a higher number of neighboring cells when compared to Spindle and Circular cells. Circular cells had the least number of neighbors. The median number of nearest

neighbors to Spindle and Circular cells was 1 and 0, respectively. Therefore, the Cobble cells are colony-loving cells, whereas the Spindle and Circular cells tend to be more scattered.

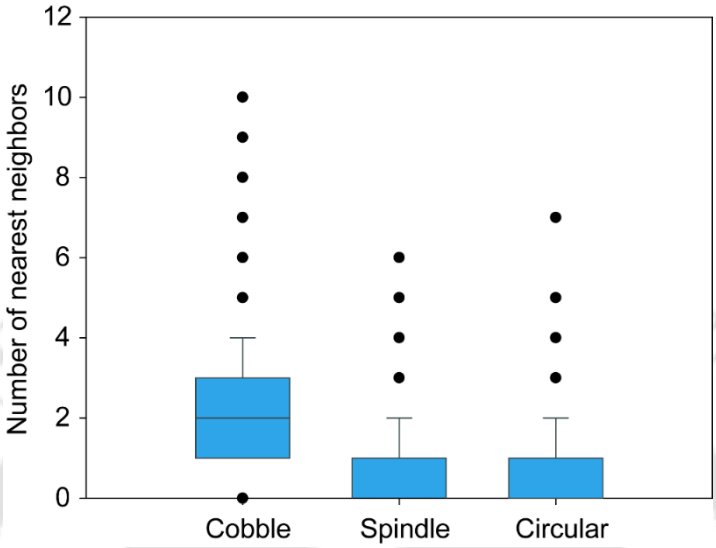



**Figure 4.6: Scattering potential of MDA-MB-468 cells.**
MDA-MB-468 cells were treated with 10 ng/mL of EGF, and the number of nearest neighbors for each cell was measured by image analysis. The differences in the median numbers of nearest neighbors between the three cell types are statistically significant (Kruskal-Wallis test, $P < 0.01$).

## 4.5. Population distribution of MDA-MB-468 cells

To study the dynamics of cell state transition, we cultured MDA-MB-468 cells for a period of 60 h and observed the phenotypic states of MDA-MB-468 cells. For experimental purposes, we discretized time in a 12 h interval. We stained the cells with HCS CellMask Red Stain and recorded the morphologies of MDA-MB-468 cells.

Through image analysis, we measured the percentage of each phenotypic state of MDA-MB-468 cells. Figure 4.7 shows the population distribution of MDA-MB-468 cells in the absence of any external stimulus. On average, 79 % of cells were in the Cobble state, 13 % of cells were in the Spindle state, and 8 % of cells were in the Circular state. The population distribution of untreated cells remained the same until

60 h. Therefore, we considered this distribution of Cobble: Spindle: Circular = 0.79: 0.13: 0.08 as the steady-state distribution.

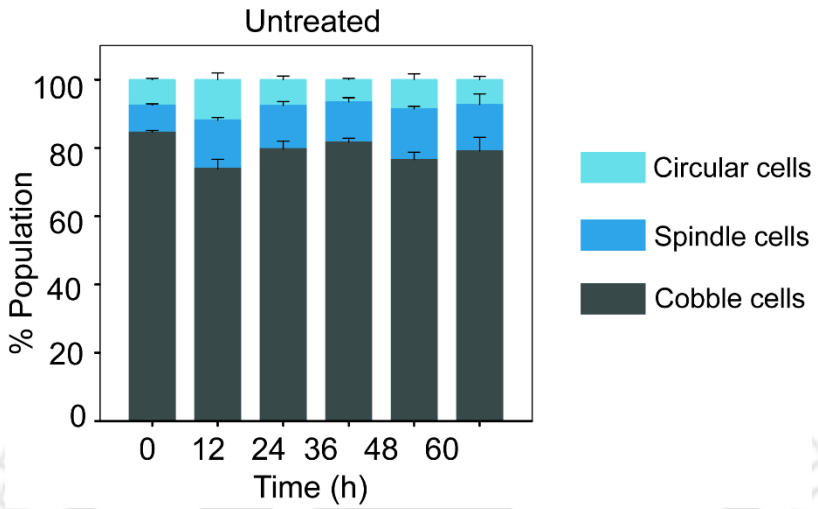



**Figure 4.7: Steady-state population distribution of MDA-MB-468 cells.**
Cells were stained and imaged using a fluorescence microscope. The population distribution of cells was estimated through image analysis. Each bar represents the mean of three independent experiments, and the error bar represents the standard deviation. The time-dependent changes in the population distribution of cells in the three morphological states were not statistically significant (Chi-square test, $P$ = 0.081). The population distribution of MDA-MB-468 cells followed a steady-state distribution in the absence of any external stimulus.

## 4.6. Dose-dependent temporal dynamics of the phenotypic states of MDA-MB-468 cells

We treated the cells with different doses of EGF for a period of 60 h and measured the population distribution of MDA-MB-468 cells. At a lower dose of EGF (1 ng/mL), only a marginal increase in Spindle cells was observed (Figure 4.8a). Whereas at a higher dose of EGF treatment (25 ng/mL), there was a rise in the Circular cell state by 12 h, and the population distribution remained the same state till 60 h (Figure 4.8b).

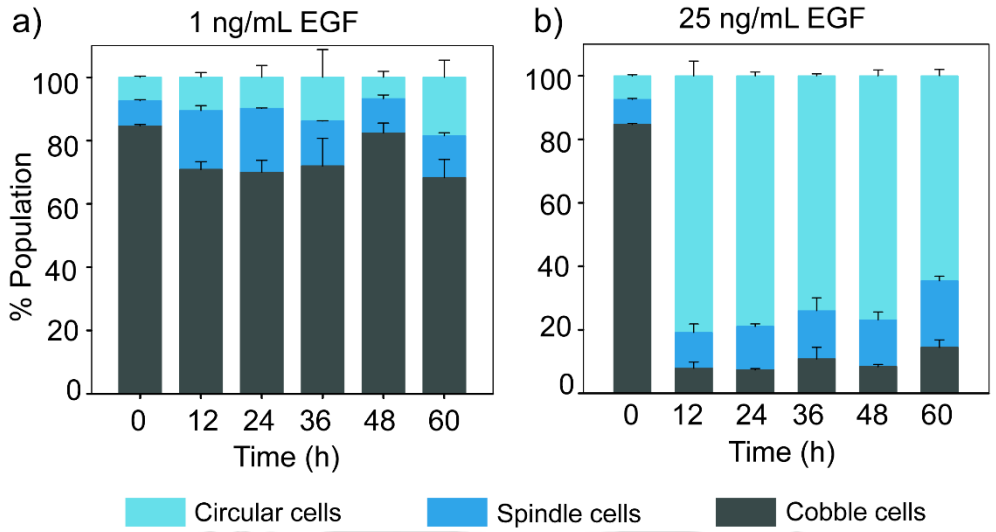



**Figure 4.8: EGF-induced temporal dynamics of the population distribution of MDA-MB-468 cells.**
Cells were treated with EGF, a) 1 ng/mL, and b) 25 ng/mL for a period of 60 h. The cells were stained and imaged using a fluorescence microscope. The population distribution of cells was estimated through image analysis. Each bar represents the mean of three independent experiments, and the error bar represents the standard deviation. The time-dependent changes in the population distribution of cells in (a) and (b) were statistically significant (Chi-square test, $P < 0.001$).

At moderate doses of EGF treatment (5 and 10 ng/mL), there was an initial increase in Circular cells, followed by a hike in the Spindle cell population. Eventually, the population distribution returned to the initial steady-state distribution (Figure 4.9).

EGF treatment disturbed the steady-state distribution of cells. Depending on the dose of EGF, the temporal dynamics of the population distribution varied. The changes in population distribution were reversible in case of moderate doses of EGF treatment (5 and 10 ng/mL) and irreversible in the case of a higher dose of EGF treatment (25 ng/mL). These observed changes in the cell state distribution can happen through: a) transition of cells from one state to the other state; b) extensive cell proliferation of one cell type, and extensive cell death of another cell type; c) both cell state transition and cell proliferation and death.

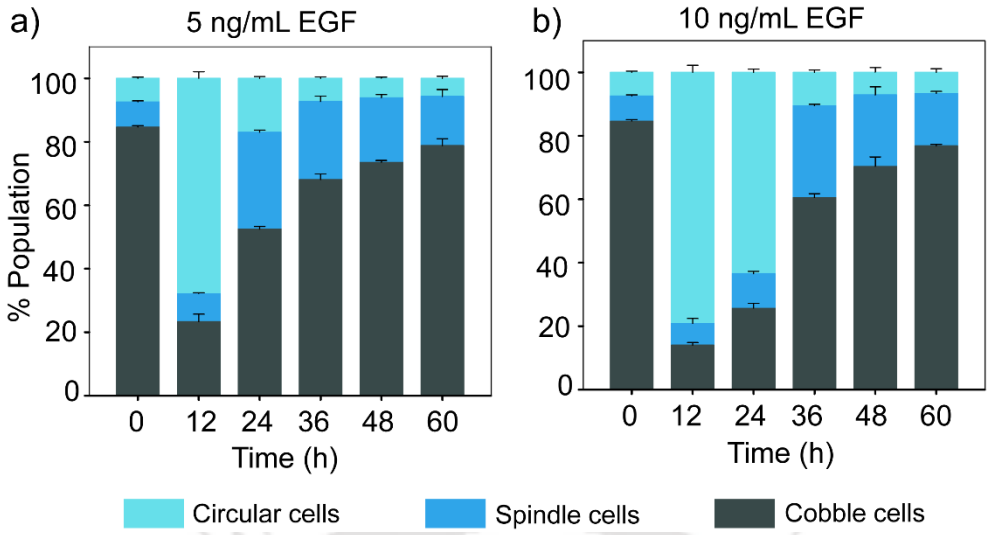



**Figure 4.9: EGF-induced reversible change in the population distribution of MDA-MB-468 cells.**
Cells were treated with EGF, a) 5 ng/mL, and b) 10 ng/mL for a period of 60 h. The cells were stained and imaged using a fluorescence microscope. The population distribution of cells was estimated through image analysis. Each bar represents the mean of three independent experiments, and the error bar represents the standard deviation. The time-dependent changes in the population distribution of the three morphological states in (a) and (b) were statistically significant (Chi-square test, $P < 0.001$).

## 4.7. Effect of EGF on cell proliferation and cell death

EGF is known to induce cell proliferation (*151, 152*) as well as cell death (*153-155*). We measured both cell proliferation and cell death in our experimental system. We performed a propidium iodide (PI)-based plate reader assay to measure the total cell number (live + dead). Figure 4.10 shows the fold change in total cell number for different doses of EGF at varying time points. An increase in total cell number is a measure of cell proliferation since the total cell number includes both live and dead cells accumulated till that time point. The total cell number cannot decrease as we are not removing any cells, and if there is no change in total cell number, then there is no cell division. We observed cell division in the absence of EGF, and cell proliferation decreased with an increase in the dose of EGF treatment.

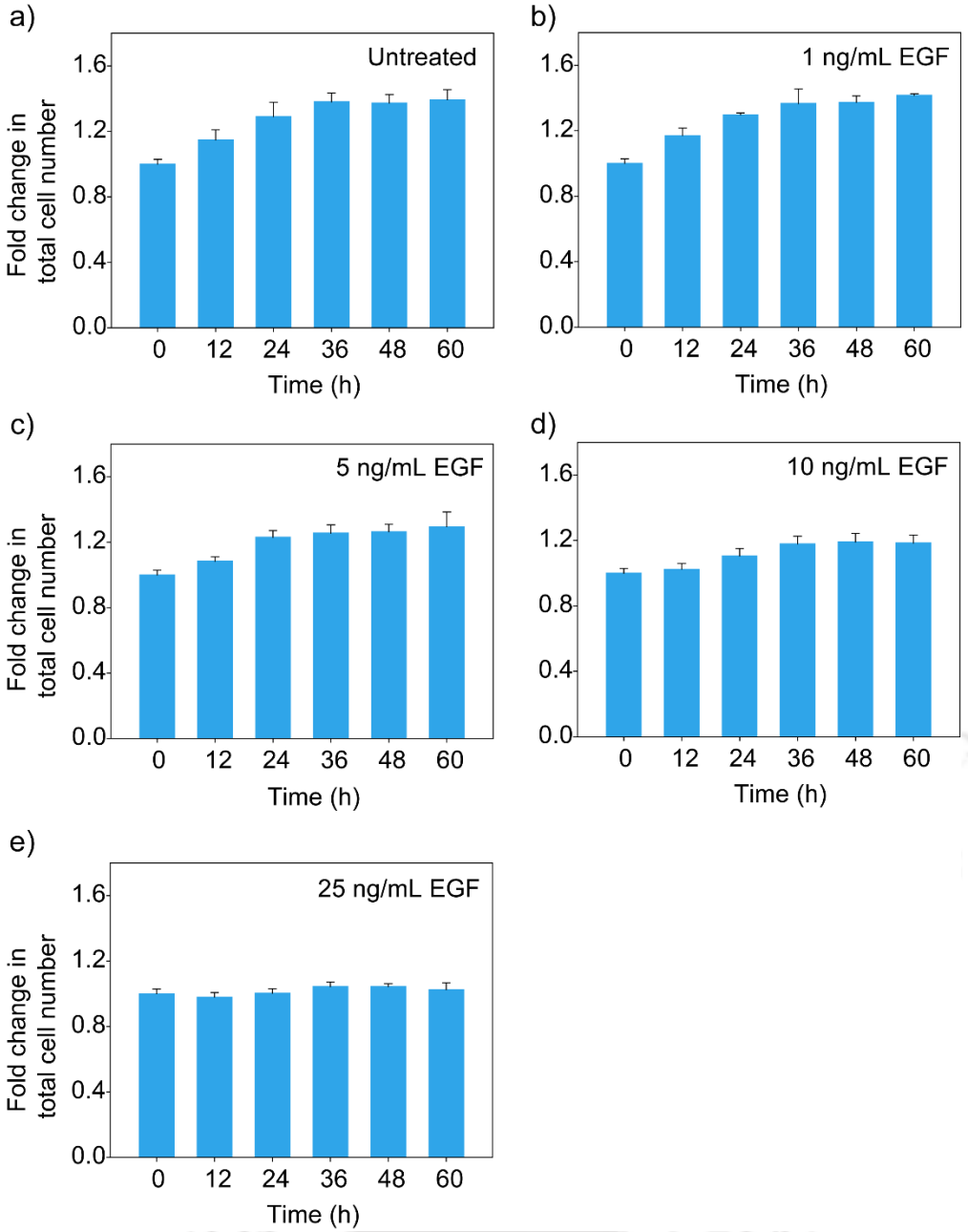



**Figure 4.10: Dose-dependent effect of EGF on total cell number.**
Cells were treated with EGF, and the fold change in total cell number (live + dead) was measured. Each data point represents the mean of three independent assays, and error bars indicate standard deviation. Fold change in cell number with time was statistically significant only for untreated, 1 ng/mL, 5 ng/mL and 10 ng/mL EGF treated cells (Kruskal-Wallis Analysis of Variance, $P < 0.05$). There was no statistically significant difference between untreated and 1 ng/mL EGF treated cells.

Through a similar PI-based plate reader assay, we measured fold change in the dead cell population (Figure 4.11). The percentage of cell death increased with an increase in the dose of EGF treatment.

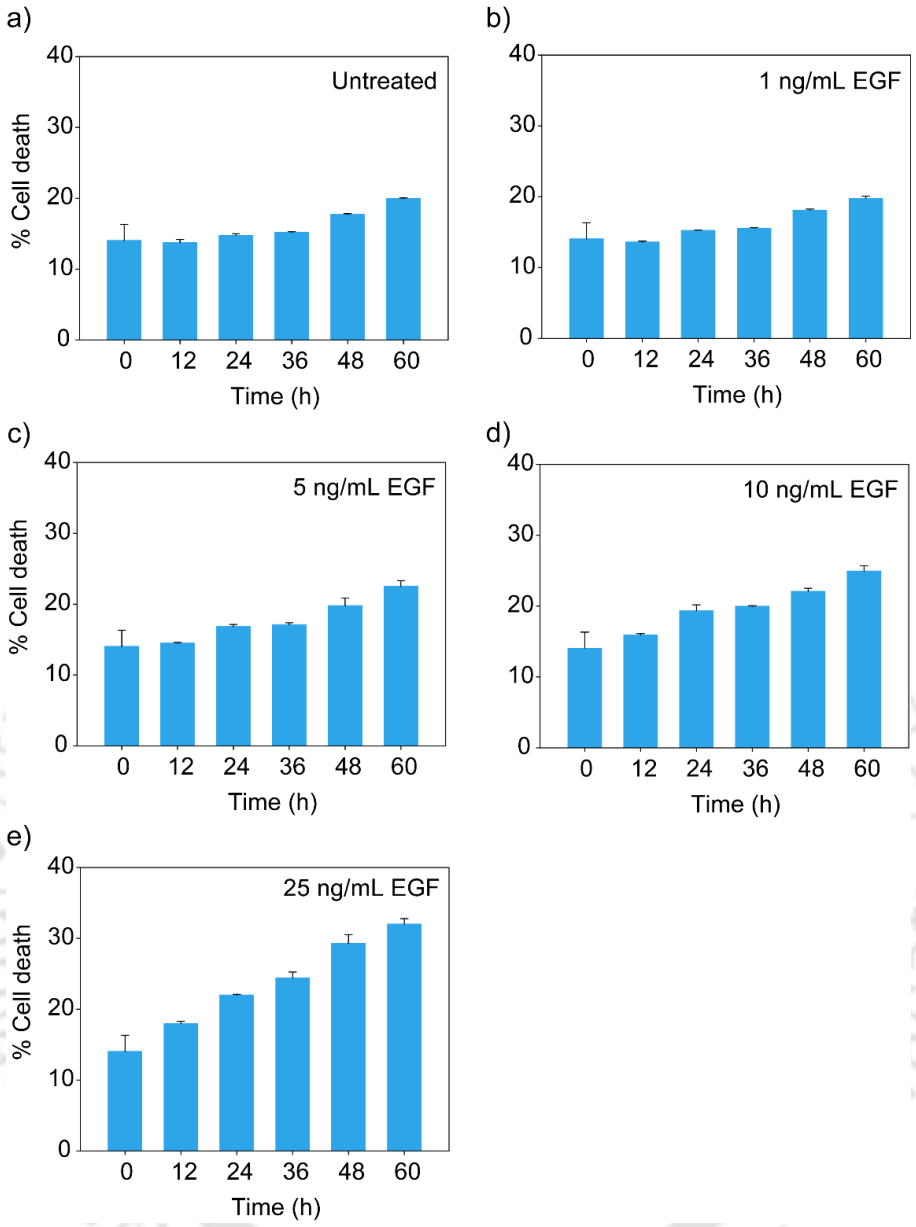



**Figure 4.11: Dose-dependent effect of EGF on cell death.**
Cells were treated with EGF, and the percentage of cell death was measured. Each data point represents the mean of three independent assays, and error bars indicate standard deviation. The time-dependent change in cell death was statistically significant in all EGF treated cases (Analysis of variance, $P < 0.01$).

## 4.8. The cell state transition model

To understand the evolution of each cell type in EGF-induced EMT, we developed a mathematical model. Many existing state transition models consider the transition rate from one state to another state as constant across time (*5*, *6*, *54*). Existing models

do not consider cell birth and death. However, in our experimental system, the population is not conserved. We observed a considerable amount of cell birth and death (Figure 4.10 and Figure 4.11). Also, we observed reversibility in the population distribution of cells (Figure 4.9). Therefore, we cannot use a constant transition rate.

We created a discrete-time population dynamic model that considers cell state transition from one state to the other state as well as cell birth and death. In our experimental system, we discretized time in an interval of 12 h till 60 h. We already measured the distribution of different cell types through image analysis and the fold change in total cell number. Using these data in the mathematical model, we constructed the evolutionary trajectories of each cell type during EGF-induced EMT.

### 4.8.1. The mathematical model

In our experiments, we observed three distinct cell-types based on morphology: Cobble (CB), Spindle (ES), and Circular (CR). Each of these cell types is considered as an individual cell state, and a cell can transit from one state to another. The total number of cells in our experimental system varies with time as cells divide and die. Therefore, we considered the death and birth of cells. To accommodate cell death in our model, we used an absorbing state called Dead state (DD). The state transition model is graphically represented in Figure 4.12.

Mathematically this state transition model is represented by the following set of conservation equations for a time interval $[t, t+\Delta t]$:

$$N_{CB}(t+\Delta t) = N_{CB}(t)\times F_{CB-CB}(t) + N_{ES}(t)\times F_{ES-CB}(t) + N_{CR}(t)\times F_{CR-CB}(t) + N_{CB}(t)\times q_{CB}(t) \quad (4.1)$$

$$N_{ES}(t+\Delta t) = N_{CB}(t)\times F_{CB-ES}(t) + N_{ES}(t)\times F_{ES-ES}(t) + N_{CR}(t)\times F_{CR-ES}(t) + N_{ES}(t)\times q_{ES}(t) \quad (4.2)$$

$$N_{CR}(t+\Delta t) = N_{CB}(t)\times F_{CB-CR}(t) + N_{ES}(t)\times F_{ES-CR}(t) + N_{CR}(t)\times F_{CR-CR}(t) + N_{CR}(t)\times q_{CR}(t) \quad (4.3)$$

$$N_{DD}(t+\Delta t) = N_{CB}(t)\times F_{CB-DD}(t) + N_{ES}(t)\times F_{ES-DD}(t) + N_{CR}(t)\times F_{CR-DD}(t) + N_{DD}(t)\times F_{DD-DD}(t) \quad (4.4)$$

where $N_i(t)$ and $N_i(t+\Delta t)$ are the total number of cells (live as well as dead) in the $i^{th}$ state at $time = t$ and $time = t + \Delta t$ respectively; $i$ = *CB, ES, CR, DD*.

$F_{i-j}(t)$ is the fraction of cells of $i^{th}$ state moving to the $j^{th}$ state in the interval [*t, t+ Δt*]; $i$ = *CB, ES,* and *CR*; $j$ = *CB, ES, CR,* and *DD*.

$q_i(t)$ is the fraction of cells in $i^{th}$ state dividing in the interval [*t, t+ Δt*]; $i$ = *CB, ES,* and *CR*.

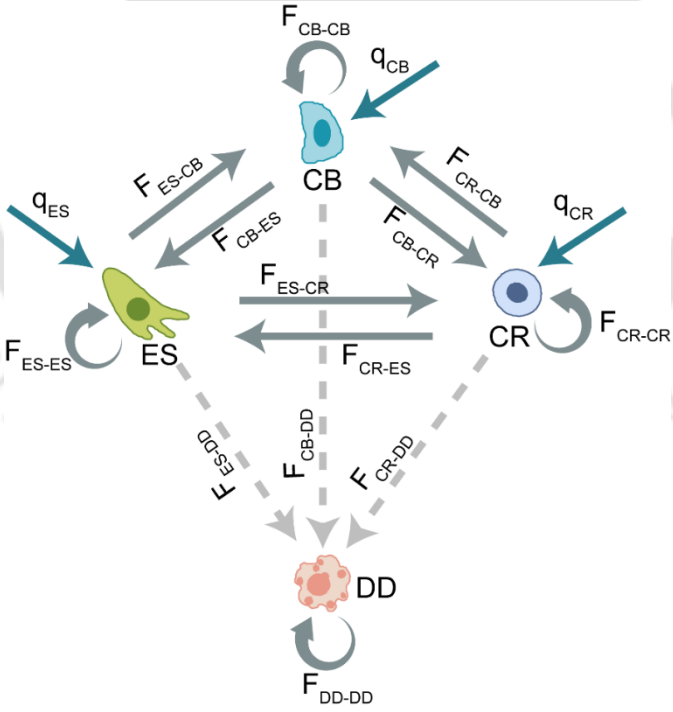



**Figure 4.12: State transition model.**
A live cell can be in any of the three morphological states CB, ES, and CR. Any cell can die and move to the dead state (DD). Birth of a new cell in a state is shown by the arrow without a source. $F_{i-j}$ represents the fraction of cells in the $i^{th}$ state, moving to the $j^{th}$ state in a time interval. $q_i$ is the fraction of cells in $i^{th}$ state dividing in a time interval.

The system has the following constraints:

a) $0 \leq F_{i-j}(t) \leq 1$, for i = CB, ES and CR; j = CB, ES, CR and DD.

b) $\sum_{j=CB,ES,CR,DD} F_{i-j}(t) = 1$, for i = CB, ES, CR.

c) $F_{DD-j}(t) = 0 \; and \; F_{DD-DD}(t) = 1$ for j = CB, ES, CR.

d) $0 \leq q_i(t) \leq 1$, for $i = CB, ES,$ and $CR$.

Equation (4.1) can be re-written in terms of the fraction of cells in each state,

$$\frac{N_{CB}(t+\Delta t)}{N(t)} = \frac{N_{CB}(t)}{N(t)} \times F_{CB-CB}(t) + \frac{N_{ES}(t)}{N(t)} \times F_{ES-CB}(t) + \frac{N_{CR}(t)}{N(t)} \times F_{CR-CB}(t) + \frac{N_{CB}(t)}{N(t)} \times q_{CB}$$

$$\frac{N_{CB}(t+\Delta t)}{N(t)} = f_{CB}(t) \times F_{CB-CB}(t) + f_{ES}(t) \times F_{ES-CB}(t) + f_{CR}(t) \times F_{CR-CB}(t) + f_{CB}(t) \times q_{CB} \quad (4.5)$$

Also,

$$\frac{N_{CB}(t+\Delta t)}{N(t)} = \frac{N(t+\Delta t)}{N(t)} \times \frac{N_{CB}(t+\Delta t)}{N(t+\Delta t)} = fd(t) \times f_{CB}(t+\Delta t) \quad (4.6)$$

Here, Fold change in the total number of cells (live and dead) in the interval [$t$, $t+\Delta t$],

$$fd(t) = \frac{N(t+\Delta t)}{N(t)}$$

$f_i(t)$ and $f_i(t+\Delta t)$ are fraction of cells in the $i^{th}$ state at time $t$ and $(t+\Delta t)$ respectively; $i$ = $CB, ES, CR, DD$.

From equation (4.5) and (4.6),

$$fd(t)\times f_{CB}(t+\Delta t)=f_{CB}(t)\times F_{CB-CB}(t)+f_{ES}(t)\times F_{ES-CB}(t)+f_{CR}(t)\times F_{CR-CB}(t)+f_{CB}(t)\times q_{CB}(t) \quad (4.7)$$

Similarly, equation (4.2) - (4.4) can be re-written as,

$$fd(t)\times f_{ES}(t+\Delta t)=f_{CB}(t)\times F_{CB-ES}(t)+f_{ES}(t)\times F_{ES-ES}(t)+f_{CR}(t)\times F_{CR-ES}(t)+f_{ES}(t)\times q_{ES}(t) \quad (4.8)$$

$$fd(t)\times f_{CR}(t+\Delta t)=f_{CB}(t)\times F_{CB-CR}(t)+f_{ES}(t)\times F_{ES-CR}(t)+f_{CR}(t)\times F_{CR-CR}(t)+f_{CR}(t)\times q_{CR}(t) \quad (4.9)$$

$$fd(t)\times f_{DD}(t+\Delta t)=f_{CB}(t)\times F_{CB-DD}(t)+f_{ES}(t)\times F_{ES-DD}(t)+f_{CR}(t)\times F_{CR-DD}(t)+f_{DD}(t)\times F_{DD-DD}(t) \quad (4.10)$$

Summation of equations (4.7) - (4.10) gives the overall conservation equation of the system,

$$fd(t)=1+[f_{CB}(t)\times q_{CB}(t)+f_{ES}(t)\times q_{ES}(t)+f_{CR}(t)\times q_{CR}(t)] \quad (4.11)$$

### 4.8.2. Estimation of model parameters for the state transition model

We measured the fraction of cells in different cell states (CB, ES, CR) at different time points (t = 0, 12 h, 24 h, 36 h, 48 h, and 60 h) by image analysis (Figure 4.7 - Figure 4.9). We also measured the fraction of dead cells and the fold change in the total cell number (live as well as dead) at each time point (Figure 4.10 and Figure 4.11). We used these data to estimate the model parameters.

### 4.8.3. Estimation of the fraction of dividing cells

Our experimental observations are at six time points (t = 0, 12 h, 24 h, 36 h, 48 h, 60 h). We used equation (4.11), to estimate the fraction of cells dividing at each time interval. Equation (4.11) can be written in matrix form,

$$\left(fd(t)-1\right)=\left(f_{CB}(t)\quad f_{ES}(t)\quad f_{CR}(t)\right)\times\begin{pmatrix}q_{CB}(t)\\q_{ES}(t)\\q_{CR}(t)\end{pmatrix}$$

In vector notation, this equation is written as,

$$K(t)=\boldsymbol{F}(t)\times\boldsymbol{q}(t)$$

We estimated the unknown $\boldsymbol{q}(t)$, through linear least-square optimization. We used the Trust-Region-Reflective Algorithm implemented in MATLAB *lsqlin* function (*169*). $K(t)$ is the observed data that deviates from the true value by an error margin. Therefore, the above equation becomes,

$$K(t)=\boldsymbol{F}(t)\times\boldsymbol{q}(t)+e(t)$$

where $e(t)$ is the residual, which denotes the difference between the experimental and the predicted data.

The above equation can be transformed to estimate the sum of square error as follows,

$$e(t)^T\times e(t)=(K(t)-\boldsymbol{F}(t)\times\boldsymbol{q}(t))^T\times(K(t)-\boldsymbol{F}(t)\times\boldsymbol{q}(t)) \quad (4.12)$$

The above equation is the objective function that yields the sum of the squared error. The unknown $\boldsymbol{q}(t)$ is estimated by minimizing the objective function, $\underset{\boldsymbol{q}(t)}{min}[e(t)^T\times e(t)]$

The estimated $\boldsymbol{q_i}$ values are given in Appendix Table A-1.

### 4.8.4. Estimation of the fractional flow of cells from one state to other states

Cell state transition in our experiment is reversible, and key signaling processes like phosphorylation of EGFR changes with time. Also, we did not observe any steady-state for cell state transition data in the interval of 0 to 60 h. Therefore, we cannot

consider a constant flow rate of cells from one state to another. We estimated the fractional flow of cells from one state to another state for each of the time intervals.

Equations (4.7) - (4.10) can be written in matrix format as follows,

$$\begin{pmatrix} fd \times f_{CB} \\ fd \times f_{ES} \\ fd \times f_{CR} \\ fd \times f_{DD} \end{pmatrix}_{t+\Delta t} = \begin{pmatrix} F_{CB-CB} + q_{CB} & F_{ES-CB} & F_{CR-CB} & 0 \\ F_{CB-ES} & F_{ES-ES} + q_{ES} & F_{CR-ES} & 0 \\ F_{CB-CR} & F_{ES-CR} & F_{CR-CR} + q_{CR} & 0 \\ F_{CB-DD} & F_{ES-DD} & F_{CR-DD} & 1 \end{pmatrix}_t + \begin{pmatrix} f_{CB} \\ f_{ES} \\ f_{CR} \\ f_{DD} \end{pmatrix}_t$$

Following the constraint of the model, $\sum_{j=CB,ES,CR,DD} F_{i-j}(t) = 1$, where $i$ = *CB* or *ES* or *CR*, we do not need to estimate the fractional flow of cells to dead state separately. Therefore, the above matrix can be reduced to,

$$\begin{pmatrix} fd \times f_{CB} \\ fd \times f_{ES} \\ fd \times f_{CR} \end{pmatrix}_{t+\Delta t} = \left[ \begin{pmatrix} F_{CB-CB} & F_{ES-CB} & F_{CR-CB} \\ F_{CB-ES} & F_{ES-ES} & F_{CR-ES} \\ F_{CB-CR} & F_{ES-CR} & F_{CR-CR} \end{pmatrix}_t + \begin{pmatrix} q_{CB} & 0 & 0 \\ 0 & q_{ES} & 0 \\ 0 & 0 & q_{CR} \end{pmatrix}_t \right] \times \begin{pmatrix} f_{CB} \\ f_{ES} \\ f_{CR} \end{pmatrix}_t$$

This can be written in vector notion as,

$$\boldsymbol{C}(t+\Delta t) = [\boldsymbol{A}(t) + \boldsymbol{Q}(t)] \times \boldsymbol{B}(t) \tag{4.13}$$

We had estimated the elements of $\boldsymbol{Q}(t)$ in section 4.8.3. Using equation (4.13), we estimated $\boldsymbol{A}(t)$for each 12 h interval, from the experimental data through parameter optimization. The objective function for this optimization can be written as $\underset{\boldsymbol{A}(t)}{min}[\boldsymbol{e}(t)^{\boldsymbol{T}}\boldsymbol{e}(t)]$, where $\boldsymbol{e}(t)$ is the residual between observed and estimated $\boldsymbol{C}(t + \Delta t)$.

The matrix $\boldsymbol{A}(t)$ has 12 unknown parameters (9 state transition parameters and three cell division parameters). The cell division parameters were estimated separately (section 4.8.3). The system is underdetermined, which can cause overfitting and can generate extremely different $\boldsymbol{A}(t)$ for each time interval. That would be unrealistic as

the cell state transition parameters in two successive time points in a real biological system will not differ drastically.

Following Chiba et al. (*170*), we assume that the difference in each element of $\boldsymbol{A}(t)$ of two consecutive time intervals is small in terms of L1-norm of the difference. This allows us to constrain the parameter space and estimate the $\boldsymbol{A}(t)$ for all the time intervals simultaneously. Therefore, two objective functions were used to estimate the fractional state transition parameters. We used a well-established multi-objective genetic algorithm, NSGA-II, proposed by Deb (*171*) that is implemented in MATLAB function *gamultiobj*. This was used to minimize both the objective functions simultaneously.

The objective functions used for the estimation of cell state transition parameters are,

Objective function 1 = $\underset{\boldsymbol{A}(t)}{min}(\sum_{t\in T}\sum(\boldsymbol{e}(t)^T\boldsymbol{e}(t)))$

Objective function 2 = $\underset{\boldsymbol{A}(t)}{min}(\sum_{t\in T}\sum(|\boldsymbol{A}(t+\Delta t)-\boldsymbol{A}(t)|))$

Here, $T$ = (0, 12 h, 24 h, 36 h, 48 h) and $\Delta t$ = 12 h.

In the case of the multi-objective function, the algorithm converges to a set of solutions called Pareto front. These solutions are all optimal since one objective function cannot be improved without sacrificing the other. If the Pareto front is convex, the knee point estimate is used to pick the best among all optimal solutions (*172*). The knee of a curve is a point on a curve, where the curvature is maximal. In our experimental system, the Pareto front of the two objective functions was convex, and we used the knee point to pick the best of the Pareto front.

We performed 1000 independent optimization, and the best parameter set in each run was decided through Pareto front analysis. The best parameter set from these 1000

optimization runs was considered as the optimized parameter set and used in further analysis. As a representative data of the estimation method, the Pareto front analysis of 10 ng/mL EGF treatment is shown in Figure 4.13. The Pareto front of one of the thousand independent optimizations is shown in Figure 4.13a. The distribution of the best solutions from all these thousand simulations is shown in Figure 4.13b.

