ON SIGNAL TRANSDUCTION IN HUMAN EMBRYONIC STEM CELLS: TOWARDS A SYSTEMS VIEW

by

Shibin Mathew

Bachelor of Chemical Engineering, University of Pune, India, 2008

Master of Chemical Engineering, Indian Institute of Science, Bangalore, India, 2010

Submitted to the Graduate Faculty of

the Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Shibin Mathew

It was defended on

January 8, 2016

and approved by

Robert Parker, PhD, Professor, Department of Chemical and Petroleum Engineering

Gilles Clermont, MD, Professor, Department of Critical Care Medicine

Yoram Vodovotz, PhD, Professor, Department of Surgery

Gerald Schatten, PhD, Professor, Department of Obstetrics, Gynecology and Reproductive Sciences

Dissertation Director: Ipsita Banerjee, PhD, Associate Professor, Department of Chemical and Petroleum Engineering

ON SIGNAL TRANSDUCTION IN HUMAN EMBRYONIC STEM CELLS: TOWARDS A SYSTEMS VIEW

Shibin Mathew, PhD

University of Pittsburgh, 2016

Human embryonic stem cells (hESC) have been a major cell source for research in regenerative medicine due to the demonstration of properties of self-renewal and efficient lineage specific differentiation, both on additions of external cues. Self-renewal provides the potential to extract large quantities of naïve cells that can then be differentiated to clinically relevant mature lineages. While there exists significant proof-of-concept to transform stem cells to the desired lineage, generating fully functional cell types is still an unmet challenge. A major reason for this is our limited understanding of the complexity of the transformation process. The overarching goal of this PhD research was to provide strategies to bring mathematical modeling into the realm of stem cell research, particularly to analyze the complex regulatory network of signaling events controlling cell fate. This work focused on the signaling pathways that in concert control the balance of self-renewal and endoderm differentiation of hESCs.

We proposed a framework for developing mechanistic understanding from disparate signaling pathways using combinations of data-driven and equation based models. As a first step, we analyzed growth factor mediated PI3K/AKT pathway that must remain highly active to inhibit differentiation in self-renewal state. Using an integrated approach of mechanistic modeling, systems analysis and experimental validation we identified the role of a regulatory process (negative feedback) in maintaining signal amplitudes and controlling the propagation of parameter uncertainty down the pathway in the self-renewal state. To analyze endoderm differentiation, biclustering with bootstrapping formulation was used to identify co-regulated transcription factor patterns under a combinatorial modulation of endoderm inducing signaling pathways. In the final step, a detailed mechanistic analysis was done to characterize the dynamic features of TGF-β/SMAD pathway for inducing endoderm. Utilizing a dynamic Bayesian network formulism, AKT mediated crosstalk connections were inferred from the detailed time series data. Modeling of competing AKT-SMAD interactions followed by parametric ensemble analysis enabled identification of plausible hypotheses that could explain experimental observations. Using our integrated approach, we can now begin to rationally optimize for desirable fate of hESCs with reduced variability and accelerate the path towards therapeutic applications of hESCs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

This section has been presented in alphabetical order

AIC: Akaike Information Criterion

AUC: Area Under the Curve

DE: Definitive Endoderm

EA: Evolutionary Algorithm

EBM: Equation Based Model

FLIP: Fluorescence Loss in Photobleaching

FRAP: Fluorescence Recovery after Photobleaching

GA: Genetic Algorithm

GRN: Gene Regulatory Network

GSA: Global Sensitivity Analysis

hESCs: Human Embryonic Stem Cells

HC: Hierarchical Clustering

ICM: Inner Cell Mass

IO: Input-output

MCMC: Markov Chain Monte Carlo

MC: Monte Carlo

ODE: Ordinary Differential Equation

PDE: Partial Differential Equation

PSC: Pluripotent Stem Cell

PSRF: Potential Scale Reduction Factor

PT: Parallel Tempering

RS-HDMR: Random Sampling High Dimensional Model Representation

RBM: Rules Based Modeling

SDE: Stochastic Differential Equation

SEBI: Sequential Evolutionary BIclustering

TF: Transcription Factor

PREFACE

I would like to thank my advisor, Dr. Ipsita Banerjee for the guidance, support and encouragement throughout my PhD process. I greatly appreciate the valuable feedbacks and ideas that have helped me overcome challenges during the stages of my PhD. Dr. Banerjee has been a scientific beacon during my PhD and she has immensely nurtured my scientific skills of collaboration and communication with researchers from diverse backgrounds. I am thankful for the opportunity I had in her lab where I metamorphosed into a “Systems Biologist”.

I would also like to thank the Chemical Engineering Department at PITT for providing a nurturing environment for developing myself as a scientist and as an individual. I am grateful to my professors and fellow students with whom I shared my first year graduate courses and I had wonderful time learning from them. I would like to thank my unnamed computing machines and servers that have borne the burden of my demands, faithfully churning out results for me to ponder and publish! Most importantly, I would like to thank my fellow Banerjee lab members, “the mensch”, who have been my trusted colleagues and from whom I learnt all things scientific and non-scientific! I am grateful to Dr. Sankar S. and Mr. Hikaru Mamiya for performing the wet lab experiments and thus making my models better and biologically meaningful! I am thankful to my previous mentors, Dr. S.R. Inamdar (VIT, Pune) and Dr. Narendra Dixit (IISC, Bangalore) who motivated me to develop my modeling and simulation skills.

Lastly, I would like to thank my parents who have done everything to make me what I am today and my sister for providing encouragement and support. I am grateful to my uncle Dr. Sam Cherian for being my motivation throughout my childhood and for providing me the inspiration to do a PhD.

1.0 INTRODUCTION

Regenerative medicine employs the replacement or replenishment of human cells, tissues, or organs in order to restore natural homeostasis (Ao *et al.*, 2011; Atala *et al.*, 2010). Development of therapies based on stem cells has been the cornerstone of research in this area for treatment of chronic illnesses like diabetes (Mimeault *et al.*, 2007). This attraction for stem cells arises from two of their unique properties, namely self-renewal and lineage specific differentiation, both on additions of external cues (Figure 1.1). While there exists significant proof-of-concept to transform stem cells to the desired lineage, generation of fully functional cell types *in vitro* that are ready for clinical applications is still an unmet challenge (Soria *et al.*, 2015). A major reason for this is our limited understanding of the complexity of the transformation process (Huang, 2011). The overall objective of this PhD research was to purport strategies to gain a mechanistic understanding of the complex behavior of stem cells by integrating the power of mathematical and computational sciences with targeted *in vitro* experiments.