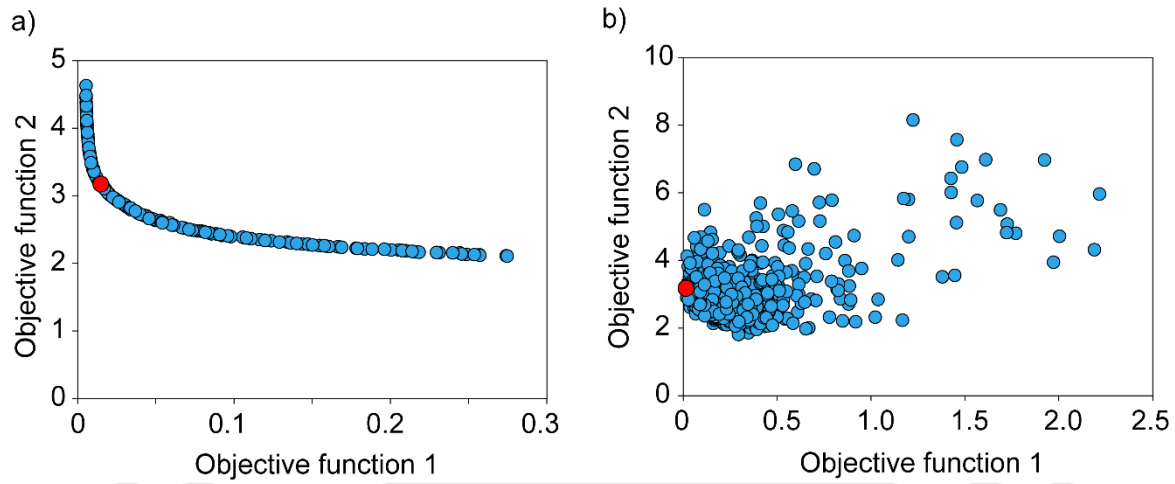



**Figure 4.13: Pareto front analysis of 10 ng/mL EGF treated experiment.**
One thousand independent optimizations were performed. a) Pareto front of one of the 1000 optimizations. Solid blue circles are the possible solutions, and the solid red circle is the best of all possible solutions. b) Each solid circle is the best solution from the Pareto front analysis of 1000 independent optimizations. The solid red circle is the solution with minimum objective function 1 and is considered the best solution of all.

The optimal state transition parameters for all doses of EGF are given in Appendix Table A-2, Table A-3, and Table A-4. Figure 4.14 shows the estimated fraction of cells in each cell state from the model using the optimal parameter set.

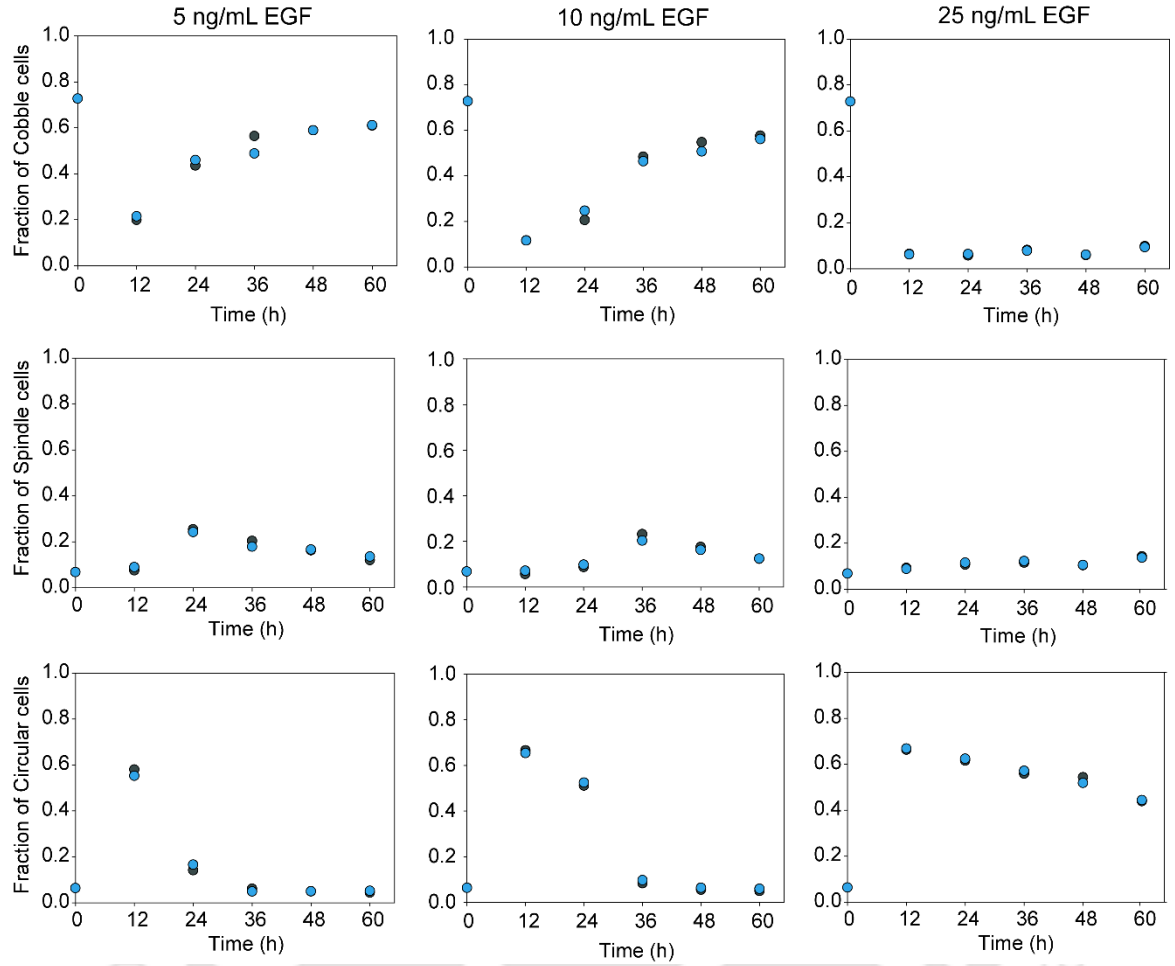



**Figure 4.14: Fraction of each cell state estimated from the state transition model.** The optimal state transition parameters were used to estimate the fraction of each cell state. The plot shows the deviation between the actual experimental values and the values estimated from the state transition model. The black circle represents the experimental values, and the blue circle represents the estimated values from the model.

### 4.8.5. Normalized flux

We visualized the cell state transitions diagrammatically in terms of normalized flux through different state transition paths (only live-cell states). Normalized flux for a cell state transition path $i - j$ in a time interval $[t, t + \Delta t]$ is defined as,

$$J_{i-j}(t) = F_{i-j}(t) \times f_i(t)$$

where $\hat{F}_{i-j}(t)$ is the fraction of live cells moving from $i^{th}$ state to $j^{th}$ state at time $t$. $\hat{f}_i(t)$ is the fraction of live cells in the $i^{th}$ state at time $t$. $i,j$ = *CB, ES,* and *CR*.

## 4.9. Trajectories of cell state transition

We estimated the normalized flux from one state to the other from the mathematical model. Normalized flux denotes the fraction of live cell transition from one state to the other. Figure 4.15 shows the cell state transition trajectories for 5 ng/mL EGF treatment. The black line highlights the dominant flux, while the other minor transitions are de-emphasized. The transition path of the dominant flux is Cobble → Circular → Cobble. The same behavior is observed in the case of 10 ng/mL EGF treatment (Figure 4.16). Here, the cells stay in the Circular state a little longer and revert to Cobble state.

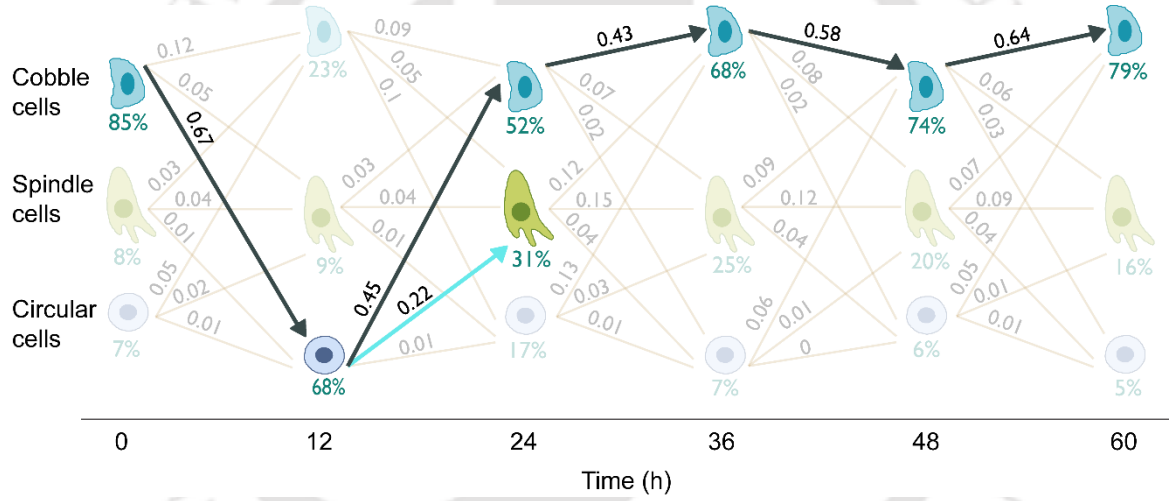



**Figure 4.15: State transition trajectories of cells treated with 5 ng/mL EGF.**
Each line represents the transition from one state to the other. The numerical values over each line are the normalized flux estimated from the state transition model. The percentage population of each cell state is mentioned right below each cell type. The black arrows indicate the dominant transition path. The blue arrow shows the next dominant transition path leading to the evolution of Spindle cells. For better visualization of the plot, the remaining minor transitions are de-emphasized.

In the case of both 5 and 10 ng/mL EGF treatment, we can see a considerable amount of Spindle cell population at 24 h and 36 h, respectively (Figure 4.15 and Figure 4.16). However, the dominant transition path does not explain the emergence of Spindle cells. Only a smaller fraction of Circular cells transitioned to the Spindle state during

the reverse state transition (blue arrow in Figure 4.15 and Figure 4.16). Therefore, the Spindle cells might have branched off from Circular cells, where the primary branch leads to Cobble state, and the second branch leads to Spindle state.

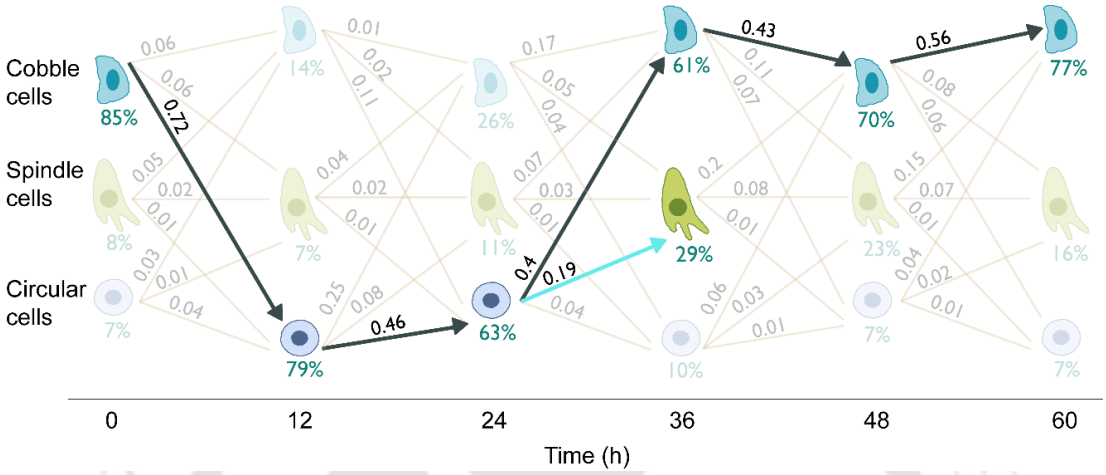



**Figure 4.16: State transition trajectories of cells treated with 10 ng/mL EGF.**
Each line represents the transition from one state to the other. The numerical values over each line are the normalized flux estimated from the state transition model. The percentage population of each cell state is mentioned right below each cell type. The black arrows indicate the dominant transition path. The blue arrow shows the next dominant transition path leading to the evolution of Spindle cells. For better visualization of the plot, the remaining minor transitions are de-emphasized.

To investigate further, we widened the time interval 24-36 h in the case of 10 ng/mL EGF treatment. Also, a sharp change in population distribution occurs at 0-12 h and 24-36 h time interval. Therefore, we recorded the population distribution of cells for every 3 h in both the time interval (Figure 4.17) and constructed the normalized flux diagram for both the cases (Figure 4.18 and Figure 4.19). In the case of 0-12 h interval, most of the cells were initially in Cobble state. The dominant transition path was Cobble → Circular (black arrows in Figure 4.18). Whereas, in the case of 24-36 h time interval, initially, cells were in the Circular state. The dominant transition path was Circular → Spindle (black arrows in Figure 4.19), and eventually, a smaller fraction of the Spindle cells moved to the Cobble state (blue arrows in Figure 4.19). Therefore, the Circular → Cobble state transition is a two-step transition through the intermediate Spindle state.

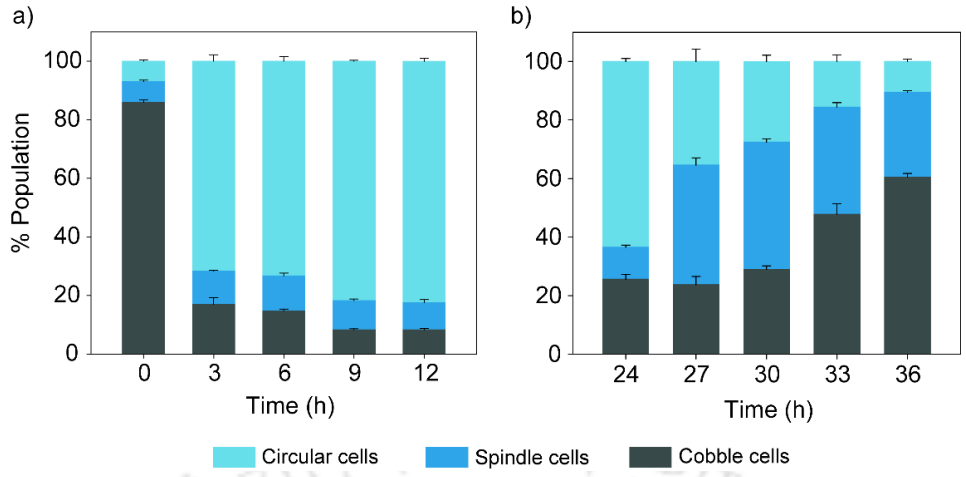



**Figure 4.17: EGF-induced change in the population distribution of cells observed at a shorter time interval.**

Cells were treated with 10 ng/mL EGF, and the cell population was observed for a period of (a) 0-12 h and (b) 24-36 h at an interval of 3 h. The Cells were stained and imaged using a fluorescence microscope. The population distribution of cells was estimated through image analysis. Each bar represents the mean of three independent experiments, and the error bar represents the standard deviation. The time-dependent changes in the distribution of the three morphological states of MDA-MB-468 cells in (a) and (b) were statistically significant (Chi-square test, $P < 0.001$).

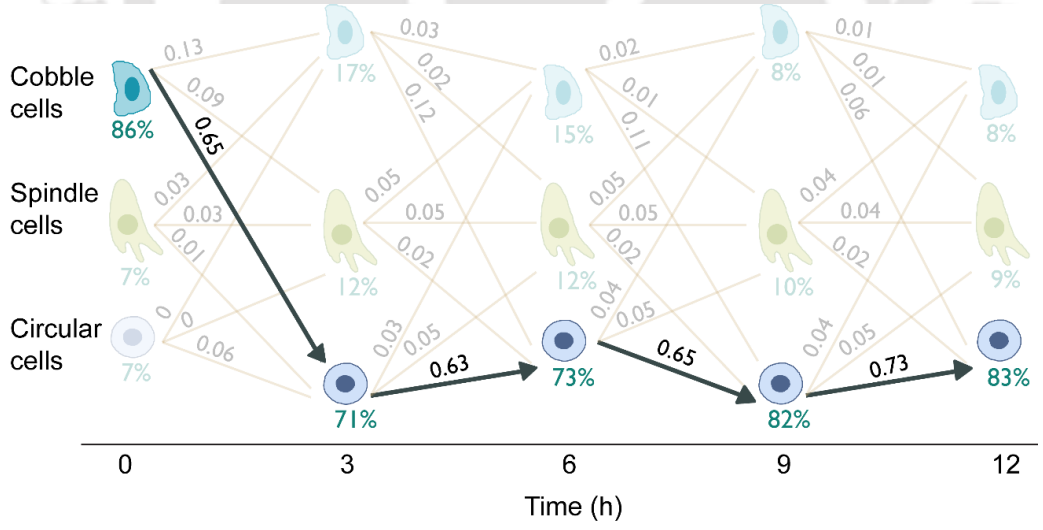



**Figure 4.18: State transition trajectories of cells treated with 10 ng/mL EGF for a period of 0-12 h.**

Each line represents the transition from one state to the other. The numerical values over each line are the normalized flux estimated from the state transition model. The percentage population of each cell state is mentioned right below each cell type. The black arrows indicate the dominant transition path. For better visualization of the plot, the remaining minor transitions are de-emphasized.

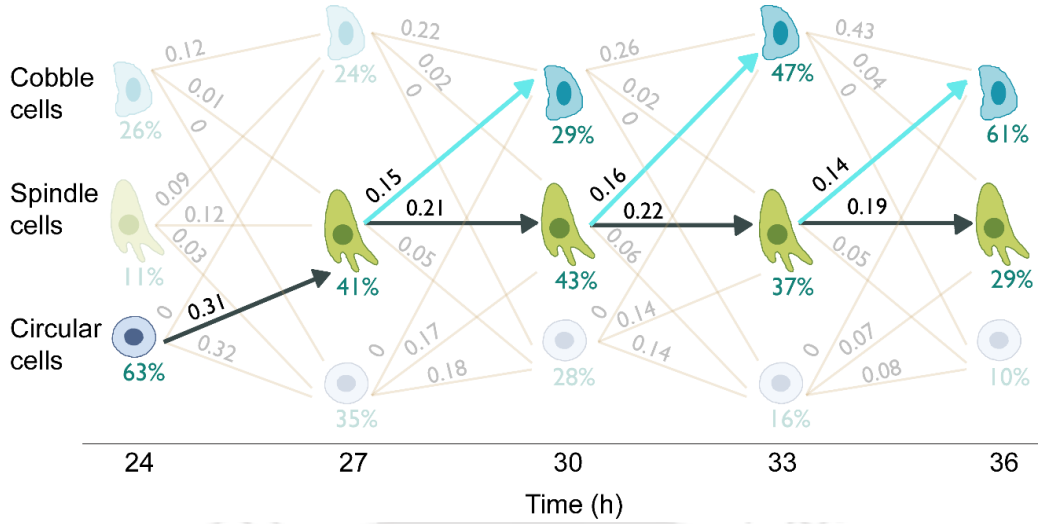



**Figure 4.19: State transition trajectories of cells treated with 10 ng/mL EGF for a period of 24-36 h.**
Each line represents the transition from one state to the other. The numerical values over each line are the normalized flux estimated from the state transition model. The percentage population of each cell state is mentioned right below each cell type. The black arrows indicate the dominant transition path. The blue arrow shows the next dominant transition path leading to the evolution of Cobble cells. For better visualization of the plot, the remaining minor transitions are de-emphasized.

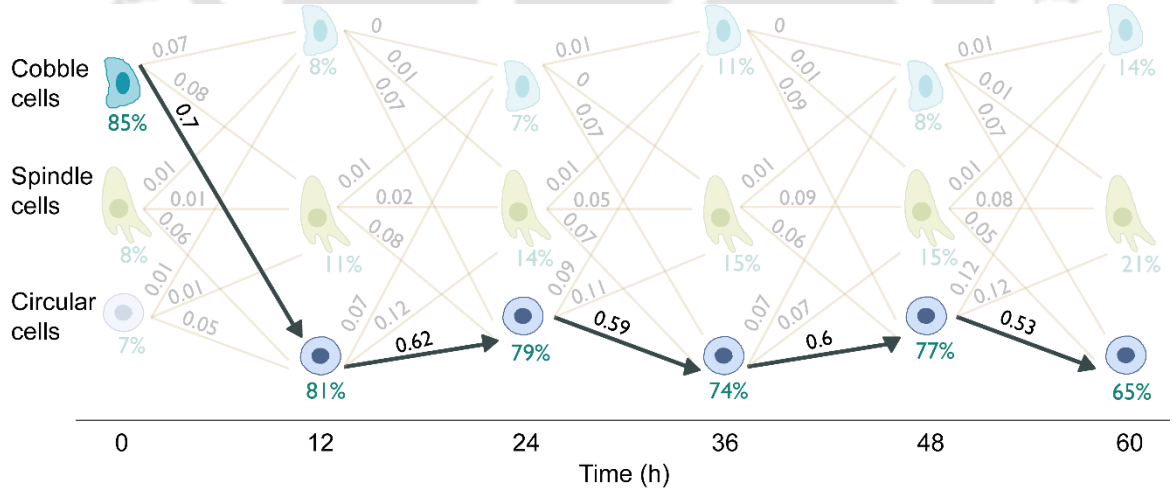



**Figure 4.20: State transition trajectories of cells treated with 25 ng/mL EGF.**
Each line represents the transition from one state to the other. The numerical values over each line are the normalized flux estimated from the state transition model. The percentage population of each cell state is mentioned right below each cell type. The black arrows indicate the dominant transition path. For better visualization of the plot, the remaining minor transitions are de-emphasized.

We also constructed the cell state transition trajectories for 25 ng/mL EGF treated samples. The dominant path was the Cobble → Circular state (black arrows in Figure 4.20). At a higher dose of EGF treatment, we did not observe any reverse transition, and therefore, there was only a marginal amount of Spindle cell population.

The highlights from the state transition model analysis are a) EGF-induced cell state transition is reversible and is dependent on the dose of EGF, b) At moderate doses of EGF treatment, the dominant path is Cobble → Circular → Spindle → Cobble. c) At a higher dose of EGF, the dominant transition path is Cobble → Circular. d) Spindle cells are predominantly formed from the Circular cells, and therefore, the evolution of Spindle cells requires a prerequisite Cobble → Circular transition.

## 4.10. The null model

The observed state transition behavior can also be explained by an extensive cell division of one cell type and extensive cell death of another cell type without any cell state transition. Therefore, we proposed an alternative null model. In the null model, we assumed that the observed changes in the distribution of cells in three morphological states originated solely from cell division and cell death, and there was no transition from one live cell state to another live cell state. The graphical representation of the null model is shown in Figure 4.21.

The model for this is just a reduced version of our complete state transition model,

$$fd(t)\times f_{CB}(t+\Delta t)=f_{CB}(t)\times F_{CB-CB}(t)+f_{CB}(t)\times q_{CB}(t) \tag{4.14}$$

$$fd(t)\times f_{ES}(t+\Delta t)=f_{ES}(t)\times F_{ES-ES}(t)+f_{ES}(t)\times q_{ES}(t) \tag{4.15}$$

$$fd(t)\times f_{CR}(t+\Delta t)=f_{CR}(t)\times F_{CR-CR}(t)+f_{CR}(t)\times q_{CR}(t) \tag{4.16}$$

$$fd(t)\times f_{DD}(t+\Delta t)=f_{CB}(t)\times F_{CB-DD}(t)+f_{ES}(t)\times F_{ES-DD}(t)+f_{CR}(t)\times F_{CR-DD}(t)+f_{DD}(t)\times F_{DD-DD}(t) \quad (4.17)$$

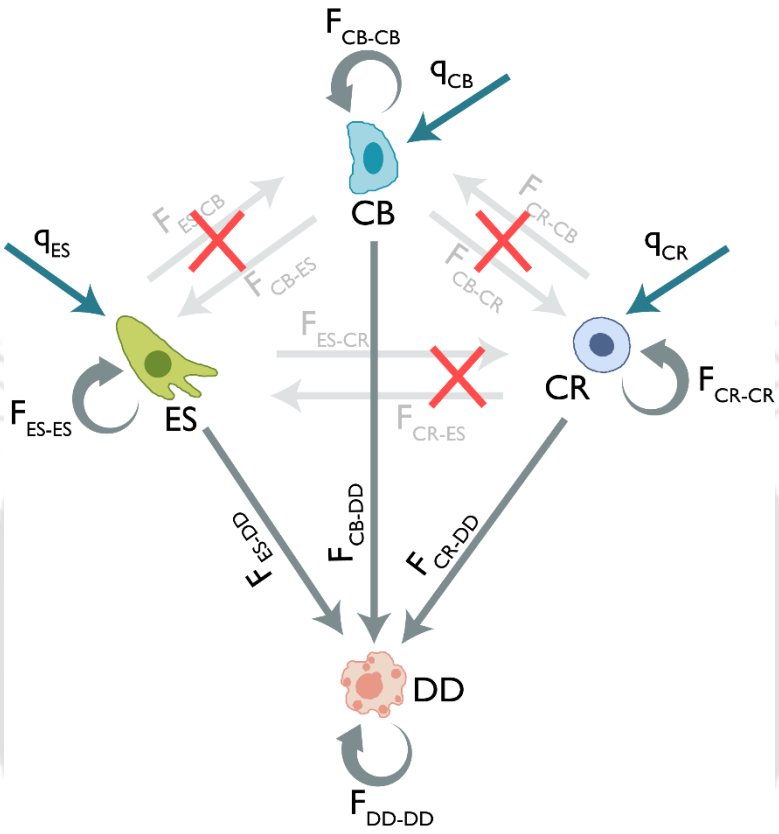



**Figure 4.21: Null model.**
A live cell can be in any of the three morphological states CB, ES, and CR. Any cell can die and move to the dead state (DD). The birth of a new cell in a state is indicated by arrows without any source. $F_{i\text{-}i}$ represents the fraction of cells in the $i^{th}$ state remaining in the same state in a time interval. $q_i$ is the fraction of cells in $i^{th}$ state dividing in a time interval.

The system has the following constraints:

a) $0 \leq F_{i-i}(t) \leq 1$, for i = CB, ES, CR, DD;

b) $0 \leq F_{i-DD}(t) \leq 1$, for i = CB, ES, CR;

c) $F_{i-i} + F_{i-DD} = 1$, for i = CB, ES, CR;

d) $F_{DD-DD}(t) = 1$

e) $0 \leq q_i(t) \leq 1$, for $i$ = CB, ES, CR.

All other notations used in section 4.8.1 remain the same.

Equations (4.14) - (4.17) can be written in matrix format as follows,

$$\begin{pmatrix} fd \times f_{CB} \\ fd \times f_{ES} \\ fd \times f_{CR} \\ fd \times f_{DD} \end{pmatrix}_{t+\Delta t} = \begin{pmatrix} F_{CB-CB} + q_{CB} & 0 & 0 & 0 \\ 0 & F_{ES-ES} + q_{ES} & 0 & 0 \\ 0 & 0 & F_{CR-CR} + q_{CR} & 0 \\ F_{CB-DD} & F_{ES-DD} & F_{CR-DD} & 1 \end{pmatrix}_t + \begin{pmatrix} f_{CB} \\ f_{ES} \\ f_{CR} \\ f_{DD} \end{pmatrix}_t$$

Following the constraint $F_{i-i} + F_{i-DD} = 1$, for $i$ = *CB, ES, CR*, we do not need to estimate the fractional flow of cells to dead state separately. Therefore, the above matrix can be reduced to,

$$\begin{pmatrix} fd \times f_{CB} \\ fd \times f_{ES} \\ fd \times f_{CR} \end{pmatrix}_{t+\Delta t} = \left[ \begin{pmatrix} F_{CB-CB} & 0 & 0 \\ 0 & F_{ES-ES} & 0 \\ 0 & 0 & F_{CR-CR} \end{pmatrix}_t + \begin{pmatrix} q_{CB} & 0 & 0 \\ 0 & q_{ES} & 0 \\ 0 & 0 & q_{CR} \end{pmatrix}_t \right] \times \begin{pmatrix} f_{CB} \\ f_{ES} \\ f_{CR} \end{pmatrix}_t$$

This can be written in vector notion as,

$$\boldsymbol{C}(t+\Delta t) = [\boldsymbol{A}(t) + \boldsymbol{Q}(t)] \times \boldsymbol{B}(t) \tag{4.18}$$

For this null model, we estimated both $\boldsymbol{A}(t)$ and $\boldsymbol{Q}(t)$ together, following the method described in section 4.8.4.

We fitted the population distribution data to the null model and obtained the cell division and cell death parameters. From these parameters, we reverse calculated the fold change in dead cells and the fold change in the total cell population. The results estimated from the null model had a very high deviation from the actual experimental

data (Figure 4.22). The null model predicted an exceptionally high cell death, which is unrealistic. Therefore, we rejected the null model.

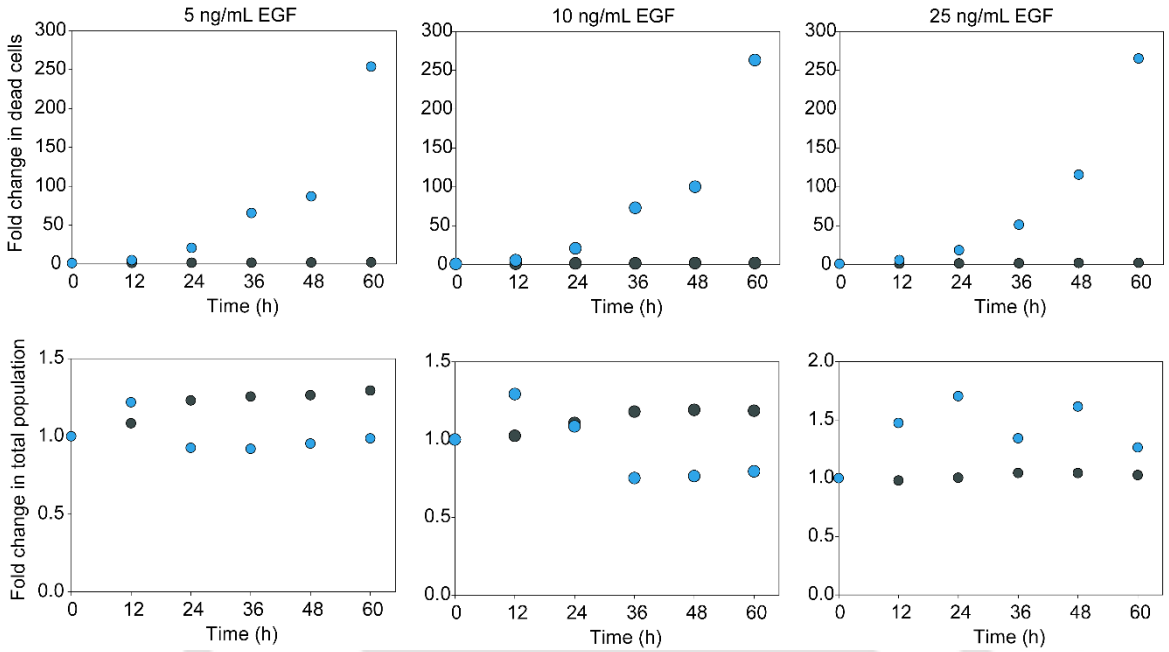



**Figure 4.22: Null model validation.**
The estimated parameters from the null model were used to calculate the fold change in total cell number as well as the fold change in the dead cell population. The plot shows the deviation between the actual experimental values and the values estimated from the null model. The black circle represents the experimental values, and the blue circle represents the estimated values from the model.

## 4.11. The Dynamics of phospho-EGFR drives the state transition

We had observed a dose-dependent temporal dynamics of EGF-induced cell state transition. To get some molecular level insight into the cell state transition, we investigated the EGF signaling. In our experimental system, EGF is the input signal. EGF binds to EGF receptor (EGFR) and phosphorylates the EGFR, followed by the internalization of the EGF-EGFR complex into the cell. The phosphorylated form of EGFR is the active form that regulates several other downstream signaling cascades.

Firstly, we checked the rate of internalization of EGF. We treated cells with different doses of EGF and measured the freely available EGF in the media through sandwich

ELISA. The internalization of EGF followed an exponential decay pattern in a dose-dependent manner (Figure 4.23). We had a control experiment to check the auto degradation of EGF. We added EGF in the media without any cells and measured the available EGF in the media. The amount of EGF remained the same in the media until 60 h (solid black circles in Figure 4.23). Therefore, the auto degradation of EGF in the cell culture medium is very less, and the decline in free EGF is primarily due to the internalization of EGF by the cells.

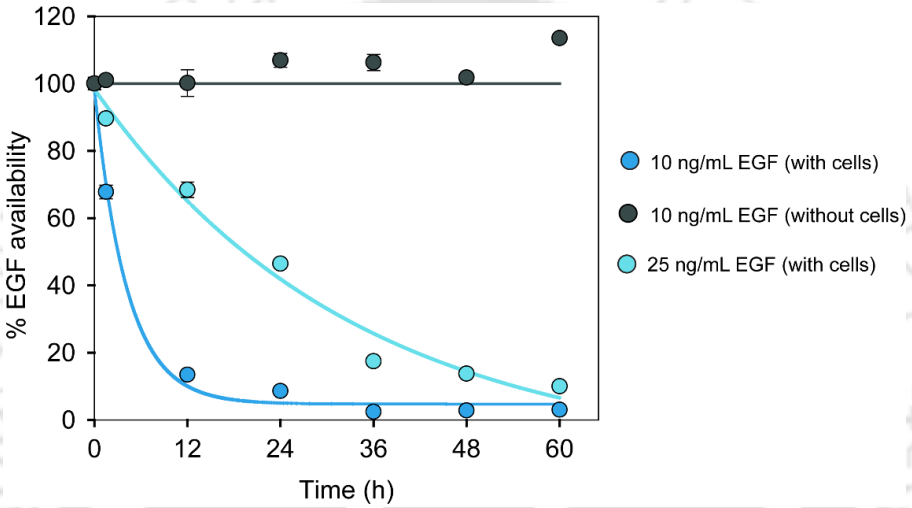



**Figure 4.23: Time-dependent EGF availability to the cells.**
EGF was added to the cells, and the time-dependent availability of free EGF in the media was measured through sandwich-ELISA. Percentage EGF availability was calculated relative to the EGF availability at time = 0.

The first signaling event in EGF signaling is the phosphorylation of the EGFR. We measured the phospho-EGFR level through western blotting. Figure 4.24b shows the quantitative plot of the temporal dynamics of phospho-EGFR. In the case of 5 and 10 ng/mL EGF treatment, phospho-EGFR showed a transient response. There was an instant rise in the phospho-EGFR level, followed by a gradual decline post 12 h of EGF treatment. Whereas, in the case of 25 ng/mL EGF treatment, the phospho-EGFR level remained very high until 36 h of EGF treatment. These data are in line with the internalization of EGF (Figure 4.23). Higher the freely available EGF in the media, the

higher is the phospho-EGFR level. The decay in the phospho-EGFR is majorly due to the unavailability of external EGF.

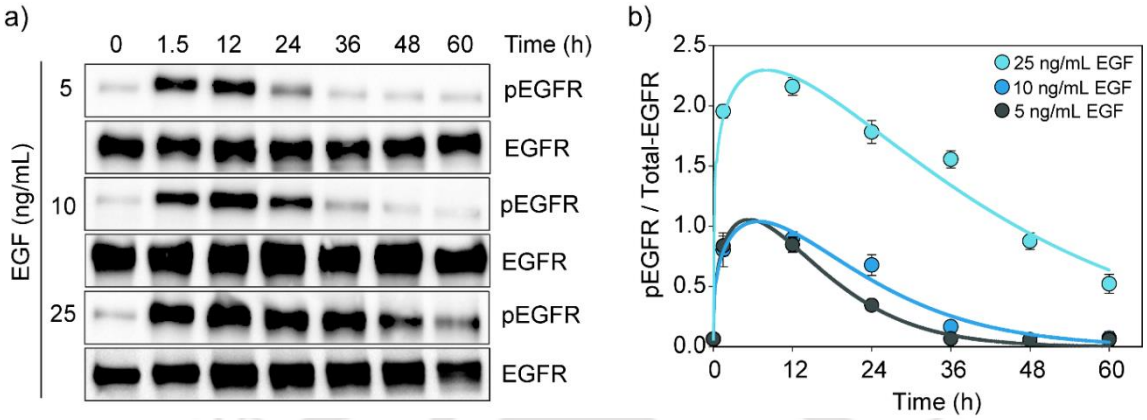



**Figure 4.24: The dose-dependent temporal dynamics of phospho-EGFR.**
Cells were treated with different doses of EGF, and the phosphorylation of EGFR was measured through western blotting. The experiment was done in triplicates, and a representative blot for each dose of EGF is shown in (a). The quantitative plot of (a) is shown in (b). Each circle represents the mean of three independent experiments, and the error bar represents standard deviation.

From the observed dynamics of phospho-EGFR and the cell state distribution, we hypothesize that the EGF signaling favors the forward transition from Cobble → Circular. The reverse transition from Circular → Cobble happens only during the decay of phospho-EGFR. The Spindle cells emerge only during the decay phase of EGF signaling and eventually become Cobble when the phospho-EGFR reaches the basal level. When there is sustained phospho-EGFR level (25 ng/mL EGF), there is no reverse transition, and the cells stay in the Circular state.

To investigate further, we treated cells with two consecutive doses of EGF. In the case of 10 ng/mL EGF treatment, the phospho-EGFR level gradually decreased after 12 h (Figure 4.25a). Therefore, to sustain the phospho-EGFR signal, we gave another pulse of 10 ng/mL EGF at 12 h. We measured both phospho-EGFR level as well the population distribution of cells (Figure 4.25). As expected, two pulses of EGF generated a prolonged phospho-EGFR signal. When the second pulse of EGF was

given, the cells remained in the Circular state for a longer duration when compared to the single pulse of EGF. These results substantiate our hypothesis that sustained phospho-EGFR favors Circular cell formation.

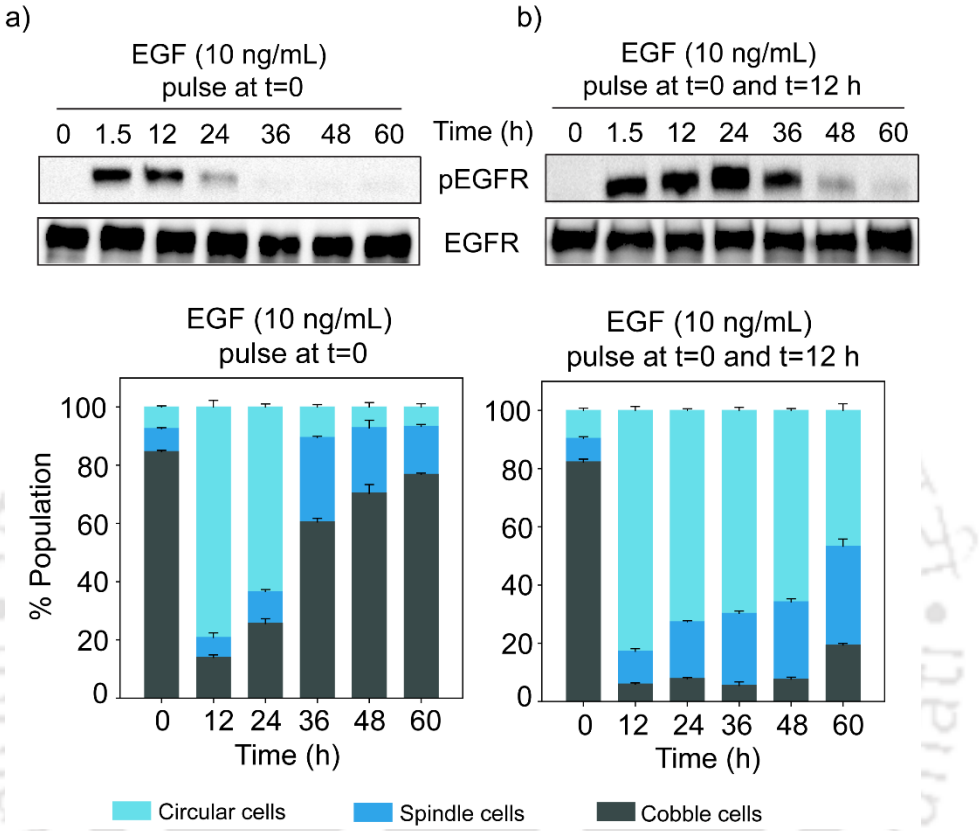



**Figure 4.25: Sustained EGF signaling favors the Circular cell state.**
Cells were treated with 10 ng/mL EGF at a) t = 0, and b) t = 0 and t = 12 h. The dynamics of phospho-EGFR was measured by western blotting (top panel). The population distribution of cell types was measured through image analysis (bottom panel). Each vertical bar represents the mean of three independent experiments, and the error bar represents standard deviation.

Further, we questioned the possibility of identifying these three different cell types based on the level of phosphorylation of EGFR so that we could connect the heterogeneity in EGFR phosphorylation to the heterogeneity in the morphology of the cells. We measured the phosphorylation status of EGFR at the single-cell level through flow cytometry. The behavior of phospho-EGFR and total-EGFR followed the same trend as observed in the western blotting (Figure 4.26 and Figure 4.27).

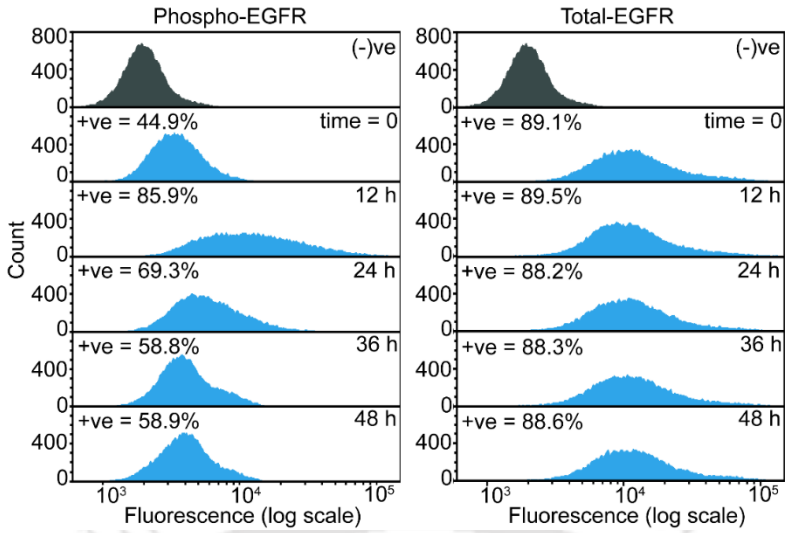



**Figure 4.26: Single-cell level measurements of phospho-EGFR and total-EGFR of 10 ng/mL EGF treated cells.**

Cells were treated with EGF, and the cells were analyzed in flow cytometry through a fluorophore-conjugated antibody. The data analysis was performed using FCS Express 5 (De Novo Software). The positive percentage population was estimated through Overton histogram subtraction. Cells stained with only secondary antibody were used as a control in histogram subtraction.

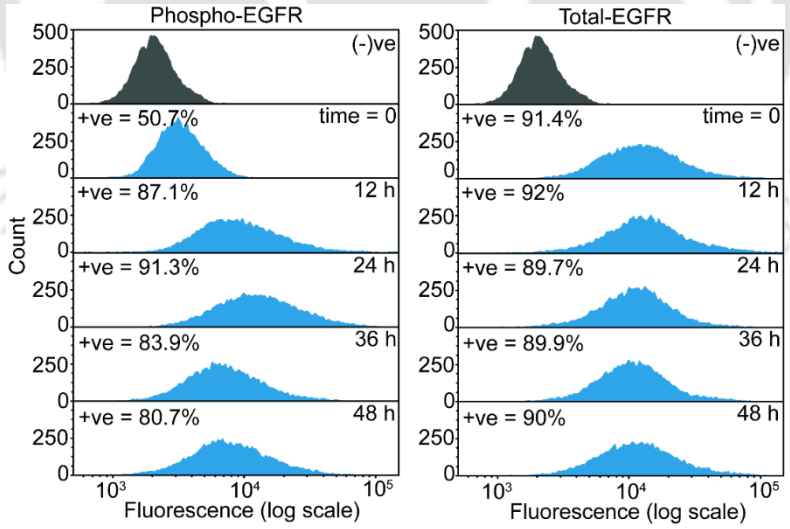



**Figure 4.27: Single-cell level measurements of phospho-EGFR and total-EGFR of 25 ng/mL EGF treated cells.**

Cells were treated with EGF, and the cells were analyzed in flow cytometry through a fluorophore-conjugated antibody. The data analysis was performed using FCS Express 5 (De Novo Software). The positive percentage population was estimated through Overton histogram subtraction. Cells stained with only secondary antibody were used as a control in histogram subtraction.

However, we could not separate the cell types based on the level of phospho-EGFR. We did not find any distinct subpopulation of phospho-EGFR. Instead, we observed a broad unimodal population of phospho-EGFR. The population of cells shifted to a high phospho-EGFR level and then declined to a low phospho-EGFR level with time in a dose-dependent fashion.

## 4.12. Adhesion signaling in cell state transition

During EMT, cells lose contact between neighboring cells and acquire migratory potential. We also observed that the Circular and Spindle cells are more scattered, and they are inherently migratory. We investigated adhesion signaling to get some insight into the migratory potential of cells and the state transition dynamics. Focal Adhesion Kinase (FAK) is a crucial molecule in adhesion signaling. Phosphorylation of FAK aids in the adhesion of cells to the extracellular matrix (*173-175*). Lu et al. (*176*) have shown that phosphorylation of EGFR induces cell migration and metastasis by dephosphorylation of FAK.

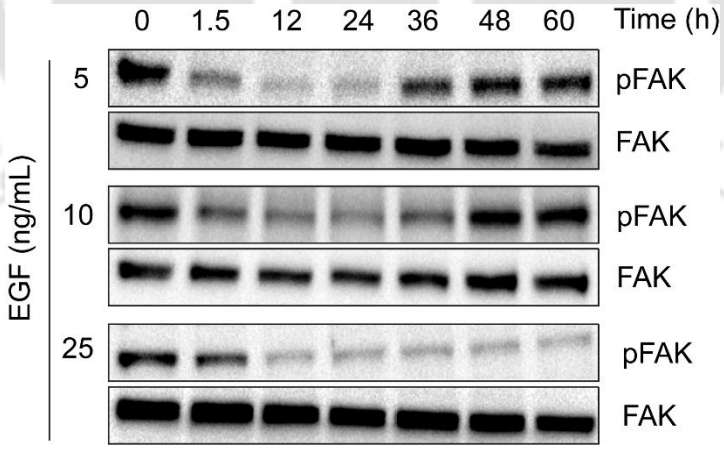



**Figure 4.28: Dose-dependent temporal dynamics of phospho-FAK.**
Cells were treated with different doses of EGF, and the dynamics of phospho-FAK were measured by western blotting.

We measured the phosphorylation status of FAK through western blotting in MDA-MB-468 cells (Figure 4.28). When there was no EGF stimulation, the phosphorylation

of FAK was very high. When cells were treated with EGF, the phospho-FAK level dropped and started rising when the EGF signal decayed. In the case of a higher dose of EGF stimulation, the phospho-FAK level remained at a lower level for a prolonged duration. Therefore, the interplay between phospho-FAK dynamics and phospho-EGFR dynamics together drives the cell state transition.

## 4.13. Ultrasensitive switch-like response in cell state transition

To substantiate the speculation that the temporal dynamics of phospho-EGFR regulates the state transition, we plotted the circular cell population against the dynamics of phospho-EGFR (Figure 4.29). The data points followed a sigmoidal trend, and we fitted it with hill function (Hill coefficient = 8.6). Figure 4.29, resembles an ultrasensitive system wherein a fractional change in the input signal triggers a drastic change in the response (*177*). Hill coefficient denotes the steepness of the sigmoidal function, and a value greater than 1 is an indicator of ultrasensitivity.

The ultrasensitive regime can be computed from the response coefficient (*178*). Response coefficient more than 1 represents ultrasensitive behavior. We computed the response coefficient from the fitted Hill function as described by Goldbeter et al. (*179*) The ultrasensitive regime is shown in Figure 4.29 (grey shaded region). In this regime, a slight shift in phosphorylation of EGFR will have a massive impact on the Circular cell population. Therefore, an ON/OFF switch determines whether a cell will be in the Circular state or not.

We further investigated this ON/OFF switch by perturbing it with Gefitinib. Gefitinib is a drug that blocks the phosphorylation of EGFR. Initially, we treated cells with a higher dose of EGF (25 ng/mL). This higher dose of EGF induced the phosphorylation of EGFR, thereby turned the switch ON. After 12 h, we added Gefitinib (0.2 μM) and turned the switch OFF. The dose of Gefitinib used in the experiments was much less than its $IC_{50}$ value (Figure 4.30).

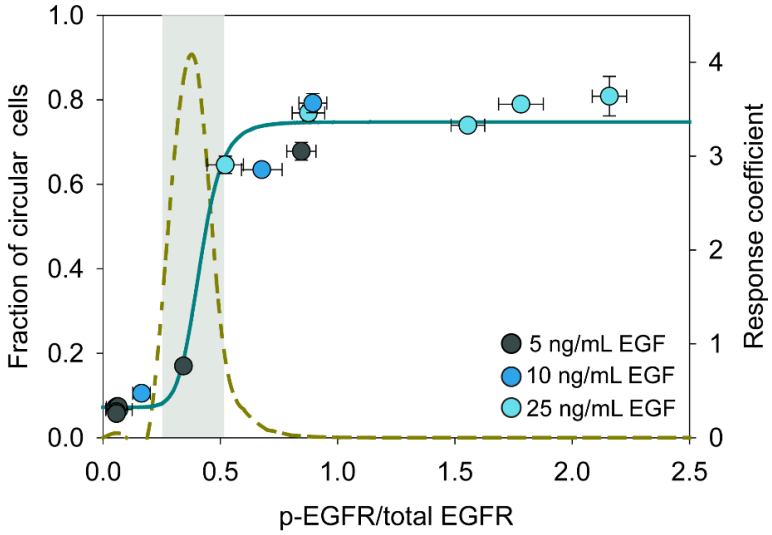



**Figure 4.29: Ultrasensitive switch-like response during EGF-induced state transition.**
The plot shows the ultrasensitive relation between the phosphorylation of EGFR and the fraction of circular cells. Normalized phospho-EGFR was measured by the densitometry of the western blots in Figure 4.24. The fraction of Circular cells was measured through image analysis. Solid circles represent the mean of three independent measurements, and the error bar indicates the standard deviation. The data were fitted to a Hill function. The dashed line indicates the response coefficient. The gray shaded portion is the ultrasensitive region (response coefficient > 1).

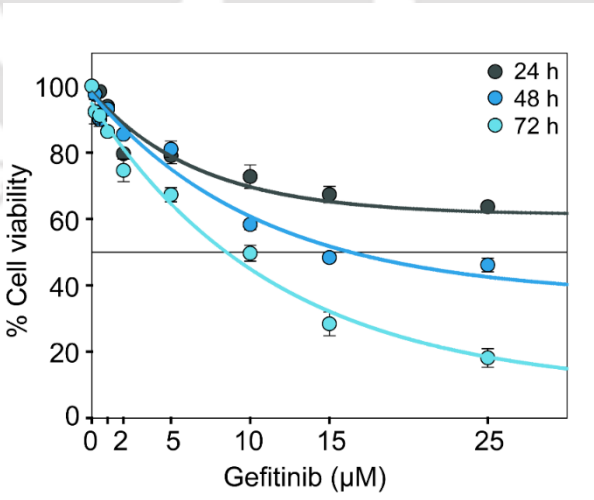



**Figure 4.30: Effect of gefitinib on cell viability.**
Cells were treated with different doses of gefitinib for varying time points, and the cell viability was measured through MTT assay. Gefitinib was dissolved in DMSO, and percentage viability was calculated relative to the cells treated with an equal amount of DMSO (without Gefitinib). The dose- and time-dependent effect of Gefitinib was statistically significant (two way ANOVA, $P < 0.01$).

On EGF treatment, cells initially became Circular as expected (Figure 4.31a – bottom panel). When Gefitinib was given at 12 h, cells started to revert from the Circular state. Almost 80 % of cells moved to Cobble state within 16 h of Gefitinib treatment (Figure 4.31b – bottom panel). The phosphorylation of EGFR dropped to the basal level immediately after Gefitinib treatment. We also measured the phosphorylation of FAK. Phospho-FAK decreased upon EGF stimulation and started to increase immediately after Gefitinib treatment (Figure 4.31b – top panel). The dynamics of phospho-FAK correlates well with the reversal of Circular cells to Cobble cells. The above experiment confirmed that there exists an ON/OFF switch that decides whether a cell will be in the circular state or not.

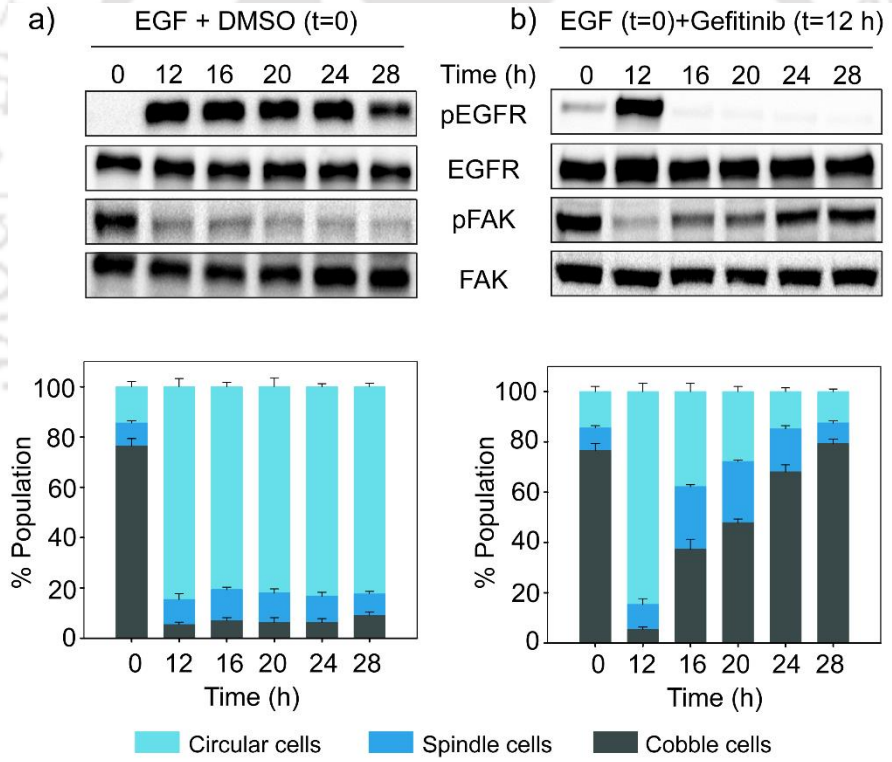



**Figure 4.31: Blockade of EGFR, turns OFF the ultrasensitive switch.**
Cells were treated with a) 25 ng/mL EGF at t = 0, b) 25 ng/mL EGF at t = 0 and 0.2 μM Gefitinib at t = 12 h. Gefitinib was dissolved in DMSO, and an equal volume of DMSO was used as vehicle control (a). The dynamics of phospho-EGFR and phospho-FAK were measured by western blotting (top panel of (a) and (b)). The population distribution of cells was measured through image analysis (bottom panel of (a) and (b)). Each bar represents the mean of three independent experiments, and the error bar represents the standard deviation.

We also checked the effect of Gefitinib on the state transition. In the case of only Gefitinib and simultaneous treatment of EGF and Gefitinib, most of the cells were Cobble as expected (Figure 4.32).

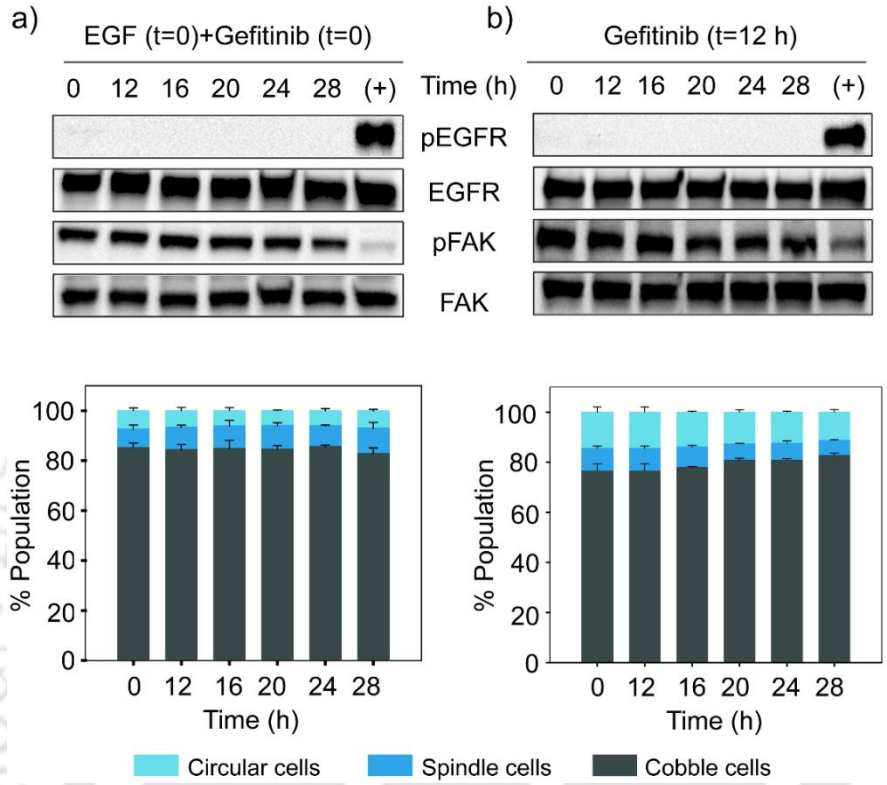



**Figure 4.32: Gefitinib does not affect the EGF-induced state transition.**
Cells were treated with a) 25 ng/mL EGF and 0.2 μM Gefitinib at t = 0, b) 0.2 μM Gefitinib at t = 12 h. The dynamics of phospho-EGFR and phospho-FAK were measured by western blotting (top panel of (a) and (b)). The population distribution of cells was measured through image analysis (bottom panel of (a) and (b)). Each bar represents the mean of three independent experiments, and the error bar represents the standard deviation.

## 4.14. Discussion

In this work, we studied the cell state transition dynamics using EGF-induced EMT of MDA-MB-468 cells as an experimental system. We discretized the phenotypic states based on the morphology of MDA-MB-468 cells. We call these phenotypic states as Cobble, Spindle, and Circular.

Classically, the states of the cells were defined based on the relative expression level of various molecular markers (*6, 31, 54*). Zhang et al. (*4*) have categorized cell types in TGF-β1-induced EMT based on the relative expression level of E-cadherin and Vimentin. Several other studies on EMT have also categorized cell states based on molecular markers (*180-182*). One of the drawbacks of the marker-based classification is that they vary across different cell types and experimental conditions (*183, 184*). We faced a similar situation in our experimental system.

The increased expression level of vimentin, SNAIL1 are the hallmarks of EMT (*11, 40*). In many of the studies on EMT, the cell types were classified based on the relative expression level of vimentin (*4*). However, in our experimental system, we did not observe any significant difference in the expression level of vimentin and SNAIL1 in the three cell types. Indeed, we observed changes between EGF-treated and untreated samples but not at individual cell-type levels (Figure 4.33). Our functional assays showed that the Circular and Spindle cells (migratory cells) are functionally different from the Cobble cells (non-migratory cells).

In this case, we could not use the conventional marker-based classification of the cells. The other way around is to categorize cells based on some functional features like change in the shape of the cells, different motility patterns of cells, variation in the scattering potential of cells. Irrespective of whichever molecules drive the state transition, these changes in functional features confer a phenotypic state to the cells. Therefore, in this work, we classified cells based on the change in morphology of the cells and explored the dynamics of EGF-induced cell state transition.

The quasi-potential landscape model well explains the phenotypic state transition phenomenon. In the potential landscape, multiple basins of attractors are separated by energy barriers. Each basin confers a phenotypic state to the cells. During the state transition, the cells are attracted to these basins. Phenotypic state transition occurs

through: a) the inherent noise in the system and b) the stimulus from an external signal. In the former case, cells stochastically jump between different states because of the inherent biochemical noise and results in a steady-state distribution of cells. In the latter case, the external signal changes the energy barrier between the basins, thereby forcing the cells to move from one state to the other. This change in the potential landscape is directional and depends on the strength and duration of the external signal.

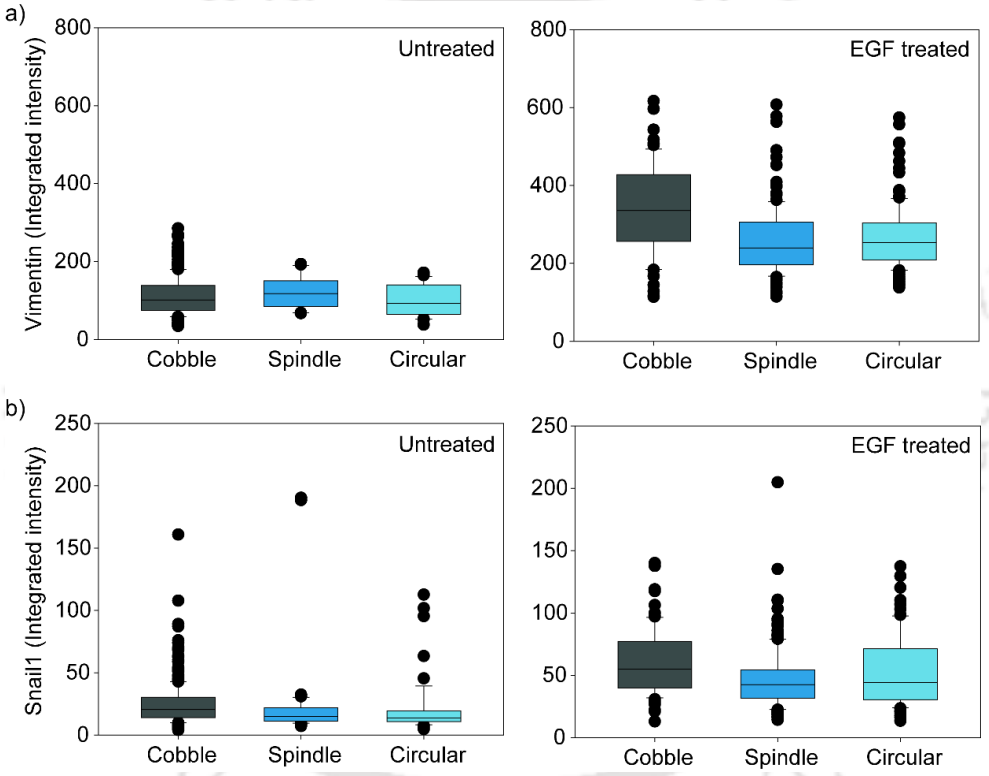



**Figure 4.33: Distribution of molecular markers of EMT in each cell type.**
Cells were grown in the presence and absence of EGF for 24 h. Cells were stained with anti-vimentin and anti-SNAIL1 antibodies conjugated with a fluorophore. The cells were imaged, and the integrated intensity of each cell was measured through image analysis. The integrated intensity is a measure of the expression level of (a) vimentin, and (b) SNAIL1.

In our experimental system, the three phenotypic states of MDA-MB-468 cells exist in a steady-state distribution in the untreated condition. When cells were treated with EGF, the steady-state distribution of cells is disturbed. Depending on the dose and the duration of EGF signaling, the distribution of cells varied. When the activation of

EGFR was short and transient, we observed a reversible population dynamics. Whereas, when there was a prolonged activation of EGFR, the cells moved from Cobble to Circular state, and the cells did not revert to the initial steady-state distribution. We had observed a similar response when we treated cells with two consecutive pulses of EGF. All these experimental observations correlate well with the quasi-potential landscape model.

In the potential landscape model, the landscape is considered as a continuous space with distinct fixed points that represents the phenotype of the cells. The landscape is defined as a function of the expression of specific molecular markers. The potential landscape model is more suited for studies, where the phenotypic states are defined based on molecular markers. In the case of the gene regulatory networks, the continuous potential landscape can be drawn using ODEs, and the state transition dynamics can be studied. However, it is difficult to construct the potential landscape when the phenotypic states are defined based on functional features. For example, in our study, we defined the phenotypic states based on the morphology of the cells. Here each phenotypic state is discrete, and the continuous potential landscape cannot be constructed

Here, we propose an alternative method to study the state transition of cells, where the phenotypic states are discrete. Each discrete phenotypic state of the cell can be considered as a discrete energy level. The Boltzmann distribution gives the relation between an energy level and the probability of staying at that energy level.

$$p_j = \frac{e^{-U_j}}{\sum_{all j} e^{-U_j}} = \frac{e^{-U_j}}{Z}$$

where $U_j$ is the energy of the $j^{th}$ state, and the $p_j$ is the probability of staying in that state. $Z$ is the normalizing constant.

From the above equation, the steady-state probability of a cell being in a particular state can be related to the energy of that state as follows,

$$U \propto -\ln(p)$$

here, $U$ is the dimensionless potential, and $p$ is the steady-state probability (*71, 72, 185*). This relation reiterates that the lowest energy states have the highest occupancy.

In our experimental system, in the absence of any external stimulus, the three phenotypic states of MDA-MB-468 cells exist in a steady-state distribution (Cobble: Spindle: Circular = 0.79: 0.13: 0.08). Therefore, we can define the potential of each discrete states as, $U_i = -\ln(f_i)$. Here, $U_i$ is the potential of the $i^{th}$ phenotypic state and $f_i$ is the fraction of cells in the $i^{th}$ state. We calculated the potential of each phenotypic state and arranged them vertically based on their potential values (horizontal lines in Figure 4.34). This way of representation resembles the Jablonski diagram that describes the molecular electronic states in spectroscopy.

The treatment of cells with EGF instantaneously activates EGFR, thereby pushing the cells from the Cobble state (lowest potential state) to the Circular state (highest potential state) (vertical green arrow in Figure 4.34). The decay of EGF signaling is a gradual process, and therefore, the relaxation of cells is also gradual. During the relaxation, the cells from the Circular state returns to the Spindle state (next highest potential state) and finally returns to the Cobble state (red vertical arrows in Figure 4.34). This transition of cells is equivalent to the transition of electrons from the ground state to the excited state, followed by relaxation through the metastable state in molecular spectroscopy.

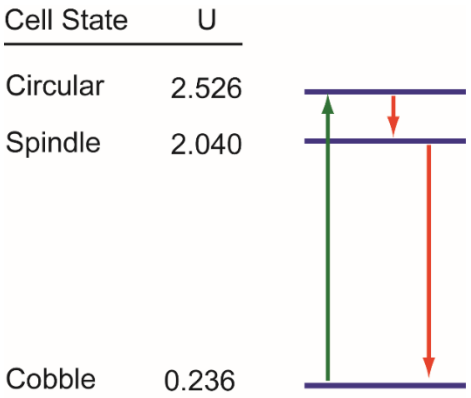



**Figure 4.34: The discrete energy-based diagram of cell state transition.**
Each phenotypic state of the cell corresponds to a discrete energy level. The horizontal blue lines represent the energy level. The potential of each state is calculated from the steady-state distribution of cells. The vertical arrows indicate the dominant transitions of cells when treated with 10 ng/mL EGF.

When there is no EGF stimulation, the cells could not move to the Circular state. Whereas when cells were treated with moderate doses of EGF, most of the cells moved to the Circular state. As the EGF signal decays, the cells returned to the Cobble state. At a higher dose of EGF, we did not observe any such behavior. The cell stayed in the higher potential state till 60 h. Thus, the dose of EGF controls the probabilistic transition of cells from a lower potential state to the higher potential state and the speed of relaxation of the cells.

The discrete energy state model has certain advantages over the conventional potential landscape model. In the potential landscape model, the phenotypic states are fixed points in the landscape. Nevertheless, in experimental biology, the phenotypic states are not defined by a unique value of the expression of genes rather by a range of expression of genes. This issue can be sorted if the phenotypes are defined as discrete states. Then, using our formulation, the potential of each state can be estimated from the steady-state data, and the state transition diagram can be constructed. Our approach is much more straightforward and susceptible to stochastic modeling to study the cell state transition.

An interesting observation in the population dynamics of MDA-MB-cells is the evolution of Spindle cells. Spindle cells were seen in considerable numbers only during the reversible population dynamics. They predominantly evolved during the decay phase of EGF signaling, and the transition path of the cells was Circular → Spindle → Cobble. Whereas during the active phase of EGF signaling, the transition path was Cobble → Circular. Therefore, the changes in the potential landscape during the active phase of EGF signaling and the decay phase of EGF signaling are different.

The transition of cells follows different paths during the rise and the decay of EGF signaling. This typical behavior of cell state transition resembles hysteresis. In hysteresis, the response of the system responds differently to the increase and decrease of the input stimulus. Hysteresis has been reported in various biological processes like cell division, cell differentiation, apoptosis (*186-188*). Terrassa et al. (*189*) have reported hysteresis in TGF-β1-induced EMT. Our observation suggests the existence of cellular memory in EGF-induced cell state transition. The system keeps track of the forward transition path from the Cobble → Circular state, and therefore, during the reversal, it returns to the Cobble state through an intermediate Spindle state.

We observed an ultrasensitive response between the phosphorylation status of EGFR and the circular cell population. The ultrasensitive switch helps cells in binary decision making. Mitogen-activated protein kinase (MAPK) is one of the significant canonical pathways activated by EGF, and the signaling cascade in MAPK is reported to exhibit ultrasensitive switch-like behavior (*190*). Melen et al. (*191*) have observed an ultrasensitive response triggered by EGFR during the embryogenesis of drosophila triggered by the activation of EGFR.

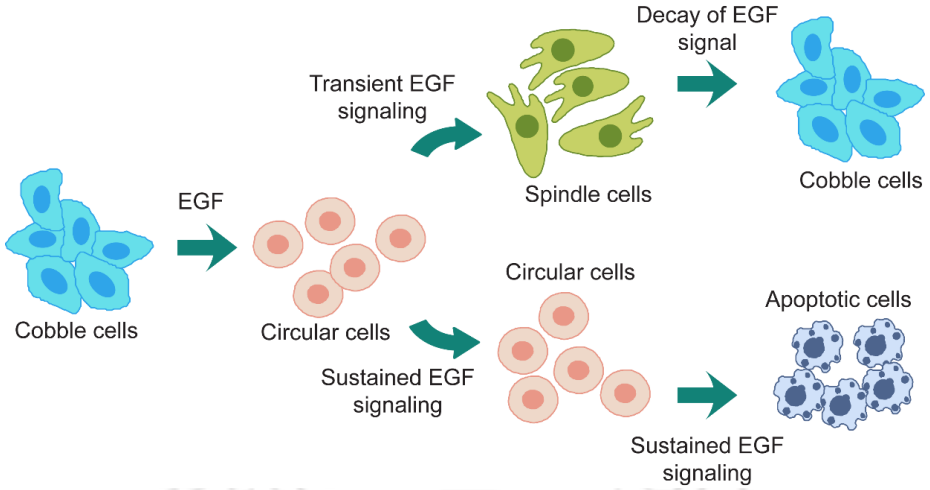



**Figure 4.35: A possible hypothesis of the cell fate decision during the EGF-induced state transition.**

We observed that the EGF signaling dephosphorylates FAK and thereby negatively regulates cell adhesion signaling. Based on our observations, we hypothesize the molecular mechanism of the EGF-induced state transition of MDA-MB-468 cells (Figure 4.35). MDA-MB-468 cells grow as a monolayer and form close cell-to-cell contact with a cobble-like appearance. When the external signal EGF is administered, EGF signaling deactivates adhesion signaling, and therefore, the cells lose contact between the neighboring cells and become circular. When the EGF signaling starts decaying, phospho-FAK slowly increases, thereby helping in establishing contacts between neighbors, and the cells become spindle. Eventually, when the EGF signaling reaches the basal level, the cells form well-established contact between neighbors and become adherent. In the case of prolonged EGF signaling, the cells could not restore adhesion signaling and therefore stays as circular cells and eventually die. A similar dual role of EGF is reported in A431 cells that express many EGF receptors like MDA-MB-468 cells. In this cell line, a lower dose of EGF stimulates cell proliferation, whereas a high dose of EGF induces cell death (*151*, *152*).

CHAPTER 5

# The Effect of Background Noise on EGF-induced Epithelial-Mesenchymal Transition

## 5.1. Introduction

The cells perform various specific cellular functions like cell division, differentiation, and cell death based on the instructions from the external cue. However, the cells are exposed to a dynamic environment, with multiple signaling molecules activating several other signaling pathways. These signaling pathways in cells are highly interconnected. Therefore, there is always a possibility of signal interference across several pathways, creating a background chatter of signals. The critical question is

whether a cell can decode the information from the input signal and perform the corresponding task in the presence of background noise.

In this chapter, we investigated how the background noise regulates signal transduction in cells. In chapter 4, we investigated the EGF-induced state transition of MDA-MB-468 cells. Here, we used the same experimental system to study the effect of background noise on cell state transition. To introduce background noise, we used a suboptimal dose of TGF-β1 in the presence of EGF. Using the concepts of information theory, we estimated the signal transmission capacity of the EGF signaling network with respect to the phenotypic states of MDA-MB-468 cells in the presence and absence of the background noise.

## 5.2. TGF-β1 modulated the EGF-induced state transition

TGF-β1 is a potent inducer of EMT in several cell lines (*4, 40, 47, 192, 193*). Firstly, we checked the effect of TGF-β1 on the phenotypic states of MDA-MB-468 cells. MDA-MB-468 cells were treated with 5 ng/mL of TGF-β1, and the cells were stained with HCS CellMask Red Stain and imaged using a fluorescence microscope. Through image analysis, we quantified the population distribution of the morphological states of MDA-MB-468 cells. We did not observe any significant difference in the population distribution of MDA-MB-468 cells in the presence and absence of TGF-β1 (Figure 5.1). 5 ng/mL of TGF- β1 did not induce any cell state transition of MDA-MB-468 cells. Therefore, we used this suboptimal dose of TGF- β1 to create background noise in the EGF-induced state transition of MDA-MB-468 cells.

We then treated cells with different doses of EGF and introduced background noise by adding TGF-β1. Here, EGF acts as the primary signal, and TGF-β1 is the background signal. We observed changes in the population distribution of cells when treated with only EGF and in the presence of both EGF and TGF-β1. There was an increase in the population of Spindle and Cobble cells when TGF-β1 was

supplemented with EGF (Figure 5.2). Though TGF-β1 did not induce any change in the population distribution of MDA-MB-468 cells (Figure 5.1), it modulated the EGF-induced state transition of MDA-MB-468 cells, when supplemented with EGF (Figure 5.2).

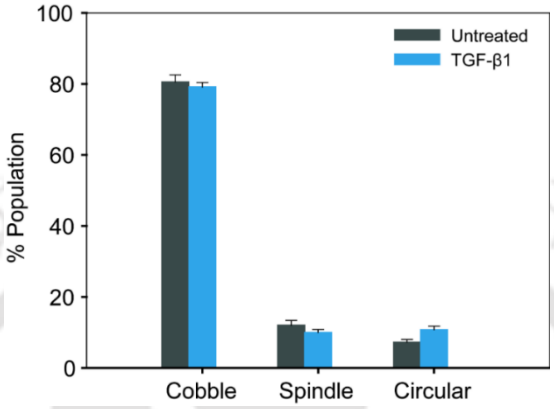



**Figure 5.1: TGF-β1 did not induce any change in the population distribution of MDA-MB-468 cells.**
Cells were treated with TGF-β1 (5 ng/mL) for 24 h. The population distribution of the three phenotypic states of MDA-MB-468 cells was quantified through image analysis. Each bar represents the mean of three independent experiments, and the error represents the standard deviation. There was no statistically significant difference between the distribution of the three phenotypic states of MDA-MB-468 cells in the presence and absence of TGF-β1 (Chi-square test, $P = 0.299$).

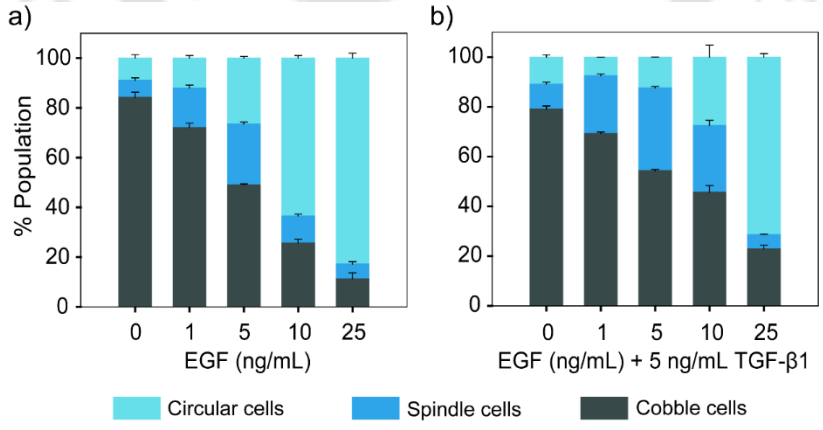



**Figure 5.2: TGF-β1 modulated the EGF-induced cell state transition.**
Cells were treated with different doses of EGF in the presence and absence of 5 ng/mL TGF-β1 for 24 h. The distribution of different phenotypic states of MDA-MB-468 cells was quantified through image analysis. Each bar represents the mean of three independent experiments, and the error bar indicates the standard deviation.

## 5.3. TGF-β1 exerts a positive synergistic effect on Spindle and Cobble cells

To understand whether the effect of TGF-β1 on EGF-induced state transition is additive or synergistic, we define a parameter Z for each cell state. Z is calculated based on the population distribution of the phenotypic states of MDA-MB-468 cells.

$$Z_{i,j} = \frac{a_{i,j} - \hat{a}_i}{\sigma_i} \tag{5.1}$$

where $i$ = Cobble, Spindle, or Circular. $j$ refers to the experimental conditions like cells treated with different doses of EGF or TGF-β1 or EGF + TGF-β1. $a_{i,j}$ is the percentage population of $i^{th}$ cell type at $j^{th}$ treatment condition. $\hat{a}_i$ is the average of the percentage population of $i^{th}$ cell type at the untreated condition. $\sigma_i$ is the standard deviation of the percentage population of $i^{th}$ cell type at the untreated condition.

The $\Delta Z$ score of the $i^{th}$ cell type is defined as,

$$\Delta Z_i = Z_{i,\,EGF+TGF-\beta1} - (Z_{i,\,EGF} + Z_{i,\,TGF-\beta1}) \tag{5.2}$$

$\Delta Z$ is a dimensionless metric that explains the deviation between the cell population treated with EGF + TGF-β1 and the cell population treated with EGF and TGF-β1 separately. A positive $\Delta Z$ score represents positive synergism, whereas a negative $\Delta Z$ score represents negative synergism. If $\Delta Z$ score = 0, then TGF-β1 exerts only an additive effect on EGF-induced state transition of MDA-MB-468 cells.

The $\Delta Z$ score analysis of our experiment is shown in Figure 5.3. For all doses of EGF, TGF-β1 had a positive synergistic effect on Cobble cells and a negative synergistic effect on Circular cells. Whereas, TGF-β1 showed both positive and negative synergism on Spindle cells depending on the dose of EGF. At a moderate dose of EGF, TGF-β1 had a positive synergistic effect, and at a higher dose of EGF, TGF-β1 had a

negative synergistic effect on Spindle cells. Therefore, the background noise (TGF-β1) does not allow the cells to stay in the Circular state. Instead, it favors the transition to Spindle and Cobble cells.

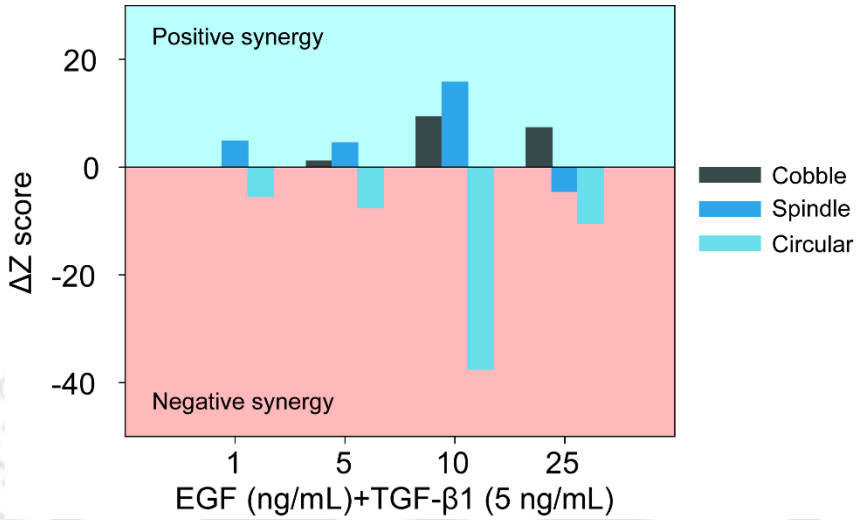



**Figure 5.3: TGF-β1 promotes the evolution of Spindle and Cobble cells.**
Cells were treated with different doses of EGF in the presence and absence of TGF-β1 for 24 h. The population distribution of cells was quantified through image analysis. Δ*Z* score was calculated for each cell type based on the population distribution data. The pale blue shaded region represents positive synergism, and the pale red shaded region represents negative synergism.

In chapter 4, we traced the evolutionary path of MDA-MB-468 cells during EGF-induced state transition. We showed that, at a moderate dose of EGF treatment, the state transition path follows Cobble → Circular → Spindle → Cobble. The Δ*Z* score for moderate doses of EGF shows that TGF-β1 promotes the Spindle and Cobble cell formation than Circular cells. From these observations, we hypothesize that TGF-β1 potentiates the reverse transition from Circular → Cobble state.

Also, in chapter 4, we showed that at a higher dose of EGF treatment, the state transition path follows Cobble → Circular state, and there is no reverse transition. Through cell cycle analysis, we observed an EGF dose-dependent increase in the sub G0/G1 population (Figure 5.4). The cells in the sub G0/G1 phase of the cell cycle are

the apoptotic cells. A higher dose of EGF is reported to induce cell death in MDA-MB-468 cells (*153-155*). Therefore, at a higher dose of EGF treatment, the cells move to the Circular state and eventually die. Also, the $\Delta Z$ score analysis shows that at a higher dose of EGF treatment, TGF-β1 exerts a negative synergism on Circular cells. Thus, TGF-β1 pushes the cells from the Circular state to Spindle and Cobble state and prevents Circular cells from undergoing apoptosis. Therefore, TGF-β1 potentiates EGF-induced reversible state transition by reducing the flux flow through the apoptotic pathway.

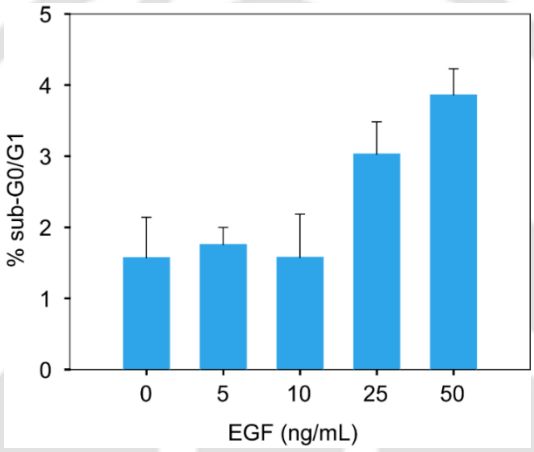



**Figure 5.4: EGF-induced apoptosis in MDA-MB-468 cells.**
Cells were treated with different doses of EGF for 48 h. The cells were stained with propidium iodide and were analyzed through flow cytometry. The sub-G0/G1 population was calculated by fitting the data to the DNA multicycle program in FCS Express. Each bar represents the average of three independent experiments, and the error bar represents the standard deviation. Higher doses of EGF (25 and 50 ng/mL) treatment showed a significant increase in the sub-G0/G1 population compared to the untreated cells (Kruskal-Wallis test, $P < 0.001$).

## 5.4. Information-theoretic analysis

We investigated the effect of background noise on signal transduction in cells using mutual information. The basics of information theory and its application in cell signaling are discussed in section 2.8 in chapter 2. In our analysis, we used mutual

information (MI) as a measure of statistical dependency between the input signal (EGF) and the output response (morphological cell state).

### 5.4.1. Estimation of mutual information from experimental data

Suderman et al. (*111*) have used a contingency table approach to estimate mutual information from experimental data. This section explains the method for contingency Table 5.1.

In Table 5.1, $x_1, x_2, \ldots x_m$ are the discrete values of the input measurement ($X$). $y_1, y_2, \ldots y_n$ are the discrete values of the output measurement ($Y$). $k_{i,j}$ are the number of $y_j$ responses for $x_i$ input condition. The total number of observed responses, $T = \sum_{i,j} k_{i,j}$; where $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$. From the contingency table, we can compute the following probabilities.

The joint probability, $p(x_i, y_j) = \frac{k_{i,j}}{T}$. where $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$.

The marginal probability of the input signal, $p(x_i) = \sum_{j=1}^{j=n} \frac{k_{i,j}}{T}$. where $i = 1, 2, \ldots, m$.

The marginal probability of the response, $p(y_j) = \sum_{i=1}^{i=m} \frac{k_{i,j}}{T}$. where $j = 1, 2, \ldots, n$.

The mutual information can be computed using the above probability values in the below equation.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \times \log_2 \frac{p(x,y)}{p(x) \times p(y)} \tag{5.3}$$

**Table 5.1. Contingency table to calculate mutual information.**

Each row represents the discretized input signal, and each column represents the discretized output response. Each entry in the table denotes the number of responses for the corresponding input signal.

| | $y_1$ | $y_2$ | ... | $y_n$ |
|---|---|---|---|---|
| $x_1$ | $k_{1,1}$ | $k_{1,2}$ | ... | $k_{1,n}$ |
| $x_2$ | $k_{2,1}$ | $k_{2,2}$ | ... | $k_{2,n}$ |
| ... | ... | ... | ... | |
| $x_m$ | $k_{m,1}$ | $k_{m,2}$ | ... | $k_{m,n}$ |

## 5.5. EGF-induced cell state transition is noisy

### 5.5.1. Mutual information of EGF-induced cell state transition

In our experimental system, the primary input signal is EGF. We introduced background noise by adding a suboptimal dose of TGF-β1, and we measured the population distribution of MDA-MB-468 cells as the response. We discretized the population of MDA-MB-468 cells in terms of the morphology of the cells (Cobble, Spindle, and Circular). Using the concepts of mutual information and entropy, we estimated the signal transduction in the population of MDA-MB-468 cells with respect to the discrete phenotypic states of the cells.

Firstly, we estimated the entropy and mutual information when cells were treated only with EGF. Here, we discretized the input signal in terms of the concentration of EGF. We treated cells with different doses of EGF (1, 5, 10, 25 ng/mL) and measured the population distribution of MDA-MB-468 cells at different time points (0, 12, 24, 36, 48, 60 h). We measured the percentage population of each phenotypic state of MDA-MB-468 cells using image analysis, and the data were discussed in chapter 4. The percentage population of each cell type represents the population behavior of MDA-MB-468 cells. Therefore, we use the percentage of each cell type as the response

variable instead of the actual cell number. The input signal, $X$ = 0, 1, 5, 10, 25 ng/mL of EGF and the response variable, $Y$ = % Cobble, % Spindle and % Circular.

We estimated the entropy and the mutual information for each time point. A representative calculation of $p(x,y), p(x)$, and $p(y)$ from the contingency table (Table 5.2) for EGF treated cells at time = 24 h is shown here (Table 5.3 - Table 5.5).

**Table 5.2. Contingency table for 24 h of EGF treated cells.**

The entries in the contingency table are the percentage of each cell type estimated through image analysis.

| *Input Signal (X)* *EGF(ng/mL)* | *Response (Y)* | | |
|---|---|---|---|
| | *Cobble* | *Spindle* | *Circular* |
| 0 | 79.77 | 12.68 | 7.55 |
| 1 | 69.95 | 20.19 | 9.86 |
| 5 | 52.44 | 30.59 | 16.97 |
| 10 | 25.65 | 10.9 | 63.45 |
| 25 | 7.35 | 13.7 | 78.95 |

From the computed marginal probability, we estimated the entropy of the input signal and the response of the EGF-induced state transition of MDA-MB-468 cells (Figure 5.5). The theoretical maximum entropy of the input and the response were 2.32 bits and 1.5 bits, respectively. As shown in Figure 5.5, the estimated entropy for the response was maximum at 24 h and was close to the theoretical limit.

As discussed in section 2.8, the upper bound of the mutual information is constrained by the $\min\big(H(X), H(Y)\big)$. Therefore, the theoretically maximum possible mutual information of EGF-induced state transition is 1.5 bits.

**Table 5.3. The joint probability of the input signal and the response for 24 h of EGF treated cells.**

| *Input Signal (X) EGF(ng/mL)* | *Response (Y)* | | |
|---|---|---|---|
| | *Cobble* | *Spindle* | *Circular* |
| 0 | 0.16 | 0.025 | 0.015 |
| 1 | 0.14 | 0.04 | 0.02 |
| 5 | 0.105 | 0.061 | 0.034 |
| 10 | 0.051 | 0.022 | 0.127 |
| 25 | 0.015 | 0.027 | 0.158 |

**Table 5.4. Marginal probability of the input signal for 24 h of EGF treated cells.**

| *The input signal (X) – EGF (ng/mL)* | | | | |
|---|---|---|---|---|
| ***0*** | ***1*** | ***5*** | ***10*** | ***25*** |
| 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |

**Table 5.5. Marginal probability of the response for 24 h of EGF treated cells.**

| *Response (Y)* | | |
|---|---|---|
| *Cobble* | *Spindle* | *Circular* |
| 0.47 | 0.176 | 0.354 |

We calculated the mutual information of EGF-induced cell state transition using equation (5.3). The maximum mutual information is 0.35 bits, which is less than the theoretical maximum possible mutual information (Figure 5.6).

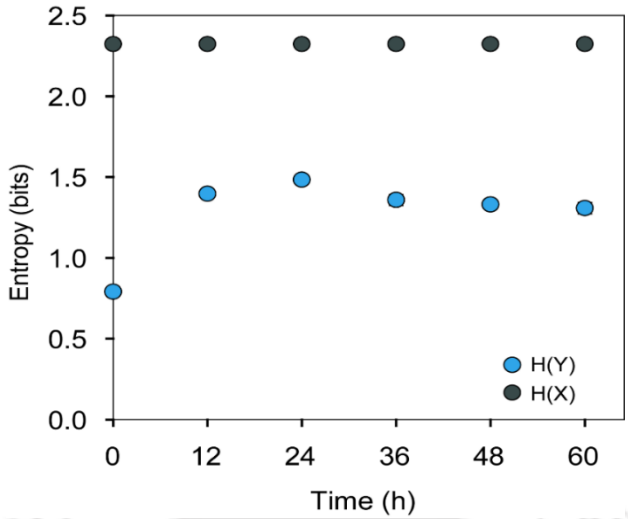



**Figure 5.5: Entropy of the input signal and the response of the EGF-induced state transition.**
Cells were treated with different doses of EGF for varying time points. The distribution of the phenotypic states of MDA-MB-468 cells was quantified through image analysis. $H(X)$ and $H(Y)$ are the entropy of the input and the response, respectively. $H(X)$ and $H(Y)$ were calculated from these data by the contingency table method.

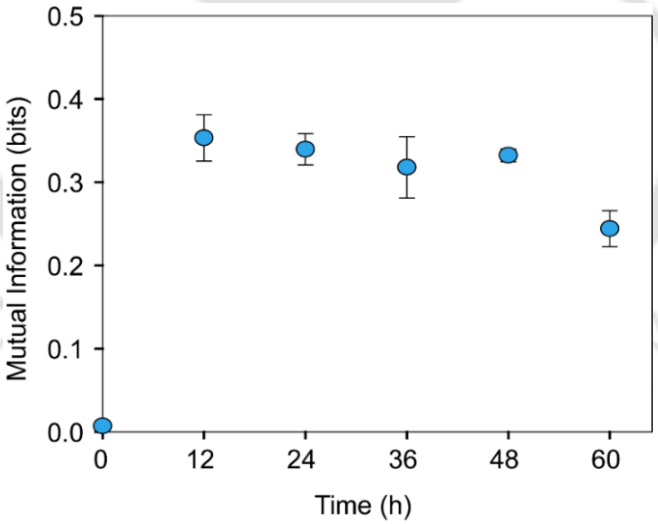



**Figure 5.6: The mutual information of the EGF-induced state transition.**
Cells were treated with different doses of EGF for varying time points. Three biological replicates were used in this experiment. The distribution of the phenotypic states of MDA-MB-468 cells was quantified through image analysis. Mutual information was calculated from these data through the contingency table method. The solid circles represent the mean of the estimated mutual information of the three biological replicates, and the error bar represents the standard deviation.

### 5.5.2. The channel capacity of EGF-induced state transition

The maximum possible information that can be propagated through a signaling pathway is defined as the channel capacity of that pathway (*194*). In our experimental system, the input signal is uniformly distributed, i.e., each dose of the input signal is equiprobable as we have performed the same number of experiments for each dose of EGF (Table 5.4). Such uniform distribution may not be valid in a real cellular system. Moreover, we do not have any information on the distribution of EGF concentration in a real cellular system. However, MI depends upon the probability distribution of the input signal. Therefore, we estimated the channel capacity by maximizing the MI over all possible probability distribution of the input (*111*, *112*).

$$C = \sup_{p(x)} I(X;Y) \tag{5.4}$$

where the supremum is taken over all possible $p(x)$.

Optimizing the input signal distribution is equivalent to optimizing the $p(x)$. Therefore, we calculated the channel capacity by optimizing the marginal probability of the input signal. Maximizing a variable is equivalent to minimizing the negative log of the variable. The optimization was performed by minimizing the objective function, $\min_{p(x)}[\log_2 I(X;Y)]$. The optimization was done in MATLAB using the interior-point algorithm (*195*) with the constrained non-linear solver, *fmincon* subject to the following equality and inequality constraints.

$$a)\ 0 \le p(x_i) \le 1$$

$$b)\ \sum_{i=1}^{m} p(x_i) = 1$$

where $i = 1, 2, \ldots, m$.

The optimization was performed as follows. $p(x)$ was varied randomly, such that the constraints mentioned above were satisfied. From this new marginal probability, a new contingency table was created such that the original population distribution of cells is preserved.

$$k'_{i,j} = \frac{p(x_i)'}{p(x_i)} \times k_{i,j}$$

where $i = 1, 2, \ldots, m; j = 1, 2, \ldots, n$. $k'_{i,j}$ is the element in the $i^{th}$ row and $j^{th}$ column of the new contingency table. $k_{i,j}$ is the element in the $i^{th}$ row and $j^{th}$ column of the existing contingency table. $p(x_i)'$ is the new marginal probability of the input signal, generated through a random number. $p(x)$ is the marginal probability calculated from the existing contingency table.

From the new contingency table, the mutual information was estimated, and the entire procedure was repeated until convergence. The mutual information calculated from the converged $p(x)$ is the channel capacity of the signaling pathway.

The estimated channel capacity is shown in Figure 5.7. Both the mutual information and the channel capacity of EGF-induced state transition are much less than the theoretical maximum possible mutual information. The maximum channel capacity was 0.5 bits and remained close to 0.5 bits for a long duration of 12-48 h. Both MI and channel capacity of EGF-induced EMT in this cellular system is lower than the theoretical limit. This shows that the signal transduction of the EGF-induced cell state transition is very noisy.

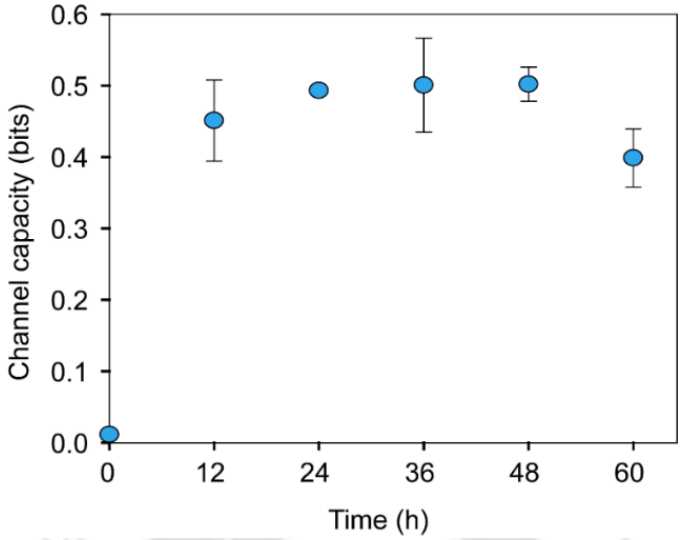



**Figure 5.7: The channel capacity of EGF-induced state transition.**
Cells were treated with different doses of EGF for varying time points. Three biological replicates were used in this experiment. The distribution of the phenotypic states of MDA-MB-468 cells was quantified through image analysis. From these data, the channel capacity was calculated by maximizing the mutual information for different input signal distribution. The solid circles represent the mean of the estimated channel capacity of the three biological replicates, and the error bar represents the standard deviation.

## 5.6. TGF-β1 amplifies the noise in EGF-induced state transition

The population distribution of the phenotypic states of MDA-MB-468 cells in the presence of TGF-β1 is shown in Figure 5.2b. From this data, we estimated the mutual information and the channel capacity of EGF-induced state transition in the presence of TGF-β1 (Figure 5.8). As the channel capacity for EGF reached the maximum by 24 h, we compared the mutual information and channel capacity of 24 h in Figure 5.8. TGF-β1 reduced mutual information and, the channel capacity followed the same pattern. The decrease in MI and channel capacity indicates that interference from TGF-β1 signaling increases the noise in EGF-induced state transition and reduces the correlation between EGF-dose and cell states.

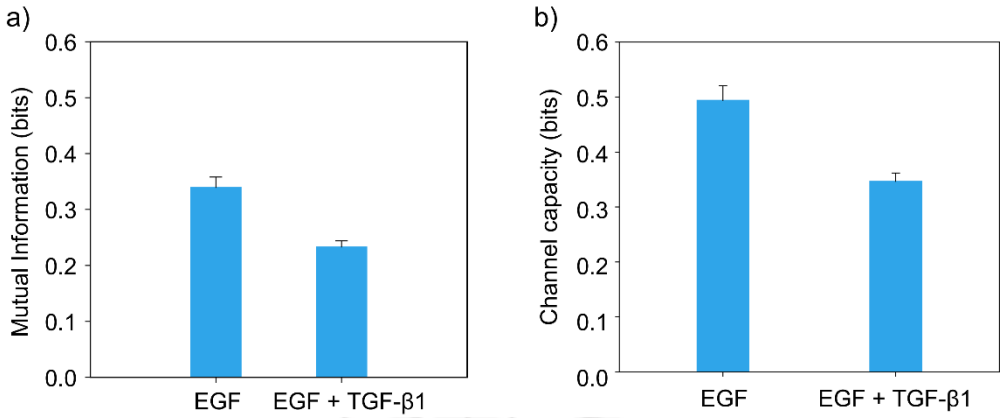



**Figure 5.8: Mutual information and channel capacity of EGF-induced state transition in the presence of TGF-β1.**
Cells were treated with different doses of EGF in the presence and absence of TGF-β1 (5 ng/mL) for 24 h. Three biological replicates were used in this experiment. The distribution of the phenotypic states of MDA-MB-468 cells was quantified through image analysis. Mutual information (a) and channel capacity (b) was calculated from these data based on the contingency table method. The vertical bars represent the mean of the three biological replicates, and the error bar represents the standard deviation. The change in (a) mutual information and (b) channel capacity in the presence and absence of TGF-β1 is statistically significant (t-test, $p = 0.001$).

## 5.7. Discussion

In this chapter, we studied how the background signal modulates the dynamics of cell state transition using EGF-induced EMT of MDA-MB-468 cells as the experimental system. We introduced background noise by adding TGF-β1 externally. Here, we explored the signal transduction of EGF signaling with respect to the phenotypic states of MDA-MB-468 cells in the presence of TGF-β1.

The information-theoretic analysis showed that EGF-induced EMT of MDA-MB-468 cells is very noisy, and TGF-β1 reduced the statistical correlation between the dose of EGF and the cell state distribution. Our observation of low MI for EGF-induced EMT is not unusual as cellular processes are noisy. Moreover, low MI and channel capacity have been reported in many experimental systems. Interestingly, Suderman et al. (*111*) had shown that in the apoptotic pathway, channel capacity is higher when one considers population-level behavior (% of cell death) in comparison to cellular level

signaling. In our work, we did not measure the channel capacity of individual signaling pathways involved in EGF signaling. Instead, we have connected the input signal (EGF) with the end response of the cell (change in morphology). Therefore, we have estimated MI and channel capacity of population-level behavior. Even then, the channel capacity was only ~30% of the theoretical maximum possible mutual information, and TGF- β1 further reduced the channel capacity.

Several authors have noted that a lower estimate of MI in biological experiments may stem from the improper choice of input signals in experiments and lack of precision in measurements (*115, 125, 127*). Both could be reasons for low MI reported in this work. An increase in the number of doses of EGF used, and the spread of the dose range may improve the estimation of MI and channel capacity. Further, we categorized cells into just three broad morphological categories. Though such categorization was performed using a machine learning tool, there is a possibility of further subcategories of cell states. Such subcategories would increase the number of possible output states and may improve the precision of our experiment.

One can use $\Delta Z$ score and MI data together to understand the change in state transition dynamics in cells co-treated with TGF-β1. We can consider cellular state transition as a noisy stochastic process where EGF signaling modulates specific state transition probabilities. Possibly, co-treatment with TGF-β1 increases noise and affects effective state transition probabilities. We observed that cells staying in Circular state for a prolonged time, eventually end up in apoptosis. Possibly TGF-β1 helps more cells to transit from Circular to Spindle or Cobble states, thereby preventing the cells from apoptosis. This change is reflected in the $\Delta Z$ score analysis. Further analysis using many doses of EGF and TGF-β1 at different time points may improve the information-theoretic study of this cellular system.

The possible molecular level explanation might be the interference between EGF and FAK signaling. In chapter 4, we showed that the decline in the phosphorylation of EGFR increases the phosphorylation of FAK, which in turn promotes the reverse transition of cells from Circular to Spindle state. Thannickal et al. (*196*) have shown that TGF-β1 promotes cell adhesion through the phosphorylation of FAK. Therefore, most probably, TGF-β1 promotes the Spindle cell formation by increasing the interference from FAK signaling. However, a further molecular-level investigation is required to confirm the hypothesis.

CHAPTER 6

# DEBay: A tool for estimation of cell-type-specific gene expression from quantitative PCR of ensemble of cells

## 6.1. Introduction

Quantitative PCR (qPCR) is a widely used technique to measure gene expression from an ensemble of cells. Cell-to-cell heterogeneity in gene expression is observed within a population of cells. A population of cells may contain several different cell types or subpopulations with distinct gene expression patterns. For example, the metastatic cancer cell population contains several cell types like Epithelial (E), Mesenchymal (M),

and Hybrid (H) with distinct gene expression patterns (*12, 197, 198*). However, as qPCR measures gene expression from an ensemble of cells, the gene expression of different cell types present in the population is obscured.

We developed a computational tool DEBay, that estimates cell-type-specific gene expression from qPCR, given the proportion of different cell types in a population. The algorithm is implemented in Python. We created an easy-to-use GUI of DEBay and is accessible as a Windows installer. The installer file and the instructions to use the tool are available in SourceForge (https://sourceforge.net/projects/debay/). DEBay would be more advantageous in experiments where the isolation of pure cell types is cumbersome. However, the proportions of various cell types can be measured through techniques like flow cytometry, quantitative image analysis, Coulter counter. This data can be used in DEBay to estimate the cell-type-specific gene expression.

## 6.2. The deconvolution algorithm

Let us consider a population of cells containing different cell types or subpopulations. The experiment is performed on this mixed cell population, and the expression of the target gene in different samples is measured by qPCR. Here, we consider two cases.

Case-1: The expression of the target gene in each cell type remains constant, but the population distribution of the cell type changes across different samples.

Case-2: The target gene expression is time-dependent. The population distribution of different cell types and their gene expression change with time.

In the following section, we discuss the mathematical formulation of both cases.

### 6.2.1. Case 1: Gene expression is independent of time

Let us consider a population of cells with $n$ different cell types. Let us assume that there are $m+1$ samples with varying proportions of these cell types. Let on an average

the $k^{th}$ cell type expresses $x_k$ number of target mRNAs. The total number of target mRNAs in the population in sample *i* is,

$$X_{T,i} = \sum_{k=1}^{n} x_k \times N_{k,i} \tag{6.1}$$

here, $i = 0, 1, 2, \ldots m$ represent different samples. $N_{k,i}$ is the number of $k^{th}$ type cells in the $i^{th}$ sample. The sample $i = 0$ is the control sample to be used for fold-change estimation in qPCR. Note that $x_k$ is constant across all samples, but the population size of each cell type, $N_{k,i}$ changes.

The $\Delta\Delta\, C_t$ method is used to measure normalized fold change of the target gene expression from quantitative PCR (*144, 145*). The fold change is calculated with respect to the control sample, followed by normalization with the reference/housekeeping gene. Here, we derive the relation between normalized fold change in the population and the target gene expression in each cell type present in the population.

The fold change in target gene expression in sample *i* is,

$$\hat{X}_{T,i} = \frac{X_{T,i}}{X_{T,0}}$$

here, $X_{T,0}$ is the total number of mRNA of target gene in the control sample.

Similarly, fold change in reference gene expression in sample *i* is,

$$\hat{X}_{R,i} = \frac{X_{R,i}}{X_{R,0}}$$

here, $X_{R,i}$ is the total number of reference mRNA in sample *i* and $X_{R,0}$ is the total number of reference mRNA in the control sample.

Normalized fold change in the target gene expression in sample *i* is,

$$Y_i = \frac{\hat{X}_{T,i}}{\hat{X}_{R,i}} = \frac{1}{\hat{X}_{R,i}} \times \frac{X_{T,i}}{X_{T,0}} \tag{6.2}$$

here, $Y_i$ is the fold change measured from qPCR by the $\Delta\Delta\ C_t$ method.

From equation (6.1) and (6.2),

$$Y_i = \frac{1}{\hat{X}_{R,i}} \times \frac{\sum_{k=1}^{n} x_k \times f_{k,i}}{\sum_{k=1}^{n} x_k \times f_{k,0}} \tag{6.3}$$

here, $f_{k,i} = \frac{N_{k,i}}{N_i}$ is the fractional population size of cell type $k$ in sample $i$. $f_{k,0}$ is the fraction of $k^{th}$ cell type in the control sample. The fractions of each cell type are usually obtained from experiments like flow cytometry, quantitative image analysis.

Let's define $\hat{g}_k$, the Normalized Gene Expression Coefficient (NGEC) of $k^{th}$ cell type as,

$$\hat{g}_k = \frac{x_k}{\sum_{k=1}^{n} x_k \times f_{k,0}} \times \frac{1}{\hat{X}_{R,i}} \tag{6.4}$$

Now, equation (6.3) can be written as,

$$Y_i = \sum_{k=1}^{n} \hat{g}_k \times f_{k,i} \tag{6.5}$$

$\hat{g}_k$ is the level of expression of the target gene in the $k^{th}$ cell type normalized to the average expression of the target gene across all cell types and the fold change in the reference gene. The above equation should satisfy the constraint, $\hat{g}_k \geq 0$. The values

of $Y_i$ and $f_{k,i}$ are obtained from experiments. $\hat{g}_k$ is the unknown parameter to be estimated from the data.

### 6.2.2. Case 2: Gene expression is dependent on time

Here, we consider that the samples are collected for different time points ($i = 0, t_1, t_2, \ldots, t_m$) and the expression of the target gene in each cell type changes with time. For a time-dependent system with $m$ discrete time points, equation (6.5) becomes,

$$Y_i = \sum_{k=1}^{n} \hat{g}_k(i) \times f_{k,i} \tag{6.6}$$

here $\hat{g}_k(i)$ is the time-dependent NGEC of the target gene in $k^{th}$ cell type.

Usually, the expression of a gene either increases or decreases or remains constant with time. Though complicated behavior like oscillation may be observed in some instances, we considered that the time-dependent gene expression pattern in each cell type follows one of the three predefined linear functions. We used three predefined functions to reduce the complexity of the problem. These three functions are:

1. Linear time-dependent increase in gene expression:

$$\hat{g}_k(i) = \theta_k + \omega_k \times i \tag{6.7}$$

where $\theta_k, \omega_k$ are constants and $\theta_k, \omega_k \geq 0$.

2. Linear time-dependent decrease in gene expression:

$$\hat{g}_k(i) = \theta_k - \omega_k \times i \tag{6.8}$$

where $\theta_k, \omega_k$ are constants and $\theta_k, \omega_k \geq 0$ $and$ $\theta_k \geq \omega_k \times t_m$.

3. Constant gene expression:

$$\hat{g}_k(i) = \theta_k \tag{6.9}$$

where $\theta_k$ is a constant and $\theta_k \geq 0$.

Here, we considered time-varying gene expression. However, the same mathematical formulation can be used where the gene expression changes to experimental conditions like cells treated with different doses of a drug. In that case, $i$ represents the dose of the drug, and $i = 0$ represents untreated cells. The mathematical formulation for NGEC and the algorithm used for its estimation remains the same as for the time-dependent problem.

### 6.2.3. Estimation of NGECs through Bayesian approach

We estimate the NGEC of each cell type through the Bayesian approach. Here, we consider the unknown parameters as a random variable and estimate the probability distribution of the parameters given the experimental data. Using Bayes theorem (*199*) for our problem, we get

$$P(\hat{g}_1, \hat{g}_2, \ldots \hat{g}_n, \sigma^2 \mid Y_1, Y_2, \ldots Y_m) \propto P(Y_1, Y_2, \ldots Y_m \mid \hat{g}_1, \hat{g}_2, \ldots \hat{g}_n, \sigma^2) \times P(\hat{g}_1, \hat{g}_2, \ldots \hat{g}_n, \sigma^2)$$

The term on the left-hand side is the posterior distribution of the unknown parameters. The first term on the right-hand side is the data likelihood, and the second term is the prior distribution of the unknown parameters. $\hat{g}_1, \hat{g}_2, \ldots \hat{g}_n$ are the NGECs of each cell type. $Y_1, Y_2, \ldots Y_m$ are the normalized fold change of the target gene expression at various experimental conditions. $\sigma^2$ is the variance in the observed fold change in the target gene expression. We assume that the variance is constant across all experimental conditions.

The vector notation of the above equation is,

$$P(\hat{\boldsymbol{g}},\sigma^2|\boldsymbol{Y}) \propto P(\boldsymbol{Y}|\hat{\boldsymbol{g}},\sigma^2)\times P(\hat{\boldsymbol{g}},\sigma^2) \tag{6.10}$$

where $\hat{\boldsymbol{g}} = (\hat{g}_1, \hat{g}_2, \ldots \hat{g}_n)$; $\boldsymbol{Y} = (Y_1, Y_2, \ldots Y_m)$. $P(\hat{\boldsymbol{g}},\sigma^2|\boldsymbol{Y})$ is the posterior distribution of the unknown parameters. $P(\boldsymbol{Y}|\hat{\boldsymbol{g}},\sigma^2)$ is the data likelihood. $P(\hat{\boldsymbol{g}},\sigma^2)$ is the prior distribution of the unknown parameters.

In this formulation, both $\hat{\boldsymbol{g}}$ and $\sigma^2$ are unknown parameters. We assume that the experimental observations ($\boldsymbol{Y}$) are normally distributed around the true mean with some unknown variance ($\sigma^2$) (*199-201*). For *m+1* different samples and *n* different cell types, the data likelihood is,

$$P(\boldsymbol{Y}|\hat{\boldsymbol{g}},\sigma^2) = \prod_{i=0}^{m} \frac{1}{\sqrt{2\times\pi\times\sigma^2}} \times e^{\frac{-\left(Y_i - \left(\sum_{k=1}^{n} \hat{g}_k \times f_{k,i}\right)\right)^2}{2\times\sigma^2}} \tag{6.11}$$

The NGECs cannot be negative. Therefore, we used a truncated normal prior distribution (0 to +∞) for $\hat{\boldsymbol{g}}$ and an inverse gamma prior distribution for the variance (*202*). We employed a hierarchical Bayesian approach where the prior distributions of the NGECs are sampled from a hyperprior (*199*). We made use of conjugate-prior to the normal data likelihood, as described by Murphy (*203*) and Clyde et al. (*204*) The priors are defined as follows,

$$P\left(\hat{\boldsymbol{g}},\sigma^2\right) = P\left(\hat{\boldsymbol{g}}|\mu_{hyper}\right)\times P\left(\mu_{hyper}\right)\times P\left(\sigma^2\right)$$

$$P\left(\mu_{hyper}\right) \sim N\left(\mu_0, \sigma_0^2 = \frac{\sigma^2}{n_0}\right), \; bound \;\; from \; 0 \; to \; +\infty$$

$$P\left(\sigma^2\right) \sim \Gamma^{-1}\left(\alpha,\beta\right)$$

$$P\left(\hat{\boldsymbol{g}}|\mu_{hyper}\right) \sim N\left(\mu_{hyper}, \sigma_0^2 = \frac{\sigma^2}{n_0}\right), \; bound \;\; from \; 0 \; to \; +\infty$$

$P(\mu_{hyper})$ is the distribution of the hyperprior from which the prior distribution of $\hat{\boldsymbol{g}}$ are sampled. $\alpha$ *and* $\beta$ are the parameters that control the height and width of the

inverse gamma distribution, respectively. $n_0$ is the scale parameter that controls the variance in the prior ($\sigma_0^2$) relative to the variance in the data ($\sigma^2$). The graphical structure of the hierarchical model is shown in Figure 6.1a. In the GUI of DEBay, the user can define $\alpha, \beta, \mu_0,$ and $n_0$.

The posterior distribution is estimated by Markov Chain Monte Carlo (MCMC) using the NUT sampler (*205, 206*). Complete estimation steps were implemented in Python through the PyMC3 package (*207*).

For the time-dependent system discussed in section 6.2.2, we defined $\hat{\boldsymbol{g}}$ as a function of time, in terms of unknown parameters ($\boldsymbol{\theta}, \boldsymbol{\omega}$). The priors are defined as follows,

$$P\left(\hat{\boldsymbol{g}}, \sigma^2\right) = P\left(\boldsymbol{\theta} \mid \mu_{hyper}\right) \times P\left(\boldsymbol{\omega} \mid \mu_{hyper}\right) \times P\left(\mu_{hyper}\right) \times P\left(\sigma^2\right)$$

$$P\left(\mu_{hyper}\right) \sim N\left(\mu_0, \sigma_0^2 = \frac{\sigma^2}{n_0}\right), \; bound \;\; from \; 0 \; to \; +\infty$$

$$P\left(\sigma^2\right) \sim \Gamma^{-1}\left(\alpha, \beta\right)$$

$$P\left(\boldsymbol{\theta} \mid \mu_{hyper}\right) \sim N\left(\mu_{hyper}, \sigma_0^2 = \frac{\sigma^2}{n_0}\right), \begin{cases} bound \;\; from \; 0 \; to \; +\infty \; if \; \hat{\boldsymbol{g}}(i) = \boldsymbol{\theta} + \boldsymbol{\omega} \times i; \hat{\boldsymbol{g}}(i) = \boldsymbol{\theta} \\ bound \;\; from \; \boldsymbol{\omega} \times t_m \; to \; +\infty \; if \; \hat{\boldsymbol{g}}(i) = \boldsymbol{\theta} - \boldsymbol{\omega} \times i \end{cases}$$

$$P\left(\boldsymbol{\omega} \mid \mu_{hyper}\right) \sim N\left(\mu_{hyper}, \sigma_0^2 = \frac{\sigma^2}{n_0}\right), \; bound \;\; from \; 0 \; to \; +\infty$$

where $\alpha, \beta, \mu_0, n_0$ are the user-defined parameters. $P(\mu_{hyper})$ is the distribution of the hyperprior from which the prior distribution of $\hat{\boldsymbol{g}}$ are sampled. $\alpha$ and $\beta$ are the parameters that control the height and width of the inverse gamma distribution, respectively. $n_0$ is the scale parameter that controls the variance in the prior ($\sigma_0^2$) relative to the variance in the data ($\sigma^2$). The graphical structure of the hierarchical model is shown in Figure 6.1b.

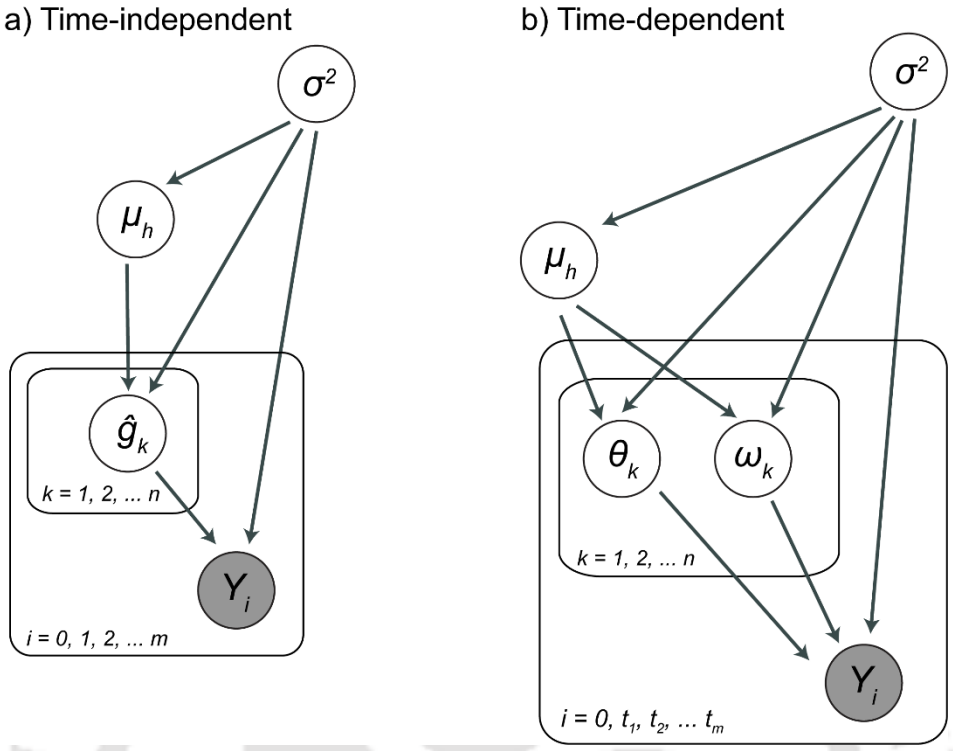



**Figure 6.1: Graphical representation of the hierarchical model**
The Bayesian hierarchical model structure of a) time-independent and b) time-dependent case. Here, $\mu_h \rightarrow \mu_{hyper}$. The white circles are the priors, and the grey circles are the observed data. The pointed arrows represent the dependency between each entity.

For the time-dependent system, the data likelihood is evaluated at the observed time points. For $m$ discrete time points and $n$ different cell types, the data likelihood is,

$$P(\boldsymbol{Y} \mid \hat{\boldsymbol{g}}, \sigma^2) = \prod_{i=0}^{t_m} \frac{1}{\sqrt{2 \times \pi \times \sigma^2}} \times e^{\frac{-\left[Y_i - \left(\sum_{k=1}^{n} \hat{g}_k(i) \times f_{k,i}\right)\right]^2}{2 \times \sigma^2}} \tag{6.12}$$

here, $\hat{g}_k(i)$ is evaluated as a function of time at the observed time points. $\hat{g}_k(i)$ can follow any of the three predefined functions (equation (6.7) - (6.9)). In general, for $n$ different cell types and three different predefined functions, the cell types and the functions can combine in $3^n$ possible ways. Each function combination is a possible model, and we estimate the posterior distribution for all models given the experimental data. The optimal model is selected based on the Bayes Information

Criterion (BIC) (*208, 209*). The model with minimum BIC is considered the optimal model.

$$BIC = -2 \times \log P\left(\boldsymbol{Y} \mid \left(\hat{\boldsymbol{g}}, \sigma^2\right)_{mean}\right) + \left(a \times \log b\right) \tag{6.13}$$

where $\log P(\boldsymbol{Y}|(\hat{\boldsymbol{g}}, \sigma^2)_{mean})$ is the log-likelihood evaluated at the mean of each parameter distribution; $a$ is the number of unknown parameters, including the variance and the hyperprior; $b$ is the number of experimental observations.

From the optimal model, we get the distribution of each parameter in the predefined functions (equations (6.7)- (6.9)). By using the random numbers from these parameter distributions in $\hat{g}_k(i)$, we get the distribution of NGEC of each cell-type for each discrete time point.

## 6.3. The graphical user interface of DEBay

We created a Graphical User Interface (GUI) of DEBay and is available as a standalone Windows installer at SourceForge (https://sourceforge.net/projects/debay/). DEBay is built with Python. The python packages used in DEBay use C libraries for faster computation. Therefore, a GCC would help to improve the performance of DEBay, but it is not mandatory. MinGW, a complete run time environment for GCC for windows is available at SourceForge (https://sourceforge.net/projects/mingw/).

DEBay takes two data – a) fold change in expression of the target gene in samples, and b) proportion of each cell type in these samples. Using these data, DEBay estimates Normalized Gene Expression Coefficient (NGEC) of each cell type using the Bayesian method of parameter estimation. DEBay can handle both time-dependent and independent gene expression cases. The workflow of DEBay is illustrated in Figure 6.2.

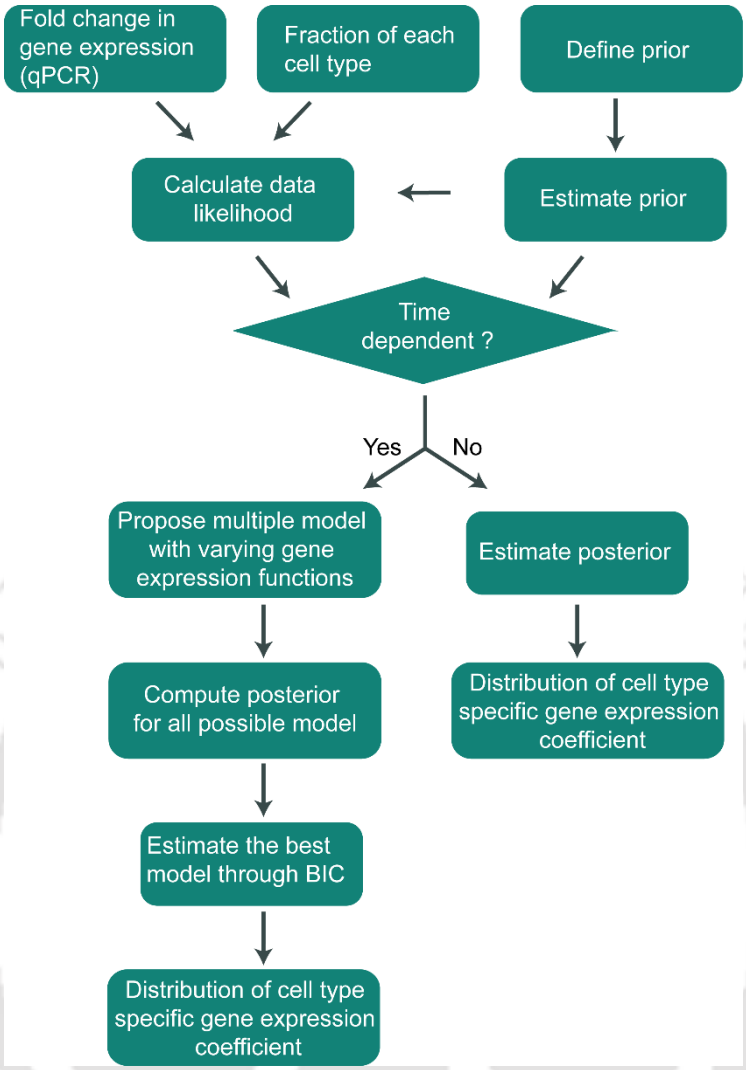



**Figure 6.2: Workflow of DEBay.**

DEBay takes the following input parameters from the user: a) Whether to use time-dependent or -independent model, b) Excel sheet containing both the population-level fold change in target gene expression and the proportions of each cell type in the population, c) Parameter values for MCMC, d) Parameter values specifying the prior distributions. The description of each input parameter and the instructions to use DEBay are given in the user manual of DEBay. A glimpse of the GUI is shown in Figure 6.3.

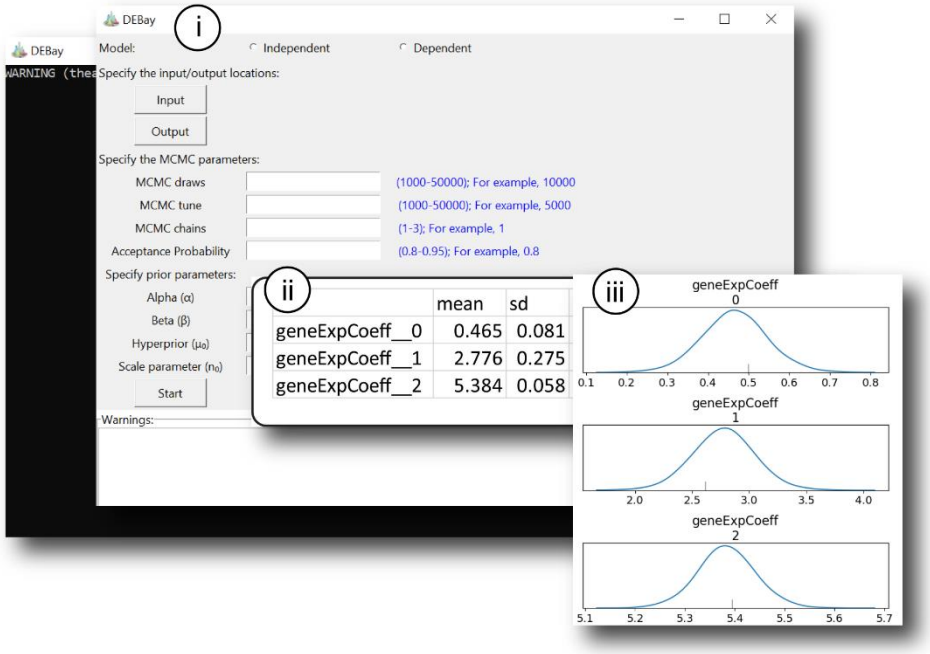



**Figure 6.3: The GUI of DEBay.**
Snapshots of the GUI (*i*) and outputs (*ii* and *iii*) generated by DEBay.

A command prompt window accompanies the user interface, and the progress of the deconvolution is shown in the command prompt. The deconvoluted NGECs are saved as a tab-delimited text file along with the statistics. The probability distributions of the gene expression coefficients are also saved in a graphical format.

## 6.4. Evaluation of DEBay with synthetic data

### 6.4.1. Generation of synthetic data

We tested DEBay with real biological data as well as synthetic data. In this section, we will discuss the synthetic data generation and testing DEBay with the synthetic data sets.

To generate synthetic data, we considered five samples ($i = S1, S2, S3, S4,$ and $S5$) having a population of $N = 10^6$ cells per sample, with four different subpopulations, A, B, C, and D.

In general, suppose for the $i^{th}$ sample, there are $N_{k,i}$ number of $k$-type cells. $N_{k,i} = N \times f_{k,i}$, where $f_{k,i}$ is the fractional size of the $k^{th}$ subpopulation in the $i^{th}$ sample. The total number of mRNAs of the target gene in $k^{th}$ subpopulation in this sample, $X_{k,i} = \sum_{j=1}^{N_{k,i}} x_{j,k,i}$. Here, $x_{j,k,i}$ is the number of mRNAs of the target gene in the $j^{th}$ cell of $k$-type in $i^{th}$ sample. Therefore, the total number of the target mRNAs in the whole population is $\sum_k X_{k,i}$.

The fold change in target gene expression in the $i^{th}$ sample was calculated with respect to the control sample. We considered $i$ = S1 as the control sample. Therefore, the fold change in expression of the target gene in the $i^{th}$ sample is

$$Y_i = \frac{\sum_k X_{k,i}}{\sum_k X_{k,S1}} = \frac{\sum_k \sum_{j=1}^{N_{k,i}} x_{j,k,i}}{\sum_k \sum_{j=1}^{N_{k,S1}} x_{j,k,S1}} \tag{6.14}$$

Usually, the change in the expression of the target gene is reported as normalized fold change, where the normalization is done with respect to the fold change in the expression of a reference/housekeeping gene. For simplification, we considered that normalization term as one. Therefore, equation (6.14) gives us the normalized fold change in the target gene in the $i^{th}$ sample.

We varied the fractional population size of each cell-type ($f_{k,i}$) such that $\sum_k f_{k,i} = 1$. The number of target mRNAs in each $k^{th}$ type cell ($x_{j,k,i}$), in a sample, was generated by repeated sampling ($N_{k,i}$ $times$) from a normal distribution $N(\mu_{k,i}, \sigma_{k,i})$. The number of copies of mRNAs of highly expressed genes in human cells varies in the range of $10^3$ - $10^5$ per cell (*210*). Therefore, $\mu_{k,i}$ was varied from $10^2$ - $10^5$.

The noise in gene expression is often quantified as $\eta_{k,i} = \frac{\sigma_{k,i}}{\mu_{k,i}}$. If we fix the level of the noise ($\eta_{k,i}$), then $\sigma_{k,i}$ can be calculated for a given $\mu_{k,i}$. For simplicity, we considered

equal noise in all the subpopulations in a sample. We generated samples with different levels of noise and calculated the fold change in expression using equation (6.14). We observed that the noise in mRNA level did not affect the calculated fold change (Figure 6.4). The effect of noise is nullified since we considered constant noise across all cell types in a population. Therefore, in all further data sets, we used $\sigma_{k,i} = 10^{-1} \times \mu_{k,i}$.

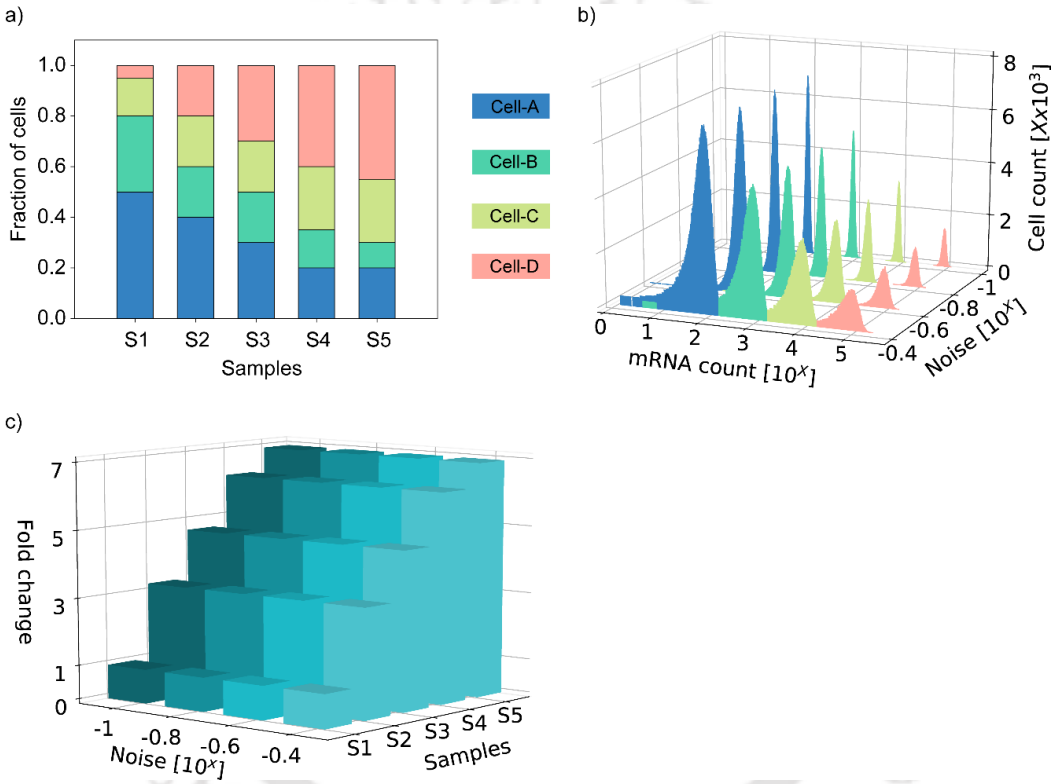



**Figure 6.4: Effect of noise in mRNA level on population-level fold change in gene expression.**
a) Each sample was a mixture of four types of cells ( A, B, C, and D). b) The mean number of mRNAs in each cell type was fixed. The number of mRNA in each cell was sampled from distributions having four different levels of noise, $\eta = \sigma/\mu$. c) Fold change in gene expression was calculated from these randomly sampled mRNA data. S1, S2, …, and S5 are different samples.

### 6.4.2. Evaluating DEBay with time-independent gene expression data

We generated 1000 independent synthetic data sets. For a particular data set, the population size of each cell-type ($f_{k,i}$) varied randomly among the samples, but the

mean level of expression of the target gene ($\mu_{k,i}$) remained the same ($\mu_{k,i}$ = constant for all $i$). We varied $f_{k,i}$ by sampling from a uniform distribution U(0,1) such that $\sum_k f_{k,i} = 1$. The mean number of target mRNAs in each cell type, $\mu_k$ was varied randomly from $10^2$-$10^5$. Accordingly, $\sigma_k$ was decided such that $\sigma_k = 10^{-1} \times \mu_k$. As mentioned earlier, repeated sampling was done from $N(\mu_k, \sigma_k)$ to generate the data for each cell in each subpopulation. These sampled data were used to calculate the population-level fold change in expression using equation (6.14).

Figure 6.5a shows the deviation of the estimated NGECs to the actual ones, relative to the standard deviation of the estimated NGECs. In all cell types, the distribution of the deviation showed a tight cluster around zero. Most of the actual NGECs were within one standard deviation of the estimated NGECs. The actual and estimated NGECs showed a robust positive correlation with $r^2 = 0.99$ (Figure 6.5b).

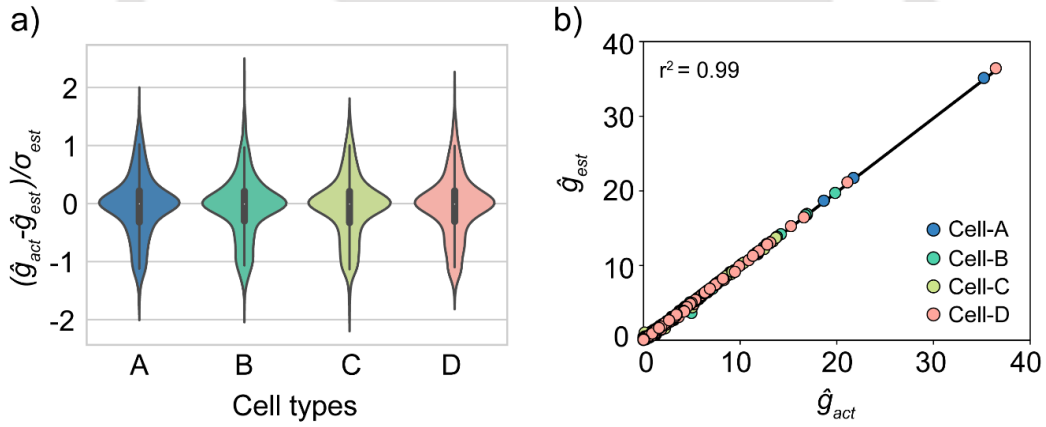



**Figure 6.5: Evaluation of the performance of DEBay for synthetic data sets.**
One thousand different synthetic data sets were generated with different population fractions and different amounts of mean mRNAs in each cell-types. a) The deviation between the mean of estimated NGECs ($\hat{g}_{est}$) and the actual NGECs ($\hat{g}_{act}$) to the standard deviation of the estimated NGECs ($\sigma_{est}$). $\hat{g}$ denotes Normalized Gene Expression Coefficient (NGEC). The deviations are tightly centered around zero. b) Correlation between the mean of estimated NGECs ($\hat{g}_{est}$) and the actual NGECs ($\hat{g}_{act}$).

The analysis of a representative synthetic data set is shown in Figure 6.6. There were five samples S1 to S5 containing four different cell types in various proportions (Figure 6.6a). The population-level fold change in target gene expression is shown in

Figure 6.6b. These two data were used as input to DEBay. The estimated mean NGECs for all the cell types (Figure 6.6c) are close to the respective actual values (Figure 6.6d) calculated algebraically using equation (2.8). Figure 6.6e shows the distribution of the estimated cell-type-specific NGECs.

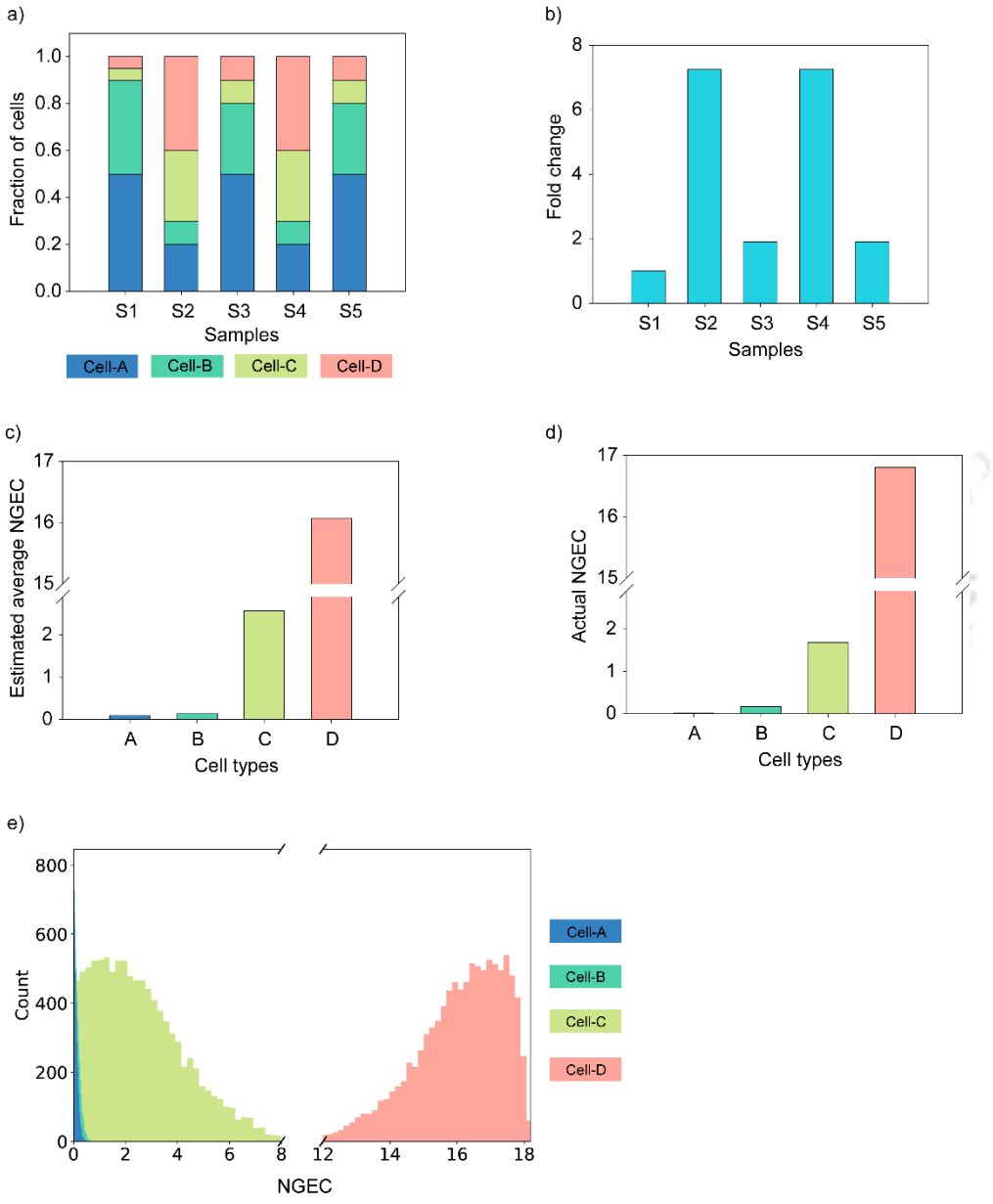



**Figure 6.6: Evaluating DEBay with time-independent synthetic data.**
The data set has five samples, and each sample is composed of four types of cells. The population size of each cell type varies among samples (a). (b) shows the fold change in expression of the target gene in different samples. The estimated average NGECs for different cell types are shown in (c). The actual values of the NGECs are shown in (d). e) Distribution of the estimated NGECs of four cell types.

### 6.4.3. Evaluating DEBay with time-dependent gene expression data

In the time-dependent case, we used three sets of synthetic data sets. Each data set has five samples, representing five-time points, from 0 to 48 h at an interval of 12 h. From equation (6.14), the fold change in the target gene expression is,

$$Y_i = \frac{\sum_k X_{k,i}}{\sum_k X_{k,0}} = \frac{\sum_k \sum_{j=1}^{N_{k,i}} x(i)_{k,j}}{\sum_k \sum_{j=1}^{N_{k,0}} x(0)_{k,j}} \tag{6.15}$$

where $i = 0, 12, 24, 36, 48\ h$. $x(i)_{k,j}$ is the time-dependent function representing the number of target mRNAs in the $j^{th}$ cell of type $k$ at the $i^{th}$ time point. Here $k$ = A, B, C, and D (different cell types).

The functions used to generate the number of target mRNAs are given in Table 6.1 - Table 6.3. The parameters in the functions ($p = (\theta, \omega)$) are generated using random numbers from, $N{\sim}(\mu_p, \sigma = 10^{-1} \times \mu_p)$. The values of $\mu_\theta, \mu_\omega$ are given in the Table 6.1 - Table 6.3. The fraction of each cell type is varied from 0 to 1, such that the summation of fractions of all cell types at a given experimental condition is 1. Using the fractions of different cell types and the mRNAs expression in each cell type in equation (6.15), the population-level fold change in target gene expression is calculated.

**Table 6.1. Expression of target mRNAs in each cell type as a function of time – SET 1.**

| *Cell types* | *mRNAs as a function of time* | $\mu_\theta$ | $\mu_\omega$ |
|---|---|---|---|
| A | $\theta + \omega \times i$ | 100 | 2000 |
| B | $\theta + \omega \times i$ | 1000 | 1300 |
| C | $\theta$ | 10000 | - |
| D | $\theta - \omega \times i$ | 100000 | 2000 |

**Table 6.2. Expression of target mRNAs in each cell type as a function of time – SET 2.**

| *Cell types* | *mRNAs as a function of time* | $\mu_\theta$ | $\mu_\omega$ |
|---|---|---|---|
| A | $\theta - \omega \times i$ | 15000 | 250 |
| B | $\theta$ | 1000 | - |
| C | $\theta$ | 500 | - |
| D | $\theta + \omega \times i$ | 5000 | 100 |

**Table 6.3. Expression of target mRNAs in each cell type as a function of time – SET 3.**

| *Cell types* | *mRNAs as a function of time* | $\mu_\theta$ | $\mu_\omega$ |
|---|---|---|---|
| A | $\theta$ | 2500 | - |
| B | $\theta + \omega \times i$ | 500 | 200 |
| C | $\theta - \omega \times i$ | 12500 | 300 |
| D | $\theta$ | 100 | - |

The synthetic data sets were loaded into DEBay, and the NGEC of each cell type were deconvoluted. The results of the deconvolution of all three synthetic data sets are shown in Figure 6.7, Figure 6.8, and Figure 6.9, respectively. In all three synthetic data sets, the time-dependent profile of the deconvoluted NGECs was very close to the actual NGECs. Figure 6.10a shows the deviation between the actual NGECs and the estimated NGECs in relative to the standard deviation of the estimated NGECs for all three synthetic data sets. In most of the cases, the actual gene expression coefficient lies within 1.5 standard deviations of the estimated gene expression coefficients. The estimated gene expression coefficients showed a reliable positive correlation to the actual gene expression coefficients with $r^2 = 0.97$ (Figure 6.10b).

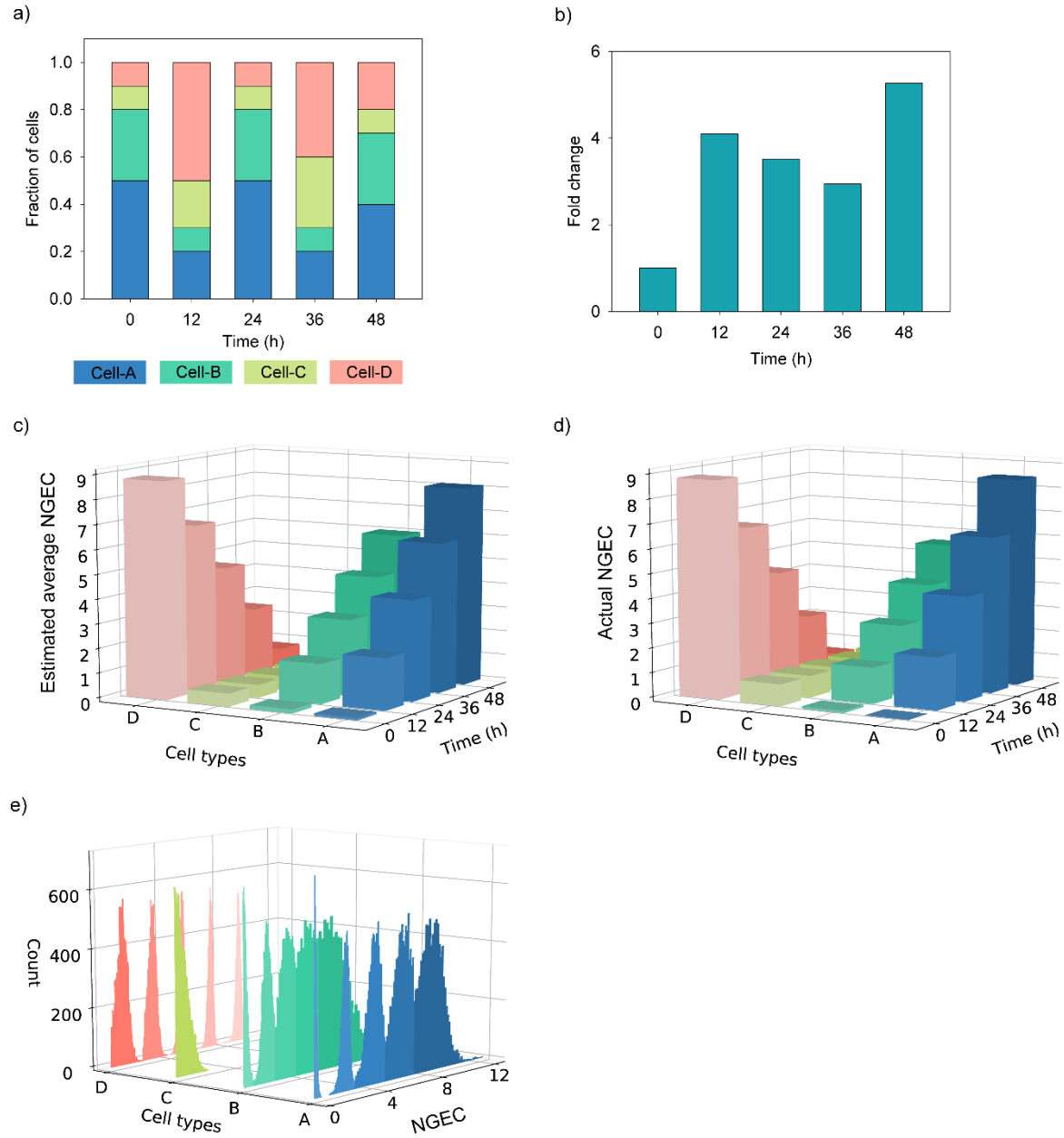



**Figure 6.7: Evaluating DEBay with time-dependent synthetic data set-1.**
The data set has five samples corresponding to five time-points. Each sample is composed of different proportions of four types of cells. (a) The change in the proportions of cell types with time. (b) Fold change in expression of the target gene in the whole population with time. Deconvolution was performed using the time-dependent model of DEBay. The estimated average NGECs and the actual NGECs for different cell types are shown in (c) and (d), respectively. e) Distribution of the estimated time-dependent NGECs. Time points are represented by increasing order of color intensities. The lowest intensity denotes the initial time point ($t = 0$), and the highest intensity denotes the end time point ($t = 48\ h$).

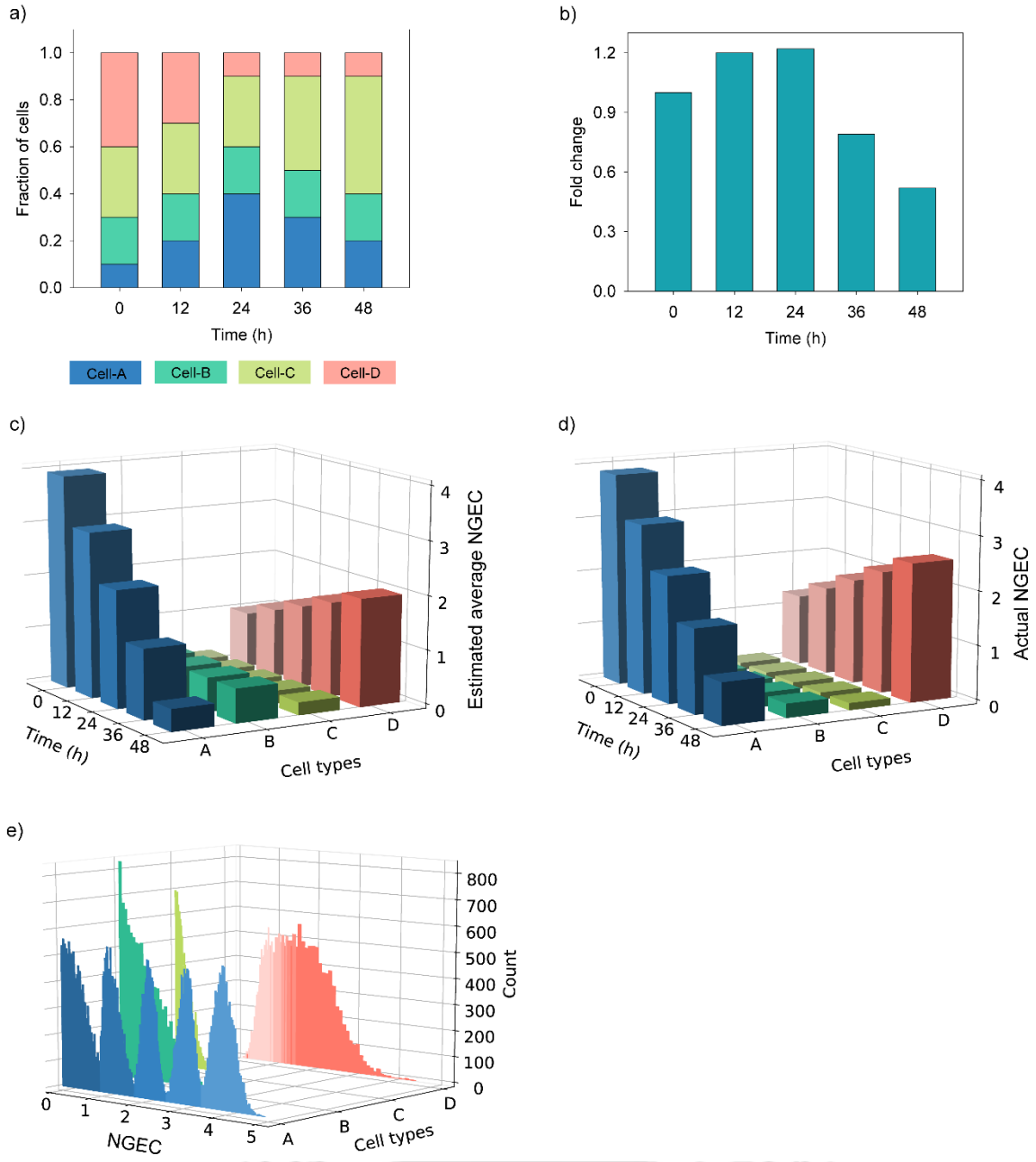



**Figure 6.8: Evaluating DEBay with time-dependent synthetic data set-2.**
The data set has five samples corresponding to five time-points. Each sample is composed of different proportions of four types of cells. (a) The change in the proportions of cell types with time. (b) Fold change in expression of the target gene in the whole population with time. Deconvolution was performed using the time-dependent model of DEBay. The estimated average NGECs and the actual NGECs for different cell types are shown in (c) and (d), respectively. e) Distribution of the estimated time-dependent NGECs. Time points are represented by increasing order of color intensities. The lowest intensity denotes the initial time point ($t = 0$), and the highest intensity denotes the end time point ($t = 48\ h$).

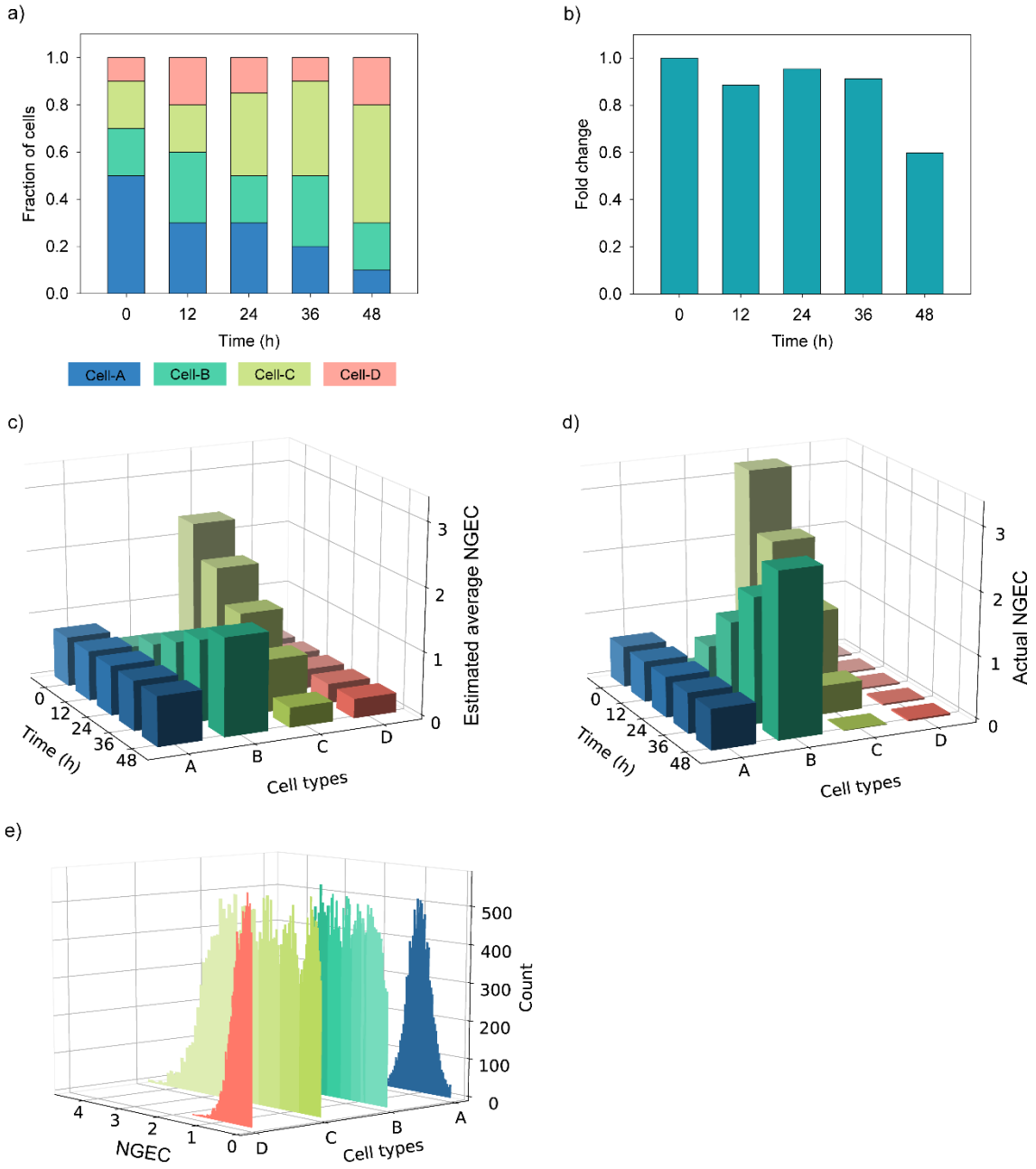



**Figure 6.9: Evaluating DEBay with time-dependent synthetic data set-3.**
The data set has five samples corresponding to five time-points. Each sample is composed of different proportions of four types of cells. (a) The change in the proportions of cell types with time. (b) Fold change in expression of the target gene in the whole population with time. Deconvolution was performed using the time-dependent model of DEBay. The estimated average NGECs and the actual NGECs for different cell types are shown in (c) and (d), respectively. e) Distribution of the estimated time-dependent NGECs. Time points are represented by increasing order of color intensities. The lowest intensity denotes the initial time point ($t = 0$), and the highest intensity denotes the end time point ($t = 48\ h$).

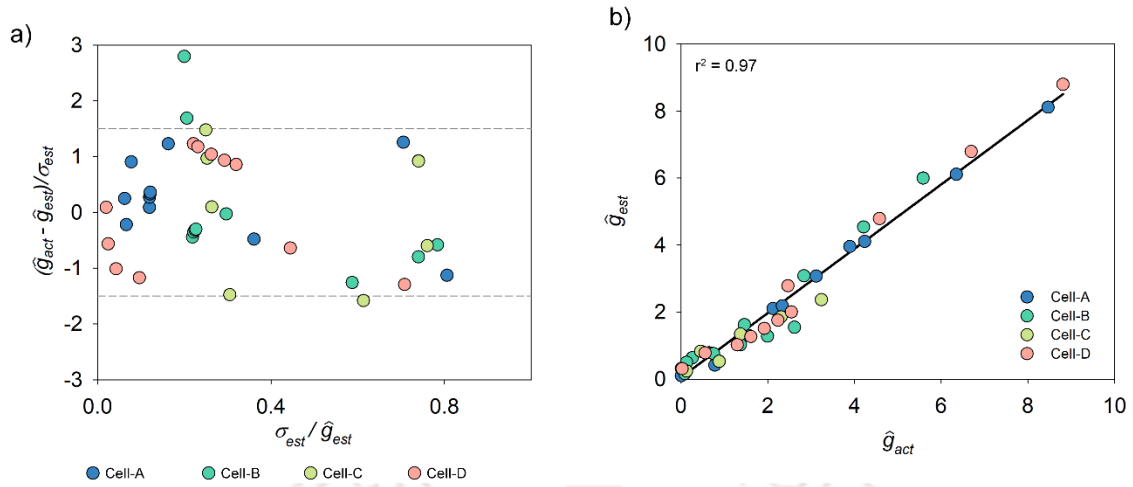



**Figure 6.10: Accuracy of the deconvoluted parameters of time-dependent synthetic data sets.**
Three time-dependent synthetic data sets were generated with different population fractions and different cell-type-specific mRNA expression levels (Table 6.1 - Table 6.3). a) The deviation between the mean of estimated NGECs ($\hat{g}_{est}$) and the actual NGECs ($\hat{g}_{act}$) to the standard deviation of the estimated NGECs ($\sigma_{est}$). $\hat{g}$ denotes Normalized Gene Expression Coefficient (NGEC). b) Correlation between the mean of estimated NGECs ($\hat{g}_{est}$) and the actual NGECs ($\hat{g}_{act}$).

## 6.5. Evaluating DEBay with real biological data

We performed a quantitative PCR (qPCR) experiment to test the software. We used three different human breast cancer cell lines, MCF-7, MDA-MB-231, and MDA-MB-468. These cell line expresses different amounts of EGFR (MDA-MB-468 > MDA-MB-231 > MCF-7) (*211*). We measured the expression of EGFR in all three cells through qPCR. The $C_t$ values of EGFR expression were normalized to the reference gene (Cyclophilin A). The normalized $C_t$ values of MDA-MB-468, MDA-MB-231, and MCF-7 were 0.9, 1.1, and 1.4, respectively. Higher the $C_t$ value, lower is the expression level.

We mixed these three cell lines in various proportions S1, S2, and S3 (Figure 6.11a). Through qPCR, we measured the fold change in EGFR expression in samples S2 and S3 with respect to S1 (Figure 6.11b). We used the fold change in EGFR in the mixed cells (S1, S2, and S3) and their proportion in DEBay and estimated the gene expression coefficient of EGFR in each cell type. Figure 6.11c shows the distribution of the

deconvoluted NGECs. The estimated NGECs of EGFR followed the same pattern, MDA-MB-468 > MDA-MB-231 > MCF-7, as measured by qPCR from pure cell types (Figure 6.11d).

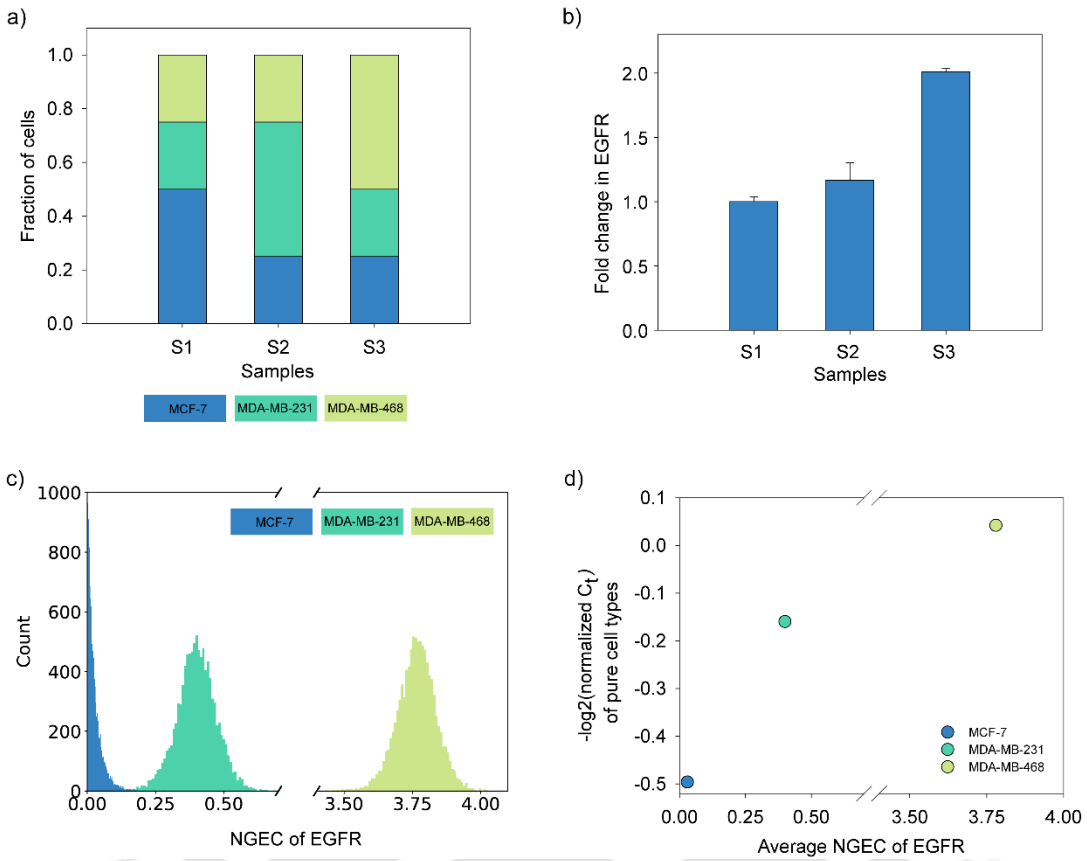



**Figure 6.11: Deconvolution of real qPCR data using DEBay.**
(a) Three samples were prepared by mixing different proportions of MCF-7, MDA-MB-486, and MDA-MB-231 cells. (b) shows normalized fold change in expression of EGFR in different samples as measured by qPCR. c) Distribution of estimated NGECs of EGFR expression in three cell types. d) shows the correspondence between the experimentally measured level of expression of EGFR in three pure cell lines and NGECs of these cells estimated by DEBay from three mixed samples. $C_t$ values of EGFR were normalized with the $C_t$ values of cyclophilin A. Normalized $C_t$ value is a proxy of gene expression level and is inversely related to the level of expression of a gene. Negative $\log_2$ transformation was used for better visualization of the data.

In chapter 4, we studied the EGF-induced morphological state transition dynamics in Epithelial to Mesenchymal Transition (EMT) of MDA-MB-468 cells. We observed three distinct morphologies of MDA-MB-468 cells - Cobble, Spindle, and Circular. The

population distribution of these cell types changed with the dose and the duration of EGF treatment. Through the Boyden-Chamber assay, we showed that the Cobble cells are non-migratory and epithelial-like, whereas Spindle and Circular cells are migratory and mesenchymal-like.

Using image analysis, we measured the population distribution of these cell-types at different time points and the population-level expression of few mesenchymal markers like vimentin and SNAIL1 through qPCR. These data are shown in Figure 6.12a and b. We used these data to estimate cell-type-specific vimentin and SNAIL1 expression. The deconvoluted data are shown in Figure 6.12c. The NGECs of vimentin and SNAIL1 for Cobble cells are extremely low and do not change with time.

On the other hand, NGECs of these two genes for the Spindle and Circular cells are very high or increased with time (Figure 6.12c). It is known that the expressions of vimentin and SNAIL1 are low in epithelial cells and high in mesenchymal cells (*11, 212*). Therefore, from the deconvoluted data, we can conclude that the Spindle and Circular cells are mesenchymal cells, and the Cobble cells are epithelial. This observation matches well with the migratory pattern of these cell types (Boyden Chamber assay).

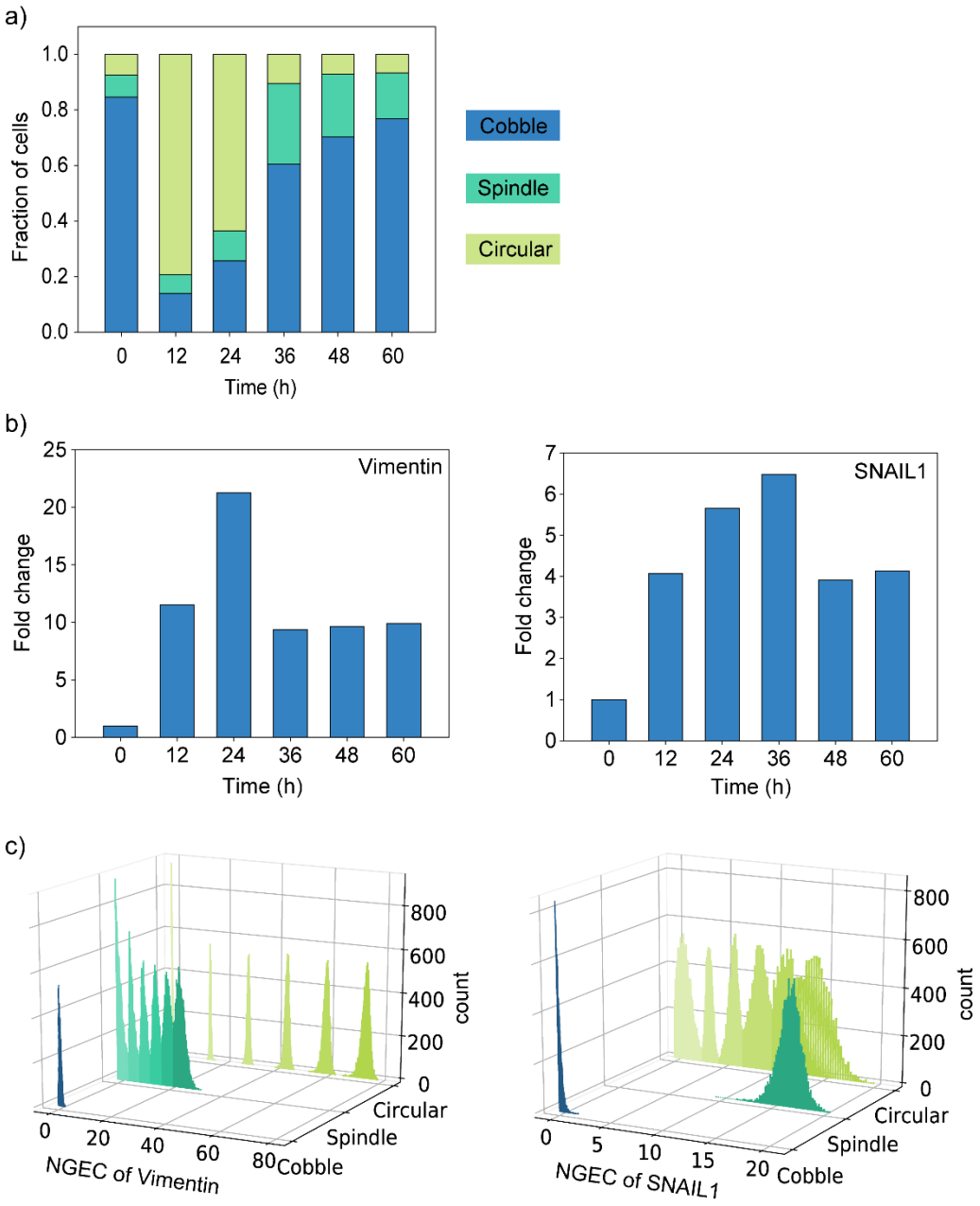



**Figure 6.12: Evaluating DEBay with real time-dependent data.**
MDA-MB-468 cells were treated with EGF for different durations, and the proportions of three cell types were measured by quantitative image analysis (a). qPCR was used to estimate the time-dependent changes in the expression of two markers of EMT, Vimentin, and SNAIL1(b). These data were deconvoluted using the time-dependent model of DEBay. c) Time-dependent distribution of NGEC of Vimentin and SNAIL1, in three cell types. Time points are represented by increasing order of color intensities. The lowest intensity denotes $t = 0$, and the highest intensity denotes $t = 60\ h$.

## 6.6. Discussion

In this chapter, we presented DEBay, a tool for computational deconvolution of quantitative PCR data to estimate time-dependent and -independent expression of the target gene in different subpopulations in an ensemble of cells. The time-dependent algorithm of DEBay can also handle qPCR data, where the expression of target gene changes with different doses of a drug. Instead of time, the corresponding experimental condition should be used.

DEBay estimates the relative expression level of the target gene in each cell type as the Normalized Gene Expression Coefficient (NGEC). We mathematically derived NGEC from the fold change in gene expression measured through qPCR. NGEC is a proxy of the normalized gene expression level in each cell type in the sample.

Existing deconvolution methods consider that the gene expression of a cell type remains constant with time or across experimental conditions. DEBay addresses this problem by considering three different models of time-dependent gene expression – linear increase, linear decrease, and constant. One could envisage various nonlinear gene expression patterns, and the algorithm of DEBay can easily be altered to accommodate such dynamics. However, an increase in the number of alternative models increases the complexity of the problem and the computation time. Further, nonlinear models suffer from the problem of overfitting when the number of data points is low. Therefore, for the current version of DEBay, we used three linear models that are most commonly observed in experiments.

Most of the deconvolution algorithms use the frequentist approach to estimate the unknown parameters (*13*, *17*, *18*, *21*, *22*, *136*, *138*). These methods converge to a point estimate of the parameter and usually report *P* value or confidence interval of the estimated parameter. This approach does not address the probability distribution of the estimated parameters. In our method, we used the Bayesian method of parameter

estimation. In this approach, the parameters are considered random variables, and we estimate the posterior distribution of the parameters based on the observed data. Through this approach, we can estimate the credible interval of the estimated parameters.

In our method, we used a hierarchical model structure. The prior distributions of cell-type-specific gene expression coefficients are defined based on a hyperprior. Through this model structure, each cell-type-specific coefficient is indirectly constrained by all the observed data through the hyperprior. The credible intervals of cell-type-specific coefficients are pulled towards the mode of the hyperprior. Therefore, the sampling of posterior probabilities becomes more efficient (*199*).

We tested our tool with real biological data and various synthetic data sets. Our algorithm performs reasonably well in all cases with minimal deviation from the actual values. We developed DEBay, keeping in mind the needs of low-throughput but widely used qPCR experiments. The GUI of DEBay is intuitive and straightforward. One can use the fold change data obtained from qPCR experiments without any correction or transformation. The output files created by DEBay are also self-explanatory.

CHAPTER 7

# Conclusion

In this study, we investigated the cellular plasticity of EGF-induced EMT of MDA-MB-468 cells. We defined the phenotype of cells based on the morphology and studied the dynamics of cell state transition using a population dynamics model. MDA-MB-468 cells exist in three different morphological forms: Cobble, Spindle, and Circular. We measured the population distribution of the three cell types at discrete time points through quantitative image analysis. In the absence of EGF, the population of cells was in a steady-state distribution, Cobble: Spindle: Circular = 0.79: 0.13: 0.08. EGF treatment disturbed the steady-state distribution.

We developed a mathematical method to estimate the time evolution of cells from the image analysis data. The data analysis showed that the EGF-induced cell state transition of MDA-MB-468 cells is either reversible or irreversible depending on the strength and the duration of the input signal. A moderate dose of EGF induced a reversible transition of cells while a higher dose of EGF resulted in an irreversible state transition. The forward transition followed Cobble → Circular state transition,

while the reverse transition followed, Circular → Spindle → Cobble. Our analysis showed that the Spindle cells evolved predominantly during the reverse transition.

Phosphorylated EGFR is a measure of active EGF signaling. Our experiments showed that the sustained EGF signaling favored the forward transition from Cobble → Circular state. The Circular cells reverted to Cobble through the intermediate Spindle state during the decay phase of EGF signaling. We further showed that the interplay between FAK signaling and EGF signaling drives the cell state transition of MDA-MB-468 cells. Our analysis showed an ultrasensitive response like behavior in the Circular cell state with the phospho-EGFR level. We experimentally proved that this switch controls the transition of cells in and out of the Circular state. When the switch is OFF, very few cells were Circular, and when the switch is ON, majority of the cells moved to Circular state.

Our experimental data are not single-cell time series data, rather discrete-time population distribution data. These types of data are called aggregate data. Our mathematical method is generic and can be used to study the state transition dynamics of any cellular system wherever aggregate data is available. In this study, we explored the state transition process at the phenomenon-level. In the future, this model can be coupled with the EGF signaling network. This would help us map the cell's molecular state to the morphological state of the cell, and we can explore several control switches governing the state transition. We also proposed an alternative method to study state transition of discrete phenotypic states. Our proposed method is simple and can be used in stochastic modeling of cell state transition.

Next, we studied the signal transduction of EGF-induced EMT in the presence and absence of background noise. We introduced background noise by adding a suboptimal dose of TGF-β1. TGF-β1 did not induce state transition of cells on its own. Instead, modulated the EGF-induced state transition. Our statistical analysis showed

that the EGF-induced state transition is inherently noisy; that is, the cells cannot distinguish different EGF doses. The addition of TGF-β1 further increased the noise and thus, reduced the signal transmission capacity.

Treatment of cells with TGF-β1, in the presence of EGF, had a negative synergistic effect on Circular cells while a positive synergistic effect on Spindle and Cobble cells. Our experiments also showed that persistent EGF signaling keeps the cells in the Circular state and pushes them towards apoptosis. Therefore, possibly the TGF-β1-induced increase in noise rescue the cells from apoptosis and pushes them to either Spindle or Cobble states, thereby favoring the reverse transition of cells from Circular → Cobble state. A molecular-level investigation is necessary to understand how the background noise (TGF-β1) helps in the decision making of cells. This investigation can be extended to several different input signals at suboptimal doses. This would provide an ideal setup to study signal transduction in cells and give us an in-depth knowledge of the role of background noise in signal transduction of cells.

We developed a computational tool, DEBay, that estimates cell-type-specific gene expression profile from quantitative PCR data of a mixed population of cells. DEBay will be more useful in experiments, where the isolation of different cell types is not possible. In MDA-MB-468 cells, we categorized the phenotypic states based on the morphology. We did not have any information on the molecular state of the cell. Moreover, we cannot isolate the different phenotypic states based on morphology. However, we had population-level gene expression data and the proportions of different cell types in the population. Using these data in DEBay, we estimated the cell-type-specific gene expression of each cell type.

Vimentin and SNAIL1 are the molecular markers of mesenchymal/migratory cells. Deconvolution of Vimentin and SNAIL1 expression showed that Cobble cells had few copies of these genes, while Spindle and Circular cells had several copies. Therefore,

the Spindle and Circular cells are migratory, while Cobble cells are non-migratory. These results were in line with the Boyden Chamber assay. Further, we evaluated the tool with several synthetic data sets, and the performance was reasonably good. DEBay is a comprehensive tool, that can deconvolute both time-dependent and -independent gene expression data.

The time-dependent model of DEBay considers that the gene expression of each cell type in a population follows any one of the three pre-defined linear functions. There is much scope for further development of the tool to incorporate non-linear gene expression profiles. DEBay can also analyze other gene expression data like cells treated with different doses of a drug. The same time-dependent model of DEBay can be used with a slight change. Instead of time, the corresponding experimental condition should be used. We created an intuitive GUI of DEBay and is openly available in Source Forge.

# Bibliography


1. R. D. Brackston *et al.*, Transition state characteristics during cell differentiation. *PLoS computational biology* **14**, e1006405 (2018).

2. M. R. Martinez *et al.*, Quantitative modeling of the terminal differentiation of B cells and mechanisms of lymphomagenesis. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 2672-2677 (2012).

3. K. Okamoto *et al.*, Single cell analysis reveals a biophysical aspect of collective cell-state transition in embryonic stem cell differentiation. *Scientific reports* **8**, 11965 (2018).

4. J. Zhang *et al.*, TGF-beta-induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. *Science signaling* **7**, ra91 (2014).

5. T. Buder *et al.*, CellTrans: An R Package to Quantify Stochastic Cell State Transitions. *Bioinformatics and biology insights* **11**, 1177932217712241 (2017).

6. P. B. Gupta *et al.*, Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell* **146**, 633-644 (2011).

7. M. Mandal *et al.*, Modeling continuum of epithelial mesenchymal transition plasticity. *Integr Biol (Camb)* **8**, 167-176 (2016).

8. M. Setty *et al.*, Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature biotechnology* **34**, 637-645 (2016).

9. W. A. Farahat, H. H. Asada, Estimation of state-transition probability matrices in asynchronous population Markov processes. *Proceedings of the 2010 American Control Conference*, 6519-6524 (2010).

10. E. Beretta *et al.*, Mathematical modelling of cancer stem cells population behavior. *Mathematical Modelling of Natural Phenomena* **7**, 279-305 (2012).

11. S. Lamouille *et al.*, Molecular mechanisms of epithelial-mesenchymal transition. *Nature reviews. Molecular cell biology* **15**, 178-196 (2014).

12. I. Pastushenko, C. Blanpain, EMT Transition States during Tumor Progression and Metastasis. *Trends in cell biology* **29**, 212-226 (2019).

13. A. R. Abbas *et al.*, Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *Nature methods* **4**, e6098 (2009).

14. K. Dimitrakopoulou *et al.*, Deblender: a semi-/unsupervised multi-operational computational method for complete deconvolution of expression data from heterogeneous samples. *BMC bioinformatics* **19**, 408 (2018).

15. R. Du *et al.*, deconvSeq: Deconvolution of Cell Mixture Distribution in Sequencing Data. *Bioinformatics (Oxford, England)* **35**, 5095–5102 (2019).

16. T. Erkkila *et al.*, Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics (Oxford, England)* **26**, 2571-2577 (2010).

17. T. Gong *et al.*, Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS One* **6**, e27156 (2011).

18. D. A. Liebner *et al.*, MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. *Bioinformatics (Oxford, England)* **30**, 682-689 (2014).

19. A. M. Newman *et al.*, Robust enumeration of cell subsets from tissue expression profiles. *Nature methods* **12**, 453-457 (2015).

20. W. Qiao *et al.*, PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS One* **8**, e1002838 (2012).

21. S. S. Shen-Orr *et al.*, Cell type-specific gene expression differences in complex tissues. *Nature methods* **7**, 287-289 (2010).

22. Y. Zhong *et al.*, Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC bioinformatics* **14**, 89 (2013).

23. R. Dawkins, *The extended phenotype*. (Oxford University Press, Oxford, 1982), vol. 8.

24. M. Pigliucci *et al.*, Phenotypic plasticity and evolution by genetic assimilation. *The Journal of experimental biology* **209**, 2362-2367 (2006).

25. G. Fusco, A. Minelli, Phenotypic plasticity in development and evolution: facts and concepts. Introduction. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **365**, 547-556 (2010).

26. P. S. Stumpf *et al.*, Stem Cell Differentiation as a Non-Markov Stochastic Process. *Cell systems* **5**, 268-282.e267 (2017).

27. A. H. Wong *et al.*, Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Human molecular genetics* **14 Spec No 1**, R11-18 (2005).

28. W. L. Tam, R. A. Weinberg, The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nature medicine* **19**, 1438-1449 (2013).

29. G. Keller, Embryonic stem cell differentiation: emergence of a new era in biology and medicine. *Genes & development* **19**, 1129-1155 (2005).

30. N. Kumar *et al.*, Stochastic modeling of phenotypic switching and chemoresistance in cancer cell populations. *Scientific reports* **9**, 10845 (2019).

31. A. O. Pisco *et al.*, Non-Darwinian dynamics in therapy-induced cancer drug resistance. *Nature communications* **4**, 2467 (2013).

32. G. Yang *et al.*, Dynamic equilibrium between cancer stem cells and non-stem cancer cells in human SW620 and MCF-7 cancer cell populations. *British journal of cancer* **106**, 1512-1519 (2012).

33. W. Wang *et al.*, Dynamics between cancer cell subpopulations reveals a model coordinating with both hierarchical and stochastic concepts. *PLoS One* **9**, e84654 (2014).

34. X. J. Tian *et al.*, Coupled reversible and irreversible bistable switches underlying TGFbeta-induced epithelial to mesenchymal transition. *Biophysical journal* **105**, 1079-1089 (2013).

35. M. Lu *et al.*, MicroRNA-based regulation of epithelial-hybrid-mesenchymal fate determination. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 18144-18149 (2013).

36. G. Moreno-Bueno *et al.*, The morphological and molecular features of the epithelial-to-mesenchymal transition. *Nature protocols* **4**, 1591-1613 (2009).

37. J. M. Lee *et al.*, The epithelial-mesenchymal transition: new insights in signaling, development, and disease. *The Journal of cell biology* **172**, 973-981 (2006).

38. D. Yao *et al.*, Mechanism of the mesenchymal-epithelial transition and its relationship with metastatic tumor formation. *Molecular cancer research : MCR* **9**, 1608-1620 (2011).

39. D. H. Kim *et al.*, Epithelial Mesenchymal Transition in Embryonic Development, Tissue Repair and Cancer: A Comprehensive Overview. *Journal of clinical medicine* **7**, 1-25 (2017).

40. J. Xu *et al.*, TGF-beta-induced epithelial to mesenchymal transition. *Cell research* **19**, 156-172 (2009).

41. I. G. Cannell *et al.*, How do microRNAs regulate gene expression? *Biochemical Society transactions* **36**, 1224-1231 (2008).

42. P. Ru *et al.*, miRNA-29b suppresses prostate cancer metastasis by regulating epithelial–mesenchymal transition signaling. *Molecular cancer therapeutics* **11**, 1166-1173 (2012).

43. J. Zhang *et al.*, miR-30 inhibits TGF-β1-induced epithelial-to-mesenchymal transition in hepatocyte by targeting Snail1. *Biochemical and biophysical research communications* **417**, 1100-1105 (2012).

44. P. A. Gregory *et al.*, The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nature cell biology* **10**, 593-601 (2008).

45. S. Brabletz, T. Brabletz, The ZEB/miR-200 feedback loop--a motor of cellular plasticity in development and cancer? *EMBO reports* **11**, 670-677 (2010).

46. H. Siemens *et al.*, miR-34 and SNAIL form a double-negative feedback loop to regulate epithelial-mesenchymal transitions. *Cell cycle (Georgetown, Tex.)* **10**, 4256-4271 (2011).

47. J. Zavadil, E. P. Bottinger, TGF-beta and epithelial-to-mesenchymal transitions. *Oncogene* **24**, 5764-5774 (2005).

48. F. M. Davis *et al.*, Non-stimulated, agonist-stimulated and store-operated Ca2+ influx in MDA-MB-468 breast cancer cells and the effect of EGF-induced EMT on calcium entry. *PLoS One* **7**, e36923 (2012).

49. F. M. Davis *et al.*, Assessment of gene expression of intracellular calcium channels, pumps and exchangers with epidermal growth factor-induced epithelial-mesenchymal transition in a breast cancer cell line. *Cancer cell international* **13**, 76 (2013).

50. F. M. Davis *et al.*, Induction of epithelial–mesenchymal transition (EMT) in breast cancer cells is calcium signal dependent. *Oncogene* **33**, 2307 (2014).

51. J. I. Yook *et al.*, A Wnt-Axin2-GSK3beta cascade regulates Snail1 activity in breast cancer cells. *Nature cell biology* **8**, 1398-1406 (2006).

52. B. P. Zhou *et al.*, Dual regulation of Snail by GSK-3beta-mediated phosphorylation in control of epithelial-mesenchymal transition. *Nature cell biology* **6**, 931-940 (2004).

53. S. Souchelnytskyi *et al.*, Phosphorylation of Ser165 in TGF-beta type I receptor modulates TGF-beta1-induced cellular responses. *The EMBO journal* **15**, 6231-6240 (1996).

54. Y. Su *et al.*, Single-cell analysis resolves the cell state transition and signaling dynamics associated with melanoma drug-induced resistance. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 13679-13684 (2017).

55. S. Ghuwalewala *et al.*, CD44(high)CD24(low) molecular signature determines the Cancer Stem Cell and EMT phenotype in Oral Squamous Cell Carcinoma. *Stem cell research* **16**, 405-417 (2016).

56. J. W. Armond *et al.*, A stochastic model dissects cell states in biological transition processes. *Scientific reports* **4**, 3692 (2014).

57. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32**, 381-386 (2014).

58. J. Chen *et al.*, Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. *Nature communications* **7**, 11988 (2016).

59. C. Sommer *et al.*, A deep learning and novelty detection framework for rapid phenotyping in high-content screening. *Molecular biology of the cell* **28**, 3428-3436 (2017).

60. F. Buggenthin *et al.*, Prospective identification of hematopoietic lineage choice by deep learning. *Nature methods* **14**, 403-406 (2017).

61. A. E. Carpenter *et al.*, CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology* **7**, R100 (2006).

62. J. C. Kimmel *et al.*, Inferring cell state by quantitative motility analysis reveals a dynamic state system and broken detailed balance. *PLoS computational biology* **14**, e1005927 (2018).

63. C. H. Waddington, *The strategy ofthe genes: a discussion of some aspects of theoretical biology*. (Allen & Unwin, London, 1957).

64. S. H. Strogatz, *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*. (CRC press, Boca Raton, 2018).

65. W. Jia *et al.*, A possible role for epigenetic feedback regulation in the dynamics of the epithelial-mesenchymal transition (EMT). *Physical biology* **16**, 066004 (2019).

66. M. K. Jolly, T. Celia-Terrassa, Dynamics of Phenotypic Heterogeneity Associated with EMT and Stemness during Cancer Progression. *Journal of clinical medicine* **8**, 1-19 (2019).

67. S. Tripathi *et al.*, A mechanism for epithelial-mesenchymal heterogeneity in a population of cancer cells. *PLoS computational biology* **16**, e1007619 (2020).

68. P. Yu *et al.*, Nanog induced intermediate state in regulating stem cell differentiation and reprogramming. *BMC systems biology* **12**, 22 (2018).

69. J. Wang *et al.*, The potential landscape of genetic circuits imposes the arrow of time in stem cell differentiation. *Biophysical journal* **99**, 29-39 (2010).

70. C. Li, J. Wang, Quantifying the underlying landscape and paths of cancer. *Journal of the Royal Society, Interface* **11**, 20140774 (2014).

71. J. Wang *et al.*, Quantifying the Waddington landscape and biological paths for development and differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 8257-8262 (2011).

72. C. Li *et al.*, Quantifying the landscape and kinetic paths for epithelial-mesenchymal transition from a core circuit. *Physical chemistry chemical physics : PCCP* **18**, 17949-17956 (2016).

73. K. Biswas *et al.*, Stability and mean residence times for hybrid epithelial/mesenchymal phenotype. *Physical biology* **16**, 025003 (2019).

74. L. J. Allen, *An introduction to stochastic processes with applications to biology*. (CRC Press, New York, 2010).

75. N. J. Higham *et al.*, On pth roots of stochastic matrices. *Linear Algebra and its Applications* **435**, 448-463 (2011).

76. T.-C. Lee *et al.*, *Estimating the parameters of the Markov probability model from aggregate time series data*. (North-Holland Publishing Company, Amsterdam, 1970).

77. I.-C. Chou, E. O. J. M. b. Voit, Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Mathematical Biosciences* **219**, 57-83 (2009).

78. A. Raue *et al.*, Data2Dynamics: a modeling environment tailored to parameter estimation in dynamical systems. *Bioinformatics (Oxford, England)* **31**, 3558-3560 (2015).

79. D. Zhou *et al.*, Population dynamics of cancer cells with cell state conversions. *Quantitative biology* **1**, 201-208 (2013).

80. K. Oda, H. Kitano, A comprehensive map of the toll-like receptor signaling network. *Molecular systems biology* **2**, 2006.0015 (2006).

81. K. Oda *et al.*, A comprehensive pathway map of epidermal growth factor receptor signaling. *Molecular systems biology* **1**, 2005.0010 (2005).

82. M. N. McClean *et al.*, Cross-talk and decision making in MAP kinase pathways. *Nature genetics* **39**, 409-414 (2007).

83. M. Natarajan *et al.*, A global analysis of cross-talk in a mammalian cellular signalling network. *Nature cell biology* **8**, 571-580 (2006).

84. A. Citri, Y. Yarden, EGF-ERBB signalling: towards the systems level. *Nature reviews. Molecular cell biology* **7**, 505-516 (2006).

85. T. A. Halsey *et al.*, A functional map of NFkappaB signaling identifies novel modulators and multiple system controls. *Genome biology* **8**, R104 (2007).

86. B. D. Manning, A. Toker, AKT/PKB Signaling: Navigating the Network. *Cell* **169**, 381-405 (2017).

87. G. Song *et al.*, The activation of Akt/PKB signaling pathway and cell survival. *Journal of cellular and molecular medicine* **9**, 59-71 (2005).

88. P. J. Murray, The JAK-STAT signaling pathway: input and output integration. *Journal of immunology (Baltimore, Md. : 1950)* **178**, 2623-2629 (2007).

89. J. E. Purvis, G. Lahav, Encoding and decoding cellular information through signaling dynamics. *Cell* **152**, 945-956 (2013).

90. S. Sasagawa *et al.*, Prediction and validation of the distinct dynamics of transient and sustained ERK activation. *Nature cell biology* **7**, 365-373 (2005).

91. U. Alon, Network motifs: theory and experimental approaches. *Nature reviews. Genetics* **8**, 450-461 (2007).

92. U. Alon, *An introduction to systems biology: design principles of biological circuits*. (CRC press, New York, 2019).

93. X. Guo, X. F. Wang, Signaling cross-talk between TGF-beta/BMP and other pathways. *Cell research* **19**, 71-88 (2009).

94. D. Javelaud, A. Mauviel, Crosstalk mechanisms between the mitogen-activated protein kinase pathways and Smad signaling downstream of TGF-beta: implications for carcinogenesis. *Oncogene* **24**, 5742-5750 (2005).

95. M. A. Schwartz, V. Baron, Interactions between mitogenic stimuli, or, a thousand and one connections. *Current opinion in cell biology* **11**, 197-202 (1999).

96. R. Linding, Multivariate signal integration. *Nature reviews. Molecular cell biology* **11**, 391 (2010).

97. J. A. Lamb *et al.*, JunD mediates survival signaling by the JNK signal transduction pathway. *Molecular cell* **11**, 1479-1489 (2003).

98. K. Lei, R. J. Davis, JNK phosphorylation of Bim-related members of the Bcl2 family induces Bax-dependent apoptosis. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 2432-2437 (2003).

99. M. T. Abreu-Martin *et al.*, Fas activates the JNK pathway in human colonic epithelial cells: lack of a direct role in apoptosis. *The American journal of physiology* **276**, G599-605 (1999).

100. K. A. Janes *et al.*, A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science (New York, N.Y.)* **310**, 1646-1653 (2005).

101. R. C. Hsueh *et al.*, Deciphering signaling outcomes from a system of complex networks. *Science signaling* **2**, ra22-ra22 (2009).

102. J. E. Ladbury, S. T. J. T. i. b. s. Arold, Noise in cellular signaling pathways: causes and effects. *Trends in biochemical sciences* **37**, 173-178 (2012).

103. N. Domedel-Puig *et al.*, Information routing driven by background chatter in a signaling network. *PLoS computational biology* **7**, e1002297 (2011).

104. M. A. Suni, V. C. Maino, Flow cytometric analysis of cell signaling proteins. *Methods in molecular biology (Clifton, N.J.)* **717**, 155-169 (2011).

105. Q. Ni *et al.*, Analyzing protein kinase dynamics in living cells with FRET reporters. *Methods (San Diego, Calif.)* **40**, 279-286 (2006).

106. C. Cohen-Saidon *et al.*, Dynamics and variability of ERK2 response to EGF in individual living cells. *Molecular cell* **36**, 885-893 (2009).

107. L. Oldach, J. Zhang, Genetically encoded fluorescent biosensors for live-cell visualization of protein phosphorylation. *Chemistry & biology* **21**, 186-197 (2014).

108. M. D. Allen, J. Zhang, Subcellular dynamics of protein kinase A activity visualized by FRET-based reporters. *Biochemical and biophysical research communications* **348**, 716-721 (2006).

109. C. E. Shannon, A mathematical theory of communication. *Bell System Technical Journal* **27**, 379-423 (1948).

110. T. M. Cover, J. A. Thomas, *Elements of Information Theory*. ( John Wiley & Sons, New York, 1991), vol. 68, pp. 69-73.

111. R. Suderman *et al.*, Fundamental trade-offs between information flow in single cells and cellular populations. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 5755-5760 (2017).

112. R. Cheong *et al.*, Information transduction capacity of noisy biochemical signaling networks. *Science (New York, N.Y.)* **334**, 354-358 (2011).

113. Z. Mousavian *et al.*, Information theory in systems biology. Part II: protein-protein interaction and signaling networks. *Seminars in cell & developmental biology* **51**, 14-23 (2016).

114. Q. Zhang *et al.*, NF-κB Dynamics Discriminate between TNF Doses in Single Cells. *Cell systems* **5**, 638-645.e635 (2017).

115. A. Kraskov *et al.*, Estimating mutual information. *Physical review. E, Statistical, nonlinear, and soft matter physics* **69**, 066138 (2004).

116. G. D. Potter *et al.*, Communication shapes sensory response in multicellular networks. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 10334-10339 (2016).

117. M. Voliotis *et al.*, Information transfer by leaky, heterogeneous, protein kinase signaling systems. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E326-333 (2014).

118. K. L. Garner *et al.*, Information Transfer in Gonadotropin-releasing Hormone (GnRH) Signaling: EXTRACELLULAR SIGNAL-REGULATED KINASE (ERK)-MEDIATED FEEDBACK LOOPS CONTROL HORMONE SENSING. *The Journal of biological chemistry* **291**, 2246-2259 (2016).

119. B. W. Andrews, P. A. Iglesias, An information-theoretic characterization of the optimal gradient sensing response of cells. *PLoS computational biology* **3**, e153 (2007).

120. C. G. Bowsher, P. S. Swain, Identifying sources of variation and the flow of information in biochemical networks. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E1320-1328 (2012).

121. P. Mehta *et al.*, Information processing and signal integration in bacterial quorum sensing. *Molecular systems biology* **5**, 325 (2009).

122. R. Ruiz *et al.*, Negative feedback increases information transmission, enabling bacteria to discriminate sublethal antibiotic concentrations. *Science advances* **4**, eaat5771 (2018).

123. J. O. Dubuis *et al.*, Positional information, in bits. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 16301-16308 (2013).

124. J. Selimkhanov *et al.*, Systems biology. Accurate information transmission through dynamic biochemical signaling networks. *Science (New York, N.Y.)* **346**, 1370-1373 (2014).

125. A. Levchenko, I. Nemenman, Cellular noise and information transmission. *Current opinion in biotechnology* **28**, 156-164 (2014).

126. C. G. Bowsher, P. S. Swain, Environmental sensing, information transfer, and cellular decision-making. *Current opinion in biotechnology* **28**, 149-155 (2014).

127. R. Suderman, E. J. Deeds, Intrinsic limits of information transmission in biochemical signalling motifs. *Interface focus* **8**, 20180039 (2018).

128. A. Dionisio *et al.*, Mutual information: a measure of dependency for nonlinear time series. *Physica A: Statistical Mechanics and its Applications* **344**, 326-329 (2004).

129. T. Speed, Mathematics. A correlation for the 21st century. *Science (New York, N.Y.)* **334**, 1502-1503 (2011).

130. E. H. J. I. Linfoot, control, An informational measure of correlation. *Information and Control* **1**, 85-89 (1957).

131. I. H. Witten *et al.*, *Data Mining: Practical Machine Learning Tools and Techniques*. (Morgan Kaufmann Publishers Inc., 2011).

132. S. Hendry *et al.*, Assessing Tumor-infiltrating Lymphocytes in Solid Tumors: A Practical Review for Pathologists and Proposal for a Standardized Method From the International Immunooncology Biomarkers Working Group: Part 1: Assessing the Host Immune Response, TILs in Invasive Breast Carcinoma and Ductal Carcinoma In Situ, Metastatic Tumor Deposits and Areas for Further Research. *Advances in anatomic pathology* **24**, 235-251 (2017).

133. A. C. Dudley, Tumor endothelial cells. *Cold Spring Harbor perspectives in medicine* **2**, a006536 (2012).

134. K. Hida *et al.*, Contribution of Tumor Endothelial Cells in Cancer Progression. *International journal of molecular sciences* **19**, 12 (2018).

135. A. Frishberg *et al.*, ImmQuant: a user-friendly tool for inferring immune cell-type composition from gene-expression data. *Bioinformatics (Oxford, England)* **32**, 3842-3843 (2016).

136. P. Lu *et al.*, Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 10370-10375 (2003).

137. A. Kuhn *et al.*, Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nature methods* **8**, 945-947 (2011).

138. H. Lahdesmaki *et al.*, In silico microdissection of microarray data from heterogeneous cell populations. *BMC bioinformatics* **6**, 54 (2005).

139. M. M. Babu, Introduction to microarray data analysis. *Computational genomics: Theory and application*, 225-249 (2004).

140. Y. Zhong, Z. Liu, Gene expression deconvolution in linear space. *Nature methods* **9**, 8-9; author reply 9 (2011).

141. W. Strober, Trypan blue exclusion test of cell viability. *Current protocols in immunology* **21**, A.3B.1-A.3B.2 (2001).

142. P. S. Aranda *et al.*, Bleach gel: a simple agarose gel for analyzing RNA quality. *Electrophoresis* **33**, 366-369 (2012).

143. C. Ramakers *et al.*, Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neuroscience letters* **339**, 62-66 (2003).

144. M. Kubista *et al.*, The real-time polymerase chain reaction. *Molecular aspects of medicine* **27**, 95-125 (2006).

145. K. J. Livak, T. D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods (San Diego, Calif.)* **25**, 402-408 (2001).

146. D. Dao *et al.*, CellProfiler Analyst: interactive data exploration, analysis and classification of large biological image sets. *Bioinformatics (Oxford, England)* **32**, 3210-3212 (2016).

147. O. H. Lowry *et al.*, Protein measurement with the Folin phenol reagent. *The Journal of biological chemistry* **193**, 265-275 (1951).

148. T. Maniatis *et al.*, *Molecular cloning: a laboratory manual*. (Cold spring harbor laboratory Cold Spring Harbor, NY, 1982), vol. 545.

149. C. A. Schneider *et al.*, NIH Image to ImageJ: 25 years of image analysis. *Nature methods* **9**, 671-675 (2012).

150. C. Riccardi, I. Nicoletti, Analysis of apoptosis by propidium iodide staining and flow cytometry. *Nature protocols* **1**, 1458-1461 (2006).

151. T. Kawamoto *et al.*, Growth stimulation of A431 cells by epidermal growth factor: identification of high-affinity receptors for epidermal growth factor by an anti-receptor monoclonal antibody. *Proceedings of the National Academy of Sciences of the United States of America* **80**, 1337-1341 (1983).

152. T. Kawamoto *et al.*, Relation of epidermal growth factor receptor concentration to growth of human epidermoid carcinoma A431 cells. *The Journal of biological chemistry* **259**, 7761-7766 (1984).

153. D. K. Armstrong *et al.*, Epidermal growth factor-mediated apoptosis of MDA-MB-468 human breast cancer cells. *Cancer research* **54**, 5280-5283 (1994).

154. P. Schaerli, R. Jaggi, EGF-induced programmed cell death of human mammary carcinoma MDA-MB-468 cells is preceded by activation AP-1. *Cellular and molecular life sciences : CMLS* **54**, 129-138 (1998).

155. K. Y. Chen *et al.*, The role of tyrosine kinase Etk/Bmx in EGF-induced apoptosis of MDA-MB-468 breast cancer cells. *Oncogene* **23**, 1854-1862 (2004).

156. W. A. Dengler *et al.*, Development of a propidium iodide fluorescence assay for proliferation and cytotoxicity assays. *Anti-cancer drugs* **6**, 522-532 (1995).

157. C. P. Wan *et al.*, A simple fluorometric assay for the determination of cell numbers. *Journal of immunological methods* **173**, 265-272 (1994).

158. T. Mosmann, Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays. *Journal of immunological methods* **65**, 55-63 (1983).

159. S. Jang *et al.*, Dynamics of embryonic stem cell differentiation inferred from single-cell transcriptomics show a series of transitions through discrete cell states. *eLife* **6**, 28 (2017).

160. C. Marr *et al.*, Single-cell gene expression profiling and cell state dynamics: collecting data, correlating data points and connecting the dots. *Current opinion in biotechnology* **39**, 207-214 (2016).

161. M. Mojtahedi *et al.*, Cell Fate Decision as High-Dimensional Critical State Transition. *PLoS biology* **14**, e2000640 (2016).

162. K. J. Chavez *et al.*, Triple negative breast cancer cell lines: one tool in the search for better treatment of triple negative breast cancer. *Breast disease* **32**, 35-48 (2010).

163. R. M. Neve *et al.*, A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer cell* **10**, 515-527 (2006).

164. A. Bonnomet *et al.*, A dynamic in vivo model of epithelial-to-mesenchymal transitions in circulating tumor cells and metastases of breast cancer. *Oncogene* **31**, 3741 (2012).

165. H. W. Lo *et al.*, Epidermal growth factor receptor cooperates with signal transducer and activator of transcription 3 to induce epithelial-mesenchymal transition in cancer cells via up-regulation of TWIST gene expression. *Cancer research* **67**, 9066-9076 (2007).

166. N. Verma *et al.*, PYK2 sustains endosomal-derived receptor signalling and enhances epithelial-to-mesenchymal transition. *Nature communications* **6**, 6064 (2015).

167. M. Wesseling *et al.*, The morphological and molecular mechanisms of epithelial/endothelial-to-mesenchymal transition and its involvement in atherosclerosis. *Vascular Pharmacology* **106**, 1-8 (2018).

168. F. M. Davis *et al.*, Remodeling of purinergic receptor-mediated Ca2+ signaling as a consequence of EGF-induced epithelial-mesenchymal transition in breast cancer cells. *PLoS One* **6**, e23464 (2011).

169. T. F. Coleman, Y. Li, A Reflective Newton Method for Minimizing a Quadratic Function Subject to Bounds on Some of the Variables. *SIAM Journal on Optimization* **6**, 1040-1058 (1992).

170. T. Chiba *et al.*, Time-Varying Transition Probability Matrix Estimation and Its Application to Brand Share Analysis. *PLoS One* **12**, e0169981 (2017).

171. K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*. (John Wiley & Sons, Inc., USA, 2001).

172. K. Deb, S. Gupta, Understanding knee points in bicriteria problems and their implications as preferred solution principles. *Engineering Optimization* **43**, 1175-1204 (2011).

173. K. Matsumoto *et al.*, Growth factor regulation of integrin-mediated cell motility. *Cancer metastasis reviews* **14**, 205-217 (1995).

174. C. E. Turner, Paxillin and focal adhesion signalling. *Nature cell biology* **2**, E231-236 (2000).

175. E. A. Clark, J. S. Brugge, Integrins and signal transduction pathways: the road taken. *Science (New York, N.Y.)* **268**, 233-239 (1995).

176. Z. Lu *et al.*, Epidermal growth factor-induced tumor cell invasion and metastasis initiated by dephosphorylation and downregulation of focal adhesion kinase. *Molecular and cellular biology* **21**, 4016-4031 (2001).

177. B. N. Kholodenko *et al.*, Quantification of information transfer via cellular signal transduction pathways. *FEBS letters* **414**, 430-434 (1997).

178. Q. Zhang *et al.*, Ultrasensitive response motifs: basic amplifiers in molecular signalling networks. *Open biology* **3**, 130031 (2013).

179. A. Goldbeter, D. E. Koshland, Sensitivity amplification in biochemical systems. *Quarterly reviews of biophysics* **15**, 555-591 (1982).

180. T. Hong *et al.*, An Ovol2-Zeb1 Mutual Inhibitory Circuit Governs Bidirectional and Multi-step Transition between Epithelial and Mesenchymal States. *PLoS computational biology* **11**, e1004569 (2015).

181. M. K. Jolly *et al.*, Stability of the hybrid epithelial/mesenchymal phenotype. *Oncotarget* **7**, 27067-27084 (2016).

182. M. K. Jolly *et al.*, Epithelial/mesenchymal plasticity: how have quantitative mathematical models helped improve our understanding? *Molecular oncology* **11**, 739-754 (2017).

183. A. Hollestelle *et al.*, Loss of E-cadherin is not a necessity for epithelial to mesenchymal transition in human breast cancer. *Breast cancer research and treatment* **138**, 47-57 (2013).

184. G. M. Nilsson *et al.*, Loss of E-cadherin expression is not a prerequisite for c-erbB2-induced epithelial-mesenchymal transition. *International journal of oncology* **45**, 82-94 (2014).

185. J. Wang *et al.*, Potential and flux landscapes quantify the stability and robustness of budding yeast cell cycle network. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 8195-8200 (2010).

186. H. H. Chang *et al.*, Multistable and multistep dynamics in neutrophil differentiation. *BMC cell biology* **7**, 11 (2006).

187. T. Eissing *et al.*, Bistability analyses of a caspase activation model for receptor-induced apoptosis. *The Journal of biological chemistry* **279**, 36892-36897 (2004).

188. W. Sha *et al.*, Hysteresis drives cell-cycle transitions in Xenopus laevis egg extracts. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 975-980 (2003).

189. T. Celia-Terrassa *et al.*, Hysteresis control of epithelial-mesenchymal transition dynamics conveys a distinct program with enhanced metastatic ability. *Nature communications* **9**, 5005 (2018).

190. C. Y. Huang, J. E. Ferrell, Jr., Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 10078-10083 (1996).

191. G. J. Melen *et al.*, Threshold responses to morphogen gradients by zero-order ultrasensitivity. *Molecular systems biology* **1**, 2005 0028 (2005).

192. J. Song, EMT or apoptosis: a decision for TGF-beta. *Cell research* **17**, 289-290 (2007).

193. Y. Yang *et al.*, Transforming growth factor-beta1 induces epithelial-to-mesenchymal transition and apoptosis via a cell cycle-dependent mechanism. *Oncogene* **25**, 7235-7244 (2006).

194. T. M. Cover, J. A. Thomas, *Elements of Information Theory*. (John Wiley & Sons, Hoboken, 2012).

195. R. H. Byrd *et al.*, A trust region method based on interior point techniques for nonlinear programming. *Mathematical Programming* **89**, 149-185 (2000).

196. V. J. Thannickal *et al.*, Myofibroblast differentiation by transforming growth factor-beta1 is dependent on cell adhesion and integrin signaling via focal adhesion kinase. *The Journal of biological chemistry* **278**, 12384-12389 (2003).

197. C. Kroger *et al.*, Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 7353-7362 (2019).

198. M. K. Jolly *et al.*, Hybrid epithelial/mesenchymal phenotypes promote metastasis and therapy resistance across carcinomas. *Pharmacology & therapeutics* **194**, 161-184 (2019).

199. J. Kruschke, *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. (Academic Press, 2014).

200. A. Raue *et al.*, Addressing parameter identifiability by model-based experimentation. *IET systems biology* **5**, 120-130 (2011).

201. A. Raue *et al.*, Lessons learned from quantitative dynamical modeling in systems biology. *PLoS One* **8**, e74335 (2013).

202. A. Gelman *et al.*, *Bayesian Data Analysis, Third Edition*. (Taylor & Francis, 2013).

203. K. P. Murphy, Conjugate Bayesian analysis of the Gaussian distribution. *def* **1**, 16 (2007).

204. M. Clyde *et al.*, *An Introduction to Bayesian Thinking-A Companion to the Statistics with R Course*. (GitHub, GitHub repository).

205. M. Betancourt, A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434*, 60 (2018).

206. M. D. Hoffman, A. J. J. o. M. L. R. Gelman, The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15**, 1593-1623 (2014).

207. J. Salvatier *et al.*, Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* **2**, e55 (2016).

208. J. Lorah, A. Womack, Value of sample size for computation of the Bayesian information criterion (BIC) in multilevel modeling. *Behavior research methods* **51**, 440-450 (2019).

209. S. I. Vrieze, Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods* **17**, 228-243 (2012).

210. R. Milo *et al.*, BioNumbers--the database of key numbers in molecular and cell biology. *Nucleic acids research* **38**, D750-753 (2010).

211. N. E. Davidson *et al.*, Epidermal growth factor receptor gene expression in estrogen receptor-positive and negative human breast cancer cell lines. *Molecular endocrinology (Baltimore, Md.)* **1**, 216-223 (1987).

212. M. Zeisberg, E. G. Neilson, Biomarkers for epithelial-mesenchymal transitions. *The Journal of clinical investigation* **119**, 1429-1437 (2009).

# Publications

- Morphological State Transition Dynamics in EGF-induced Epithelial to Mesenchymal Transition. Devaraj V., Bose B. Journal of clinical medicine, 2019, 8(7). pii: E911.

- The Mathematics of Phenotypic State Transition: Paths and Potential. Devaraj V., Bose B. Journal of the Indian Institute of Science, 2020. DOI: 10.1007/s41745-020-00173-6.

- DEBay: a computational tool for deconvolution of quantitative PCR data for estimation of cell type-specific gene expression in a mixed population. Devaraj V., Bose B. [under review] [preprint in BioRxiv https://doi.org/10.1101/2020.04.10.035642].

# Presentations

- Oral presentation on Signal Discrimination through a Negative Feedback. Devaraj V., Dutta P., Bose B. RTG Big Data Research Summer School, April 2019, Allahabad University, India.

- Poster presentation on Understanding the Cellular State Transition in Epithelial to Mesenchymal Transition. Devaraj V., Bose B. 1st IBSE International Symposium, January 2018, IIT Madras, India.

- Poster presentation on Characterization of a Negative Feedback in PI3K/Akt pathway. Devaraj V., Dutta P., Bose B. International Symposium on Systems, Synthetic and Chemical Biology, December 2017, Bose Institute, Kolkata, India.

# Appendix A

## Section A-1: Estimation of percentage dead cell population

We have measured the fold change in dead cell number as well as the fold change in total cell number using propidium iodide (fluorescence-based plate reader assay). The experiment was performed form time = 0 till 60 h at an interval of 12 h. From this data, we calculated the percentage of the dead cell population.

The fold change in dead cells at time, $t + \Delta t$ is,

$$FD_{t+\Delta t} = \frac{D_{t+\Delta t}}{D_t}$$

where $D_t$ and $D_{t+\Delta t}$ are the number of dead cells at $t$ and $t + \Delta t$ respectively.

$$FD_{t+\Delta t} = \frac{T_{t+\Delta t} \times (\%\ dead\ cells)_{t+\Delta t}}{T_t \times (\%\ dead\ cells)_t}$$

where $D_{t+\Delta t} = T_{t+\Delta t} \times (\%\ dead\ cells)_{t+\Delta t}$ and $D_t = T_t \times (\%\ dead\ cells)_t$. $T_t$ and $T_{t+\Delta t}$ are the total number of cells (live + dead) at $t$ and $t + \Delta t$ respectively.

$$FD_{t+\Delta t} = \frac{FT_{t+\Delta t} \times (\%\ dead\ cells)_{t+\Delta t}}{(\%\ dead\ cells)_t}$$

where $FT_{t+\Delta t} = \frac{T_{t+\Delta t}}{T_t}$. $FT_{t+\Delta t}$ is the fold change in the total number of cells at $t + \Delta t$.

$$(\%\ dead\ cells)_{t+\Delta t} = \left(\frac{FD_{t+\Delta t}}{FT_{t+\Delta t}}\right) \times (\%\ dead\ cells)_t$$

From the above equation, the percentage of dead cells was estimated at each time point. $FD_{t+\Delta t}$ and $FT_{t+\Delta t}$ were measured experimentally through the PI-based microplate reader assay. The percentage of dead cells at time = 0, was estimated through trypan blue staining. $(\%\ dead\ cells)_{t=0}$~14 %.

**Table A-1: The fractional cell division values estimated from the state transition model.**

| *Time interval (h)* | | *Untreated* | *1 ng/mL EGF* | *5 ng/mL EGF* | *10 ng/mL EGF* | *25 ng/mL EGF* |
|---|---|---|---|---|---|---|
| 0-12 | $q_{CB}$ | 0.174 | 0.198 | 0.093 | 0.023 | 0 |
| | $q_{ES}$ | 0.174 | 0.198 | 0.093 | 0.023 | 0 |
| | $q_{CR}$ | 0.174 | 0.198 | 0.093 | 0.023 | 0 |
| 12-24 | $q_{CB}$ | 0.139 | 0.127 | 0.164 | 0.095 | 0.024 |
| | $q_{ES}$ | 0.139 | 0.127 | 0.165 | 0.095 | 0.024 |
| | $q_{CR}$ | 0.139 | 0.127 | 0.164 | 0.095 | 0.024 |
| 24-36 | $q_{CB}$ | 0.082 | 0.059 | 0.024 | 0.087 | 0.051 |
| | $q_{ES}$ | 0.082 | 0.059 | 0.024 | 0.087 | 0.051 |
| | $q_{CR}$ | 0.082 | 0.059 | 0.024 | 0.087 | 0.051 |
| 36-48 | $q_{CB}$ | 0 | 0 | 0.012 | 0.012 | 0 |
| | $q_{ES}$ | 0 | 0 | 0.012 | 0.012 | 0 |
| | $q_{CR}$ | 0 | 0 | 0.012 | 0.012 | 0 |
| 48-60 | $q_{CB}$ | 0.012 | 0.037 | 0.025 | 0 | 0 |
| | $q_{ES}$ | 0.012 | 0.037 | 0.025 | 0 | 0 |
| | $q_{CR}$ | 0.012 | 0.037 | 0.025 | 0 | 0 |

**Table A-2: The fractional state transition values of 5 ng/mL EGF treated samples estimated from the state transition model.**

| ***Cell states (0-12 h)*** | CB | ES | CR | DD |
|---|---|---|---|---|
| CB | 0.14 | 0.06 | 0.8 | 0 |
| ES | 0.31 | 0.42 | 0.13 | 0.14 |
| CR | 0.63 | 0.26 | 0.05 | 0.06 |
| ***Cell states (12-24 h)*** | CB | ES | CR | DD |
| CB | 0.3 | 0.15 | 0.32 | 0.23 |
| ES | 0.32 | 0.43 | 0.12 | 0.13 |
| CR | 0.66 | 0.32 | 0.02 | 0 |
| ***Cell states (24-36 h)*** | CB | ES | CR | DD |
| CB | 0.81 | 0.14 | 0.04 | 0.01 |
| ES | 0.34 | 0.43 | 0.12 | 0.11 |
| CR | 0.77 | 0.2 | 0.02 | 0.01 |
| ***Cell states (36-48 h)*** | CB | ES | CR | DD |
| CB | 0.84 | 0.12 | 0.04 | 0 |
| ES | 0.34 | 0.44 | 0.14 | 0.08 |
| CR | 0.77 | 0.19 | 0.03 | 0.01 |
| ***Cell states (48-60 h)*** | CB | ES | CR | DD |
| CB | 0.87 | 0.08 | 0.04 | 0.01 |
| ES | 0.35 | 0.45 | 0.17 | 0.03 |
| CR | 0.73 | 0.2 | 0.03 | 0.04 |

**Table A-3: The fractional state transition values of 10 ng/mL EGF treated samples estimated from the state transition model.**

| ***Cell states (0-12 h)*** | CB | ES | CR | DD |
|---|---|---|---|---|
| CB | 0.07 | 0.07 | 0.86 | 0 |
| ES | 0.5 | 0.25 | 0.11 | 0.14 |
| CR | 0.32 | 0.09 | 0.5 | 0.09 |
| ***Cell states (12-24 h)*** | CB | ES | CR | DD |
| CB | 0.07 | 0.11 | 0.77 | 0.05 |
| ES | 0.5 | 0.26 | 0.11 | 0.13 |
| CR | 0.31 | 0.1 | 0.56 | 0.03 |
| ***Cell states (24-36 h)*** | CB | ES | CR | DD |
| CB | 0.56 | 0.18 | 0.11 | 0.15 |
| ES | 0.65 | 0.26 | 0.04 | 0.05 |
| CR | 0.61 | 0.3 | 0.07 | 0.02 |
| ***Cell states (36-48 h)*** | CB | ES | CR | DD |
| CB | 0.7 | 0.18 | 0.11 | 0.01 |
| ES | 0.65 | 0.26 | 0.04 | 0.05 |
| CR | 0.55 | 0.28 | 0.08 | 0.09 |
| ***Cell states (48-60 h)*** | CB | ES | CR | DD |
| CB | 0.78 | 0.12 | 0.09 | 0.01 |
| ES | 0.62 | 0.27 | 0.04 | 0.07 |
| CR | 0.51 | 0.27 | 0.08 | 0.14 |

**Table A-4: The fractional state transition values of 25 ng/mL EGF treated samples estimated from the state transition model.**

| ***Cell states (0-12 h)*** | CB | ES | CR | DD |
|---|---|---|---|---|
| CB | 0.07 | 0.09 | 0.81 | 0.03 |
| ES | 0.07 | 0.14 | 0.56 | 0.23 |
| CR | 0.1 | 0.17 | 0.7 | 0.03 |
| ***Cell states (12-24 h)*** | CB | ES | CR | DD |
| CB | 0.04 | 0.04 | 0.91 | 0.01 |
| ES | 0.07 | 0.14 | 0.56 | 0.23 |
| CR | 0.08 | 0.15 | 0.75 | 0.02 |
| ***Cell states (24-36 h)*** | CB | ES | CR | DD |
| CB | 0.04 | 0.03 | 0.91 | 0.01 |
| ES | 0.08 | 0.29 | 0.38 | 0.25 |
| CR | 0.11 | 0.14 | 0.74 | 0.01 |
| ***Cell states (36-48 h)*** | CB | ES | CR | DD |
| CB | 0.07 | 0.05 | 0.75 | 0.13 |
| ES | 0.06 | 0.46 | 0.3 | 0.19 |
| CR | 0.09 | 0.09 | 0.79 | 0.02 |
| ***Cell states (48-60 h)*** | CB | ES | CR | DD |
| CB | 0.09 | 0.05 | 0.74 | 0.12 |
| ES | 0.05 | 0.49 | 0.24 | 0.22 |
| CR | 0.15 | 0.15 | 0.69 | 0 |

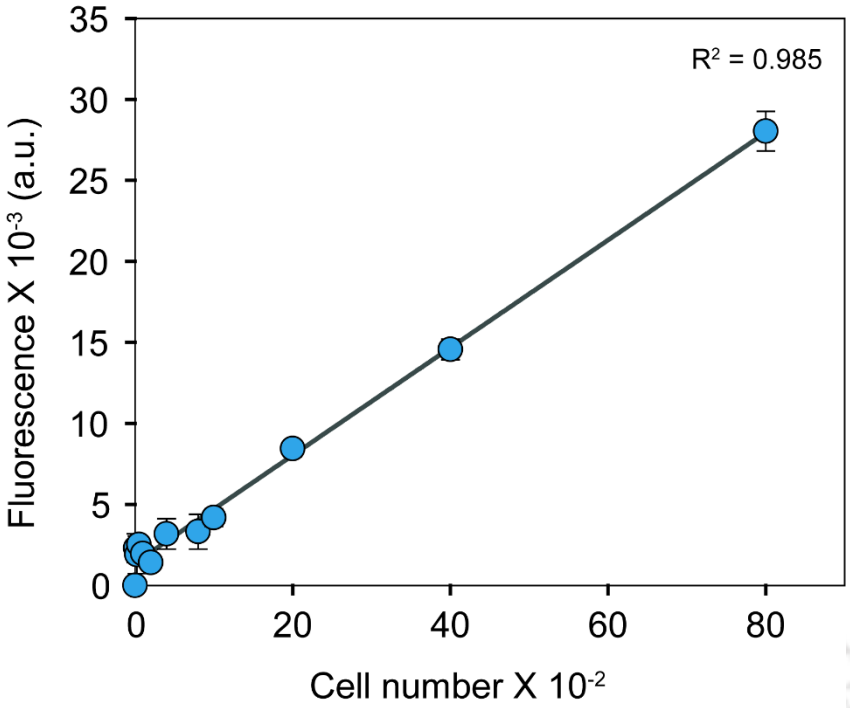



**Figure A-1: Standard curve of dead cell number estimation assay.**
The number of dead cells was estimated using a fluorescence-based plate reader assay. Cells were seeded in various density from 10 to 8000 cells. Once the cells have adhered, staining solution (final concentration: 1 µg/mL PI, 0.5 % Triton X-100) was added to the cells without removing the media. Triton X-100 permeabilizes the cell, and therefore, propidium iodide (PI) can bind to the double-stranded DNA of the cells. The amount of fluorescence from PI was considered as a measure of the DNA content, which in turn corresponds to the dead cells. The cells were incubated at room temperature for 3 h, followed by fluorescence measurement at $\lambda_{ex} = 530\ nm$ and $\lambda_{em} = 620\ nm$. The fluorescence increased linearly with the increase in cell number.

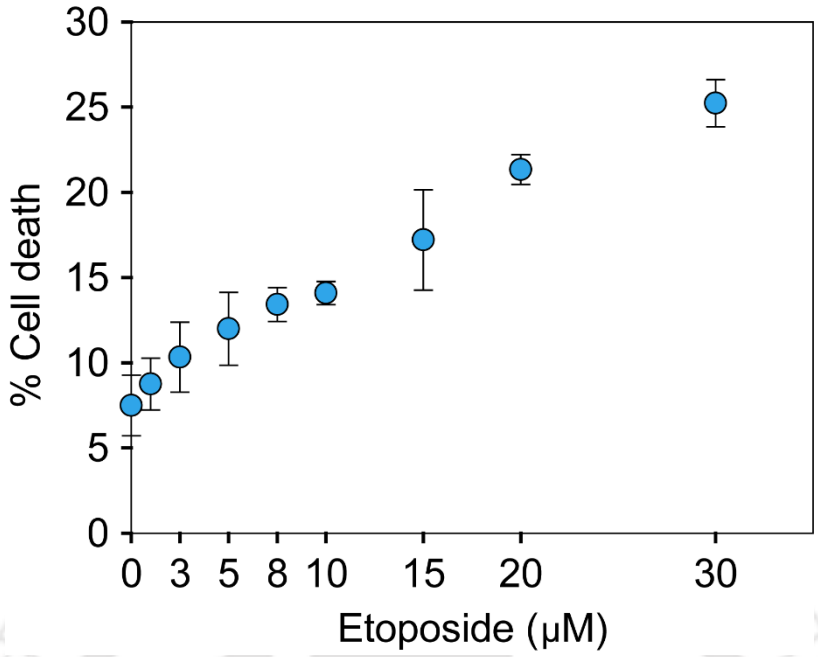



**Figure A-2: Quality check of the dead cell number estimation assay.**
The number of dead cells was estimated using a fluorescence-based plate reader assay. The cell membrane of dead cells will be compromised. Propidium iodide (PI) can penetrate into the cells and bind to the double-stranded DNA. The amount of fluorescence from PI is considered as a measure of the DNA content, which in turn corresponds to the number of dead cells. The cells were seeded in 96 well plates. The cells were treated with various doses of etoposide for 24 h. Few extra wells containing cells were fixed with 100 % methanol, to make sure 100 % cell death. These cells were used as control. After 24 h of etoposide treatment, 1 µg/mL PI (final concentration) was added to each well without removing the media. The cells were incubated at 37 °C for 10 min., followed by fluorescence measurement at $\lambda_{ex} = 530\,nm$ and $\lambda_{em} = 620\,nm$. The percentage of cell death was calculated with respect to the methanol fixed cells.

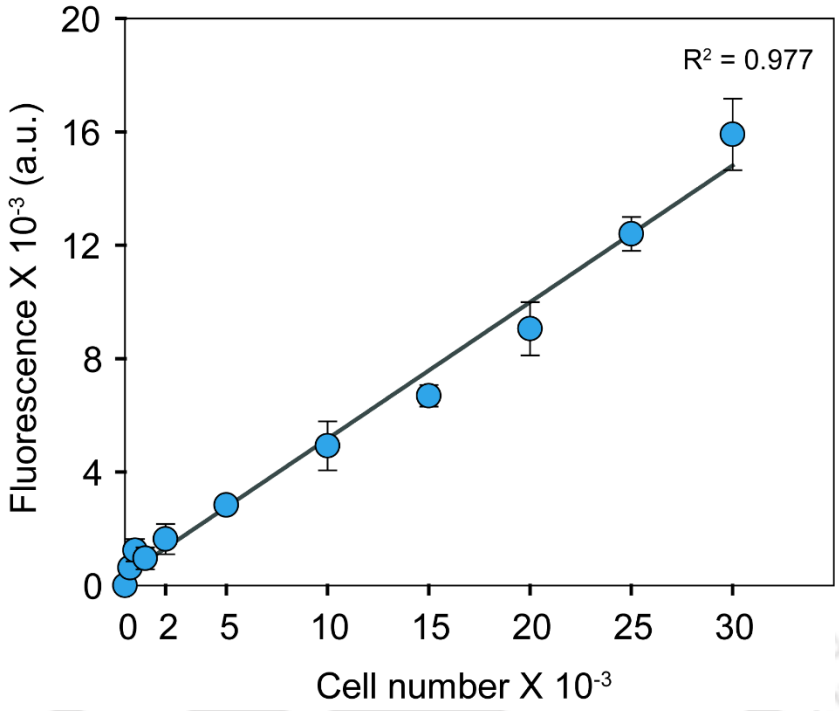



**Figure A-3: Standard curve of total cell number estimation assay.**
The total number of cells (live + dead) was estimated using a fluorescence-based plate reader assay. Cells were seeded in various density from 250 to 30000 cells. Once the cells have adhered, staining solution (final concentration: 30 µg/mL PI, 0.5 % Triton X-100, and 0.01 M EDTA) was added to the cells without removing the media. Triton X-100 permeabilizes the cell, and therefore, propidium iodide (PI) can bind to the double-stranded DNA of the cells. The amount of fluorescence from PI was considered as a measure of the DNA content, which in turn corresponds to the total number of cells (live + dead). The cells were incubated at room temperature for 6 h, followed by fluorescence measurement at $\lambda_{ex} = 530\ nm$ and $\lambda_{em} = 620\ nm$. The fluorescence increased linearly with the increase in cell number.

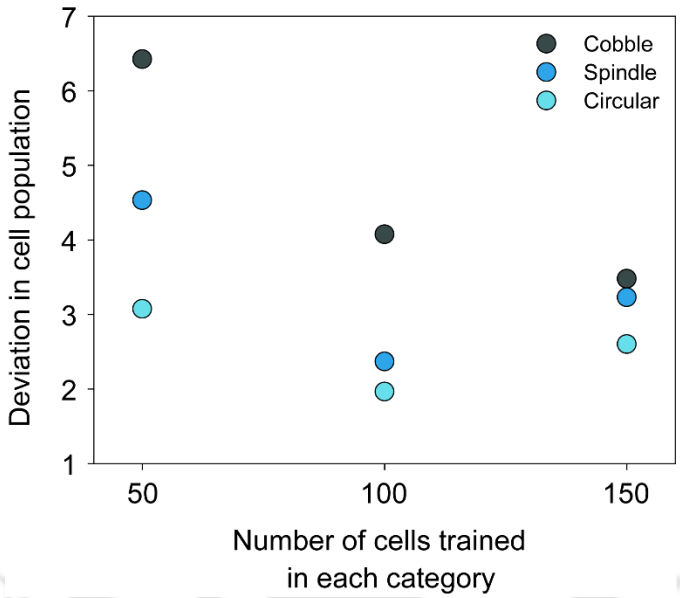



**Figure A-4: Optimal number of training objects per cell type.**
The cells were treated with different doses of EGF, such that all different cell types (Cobble, Spindle, and Circular) were observed. The cells were imaged and were analyzed using CellProfiler Analyst. The cells were classified into Cobble, Spindle, and Circular manually by the user with varying numbers of cells in each category. The algorithm was trained in this fashion, and the same procedure was repeated five times independently on the same data set. A set of rules was obtained at the end of each training. These five different rules were used to classify an entirely new set of images that were not used in training. The plot shows the standard deviation of the percentage cell population estimated from the five-independent training. The average of the standard deviation across all cell types was minimum in the case of 100 trained cells in each cell category. Therefore, in all subsequent experiments, 100 cells were trained in each cell category.

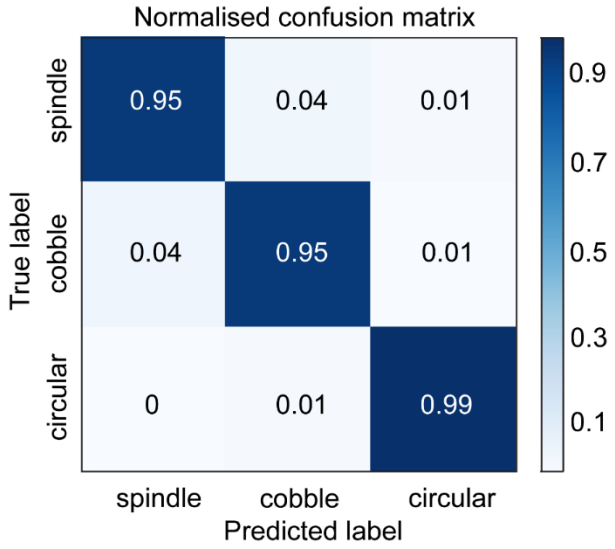



**Figure A-5: Quality check of the image classifier training.**
The MDA-MB-468 cells were categorized into three cell types based on their morphology. The cells were imaged, and the distribution of the three cell types was quantitatively estimated using image analysis tools. The tools employ a machine learning approach to classify the cells. Initially, the algorithm was trained by the user with a set of images. Then, the test samples were classified based on the training. The plot shows a sample confusion matrix of the image classifier training. The true label represents the actual cell type defined by the user, and the predicted label represents the cell type categorized by the classifier. The efficiency of the training was scored from 0 to 1. A value of 1 shows that all the cells were rightly classified by the classifier as defined by the user. The diagonal elements are the fractions of cells that were correctly classified by the classifier. The non-diagonal elements are the fractions of false-positive cells.

# Appendix B

## Section B-1: Reagents used in cell culture

### Basal media

19.5 g of DMEM with high glucose (Himedia) was dissolved in 900 mL of autoclaved double distilled water. 3.7 g of sodium bicarbonate was added and stirred until dissolved. The total volume was made to 1L with double distilled water. The media was sterilized using a vacuum filtration unit with a 0.2 μm pore size. The basal media was stored at 2-4 °C.

### Complete growth media

Basal media was supplemented with 10 % fetal bovine serum (FBS, GIBCO) and 1x Antibiotic (Antibiotic-Antimycotic, GIBCO) to constitute the complete growth media. The complete growth media was stored at 2-4 °C.

### Reduced serum media

Basal media was supplemented with 0.5 % fetal bovine serum (FBS, GIBCO) and 1x Antibiotic (Antibiotic-Antimycotic, GIBCO) to constitute the reduced serum media. The complete growth media was stored at 2-4 °C.

### Phosphate-buffered saline (PBS)

PBS was prepared in double-distilled water with the following salts: 0.137 M $NaCl$, 2.68 mM $KCl$, 7.98 mM $Na_2HPO_4$, and 1.4 mM $KH_2PO_4$.

## Section B-2: Reagents used in imaging

### Paraformaldehyde solution

4 % paraformaldehyde was prepared in PBS. The mixture was heated at 60 °C with dropwise addition of 10 N NaOH until a clear solution was formed. The solution was aliquoted into multiple vials and stored at -20 °C.

## Section B-3: Reagents used in agarose gel electrophoresis

### 50x Tris-acetate buffer (TAE)

100 mL of 50x TAE was prepared by dissolving 24.2 g of Tris base in 50 mL of distilled water. To this mixture, 5.7 mL of acetic acid, and 10 mL of 0.5 M EDTA was added. The total volume was made to 100 mL with distilled water. Whenever required, 1x TAE was prepared from 50x stock.

### Agarose gel preparation

To prepare a 1.5 % agarose gel, 0.75 g agarose was mixed in 50 mL of 1x TAE. The mixture was boiled until the agarose is completely dissolved. The solution was allowed to cool. Ethidium bromide at a final concentration of 0.5 µg/mL was added to the solution when the temperature is approximately 50 °C. The solution was poured to the gel casting tray with combs placed and was allowed to solidify.

### Bleach gel preparation

A 1.5 % bleach-agarose gel was prepared by adding 0.75 g of agarose in 50 mL of 1x TAE. 1.5 mL of 4% bleach was added to the solution and kept at room temperature for 10 min. The mixture was boiled until the agarose is completely dissolved. The solution was allowed to cool. Ethidium bromide at a final concentration of 0.5 μg/mL was added to the solution when the temperature is approximately 50 °C. The solution was poured to the gel casting tray with combs placed and was allowed to solidify.

### 6x DNA loading dye

The 6x loading dye was prepared in distilled water as per the following composition: 0.25 % bromophenol blue, 0.25 % xylene cyanol FF, and 30 % of glycerol. The aliquots were stored at -20 °C.

## Section B-4: Reagents used in western blotting

### Radioimmunoprecipitation assay (RIPA) buffer

RIPA buffer was prepared in double-distilled water with the following composition: 50 mM Tris-Cl pH 7.5, 150 mM NaCl, 1 % NP-40, 0.5 % sodium deoxycholate, 0.1 % SDS. The stock solution was stored at -20 °C. Whenever RIPA buffer was used to lyse cells, the necessary protease and phosphatase inhibitors were added. The following inhibitors were used: 1 mM sodium orthovanadate, 50 mM sodium fluoride, 1 mM PMSF (stock solution prepared in 100 % ethanol), 1 mM EDTA.

### 5x Towbin buffer

The Towbin buffer (5x) was prepared in double-distilled water with the following composition: 25 mM Tris base, 192 mM glycine, and 20 % methanol. The pH should be approximately 8.3. This buffer should be prepared fresh whenever required.

### 1x Tris-buffered saline (TBS)

TBS was prepared in double-distilled water with the following salts: 50 mM Tris base and 150 mM NaCl. The pH should be adjusted to 7.5 with 1 N HCl.

### TBST

TBS containing 0.1 % Tween 20.

### 30 % Acrylamide solution

Acrylamide solution was prepared in double-distilled water containing 29.2 % Acrylamide and 0.8 % bisacrylamide.

### Gel running buffer (10x)

The gel running buffer (10x) was prepared in double-distilled water with the following composition: 250 mM Tris base, 1.92 M glycine, and 1 % SDS. The 10x buffer was diluted to 1x with double distilled water before use.

### 4x sample loading buffer

10 mL of 4x dye is prepared by mixing the following components: 2.5 mL 1 M Tris-HCl (pH 6.8), 1 g SDS, 0.8 mL 0.1% Bromophenol Blue, 4 mL 100% glycerol, 2 mL 100% β-mercaptoethanol. Adjust the final volume to 10 mL with double distilled water.

## Section B-5: Reagents used in cell cycle analysis

### DNA extraction buffer

The DNA extraction buffer was prepared by mixing 192 mL of 0.2 M $Na_2HPO_4$ with 8 mL of 0.1 % Triton X-100. The pH was adjusted to 7.8.

## DNA staining solution

The DNA staining solution was prepared by dissolving propidium iodide to a final concentration of 20 μg/mL in PBS. DNase free RNase was added to the mixture at a final concentration of 0.2 mg/mL.

**Table B-1: List of antibodies used in western blotting.**

| *Antibody* | *Make* | *Dilution* |
|---|---|---|
| Anti-Phospho-EGF Receptor (Tyr 1068) | Cell Signaling Technology-3777 | 1:4000 |
| Anti-EGF Receptor | Cell Signaling Technology-4267 | 1:4000 |
| Anti-FAK | Cell Signaling Technology-13009 | 1:2000 |
| Anti-Phospho-FAK (Tyr 397) | Invitrogen-700255 | 1:2000 |
| Goat anti-rabbit HRP-conjugate | Cell Signaling Technology-7074P2 | 1:5000 |

**Table B-2: List of antibodies used in Immunofluorescence.**

| *Antibody* | *Make* | *Dilution* |
|---|---|---|
| Anti-Vimentin Alexa Fluor 488 conjugated | Abcam-ab195877 | 1:50 |
| Anti-SNAIL1 Alexa Fluor 488 conjugated | eBioscience-53-9859-82 | 1:50 |

**Table B-3: List of antibodies used in flow cytometry experiments.**

| *Antibody* | *Make* | *Dilution* |
|---|---|---|
| Anti-Phospho-EGF Receptor (Tyr 1068) | Cell Signaling Technology-3777 | 1:500 |
| Anti-EGF Receptor | Cell Signaling Technology-4267 | 1:500 |
| Goat anti-Rabbit Alexa Fluor 488 conjugated | Invitrogen-A-11070 | 1:1000 |

**Table B-4: List of primers used in PCR.**

| *Gene* | | *Sequence* |
|---|---|---|
| Vimentin | Forward | AGTCCACTGAGTACCGGAGAC |
| | Reverse | CATTTCACGCATCTGGCGTTC |
| Fibronectin | Forward | AGGAAGCCGAGGTTTTAACTG |
| | Reverse | AGGACGCTCATAAGTGTCACC |
| Snail 1 | Forward | TCGGAAGCCTAACTACAGCGA |
| | Reverse | AGATGAGCATTGGCAGCGAG |
| Zeb 1 | Forward | TTACACCTTTGCATACAGAACCC |
| | Reverse | TTTACGATTACACCCAGACTGC |
| EGFR | Forward | GGAGAACTGCCAGAAACTGACC |
| | Reverse | GCCTGCAGCACACTGGTTG |
| Cyclophilin A | Forward | GGGCCGCGTCTCCTTTGAGC |
| | Reverse | GGCGTGTGAAGTCACCACCC |

**Table B-5: The composition of different percentages of resolving gel.**

| | | | | | |
|---|---|---|---|---|---|
| ***6 % Resolving gel*** | ***Components*** | | ***Volume (mL)*** | | |
| | Water | 2.6 | 5.3 | 7.9 | 10.6 |
| | 30% acrylamide | 1 | 2 | 3 | 4 |
| | 1.5 M Tris (pH 8.8) | 1.3 | 2.5 | 3.8 | 5 |
| | 10% SDS | 0.05 | 0.1 | 0.15 | 0.2 |
| | 10% APS | 0.05 | 0.1 | 0.15 | 0.2 |
| | TEMED | 0.004 | 0.008 | 0.012 | 0.016 |
| | Total volume | 5 | 10 | 15 | 20 |
| ***8 % Resolving gel*** | ***Components*** | | ***Volume (mL)*** | | |
| | Water | 2.3 | 4.6 | 6.9 | 9.3 |
| | 30% acrylamide | 1.3 | 2.7 | 4 | 5.3 |
| | 1.5 M Tris (pH 8.8) | 1.3 | 2.5 | 3.8 | 5 |
| | 10% SDS | 0.05 | 0.1 | 0.15 | 0.2 |
| | 10% APS | 0.05 | 0.1 | 0.15 | 0.2 |
| | TEMED | 0.003 | 0.006 | 0.009 | 0.012 |
| | Total volume | 5 | 10 | 15 | 20 |
| ***10 % Resolving gel*** | ***Components*** | | ***Volume (mL)*** | | |
| | Water | 1.9 | 4 | 5.9 | 7.9 |
| | 30% acrylamide | 1.7 | 3.3 | 5 | 6.7 |
| | 1.5 M Tris (pH 8.8) | 1.3 | 2.5 | 3.8 | 5 |
| | 10% SDS | 0.05 | 0.1 | 0.15 | 0.2 |
| | 10% APS | 0.05 | 0.1 | 0.15 | 0.2 |
| | TEMED | 0.002 | 0.004 | 0.006 | 0.008 |

| | | | | | |
|---|---|---|---|---|---|
| | Total volume | 5 | 10 | 15 | 20 |
| ***12 % Resolving gel*** | ***Components*** | ***Volume (mL)*** | | | |
| | Water | 1.6 | 3.3 | 4.9 | 6.6 |
| | 30% acrylamide | 2 | 4 | 6 | 8 |
| | 1.5 M Tris (pH 8.8) | 1.3 | 2.5 | 3.8 | 5 |
| | 10% SDS | 0.05 | 0.1 | 0.15 | 0.2 |
| | 10% APS | 0.05 | 0.1 | 0.15 | 0.2 |
| | TEMED | 0.002 | 0.004 | 0.006 | 0.008 |
| | Total volume | 5 | 10 | 15 | 20 |
| ***15 % Resolving gel*** | ***Components*** | ***Volume (mL)*** | | | |
| | Water | 1.1 | 2.3 | 3.4 | 6.6 |
| | 30% acrylamide | 2.5 | 5 | 7.5 | 8 |
| | 1.5 M Tris (pH 8.8) | 1.3 | 2.5 | 3.8 | 5 |
| | 10% SDS | 0.05 | 0.1 | 0.15 | 0.2 |
| | 10% APS | 0.05 | 0.1 | 0.15 | 0.2 |
| | TEMED | 0.002 | 0.004 | 0.006 | 0.008 |
| | Total volume | 5 | 10 | 15 | 20 |

**Table B-6: The composition of stacking gel.**

| | *Components* | *Volume (mL)* | | | |
|---|---|---|---|---|---|
| ***5 % Stacking gel*** | Water | 2.7 | 3.4 | 5.5 | 6.8 |
| | 30 % acrylamide | 0.67 | 0.83 | 1.3 | 1.7 |
| | 1.0 M Tris (pH 6.8) | 0.5 | 0.63 | 1 | 1.25 |
| | 10 % SDS | 0.04 | 0.05 | 0.08 | 0.1 |
| | 10 % APS | 0.04 | 0.05 | 0.08 | 0.1 |
| | TEMED | 0.004 | 0.005 | 0.008 | 0.01 |
| | Total volume | 4 | 5 | 8 | 10 |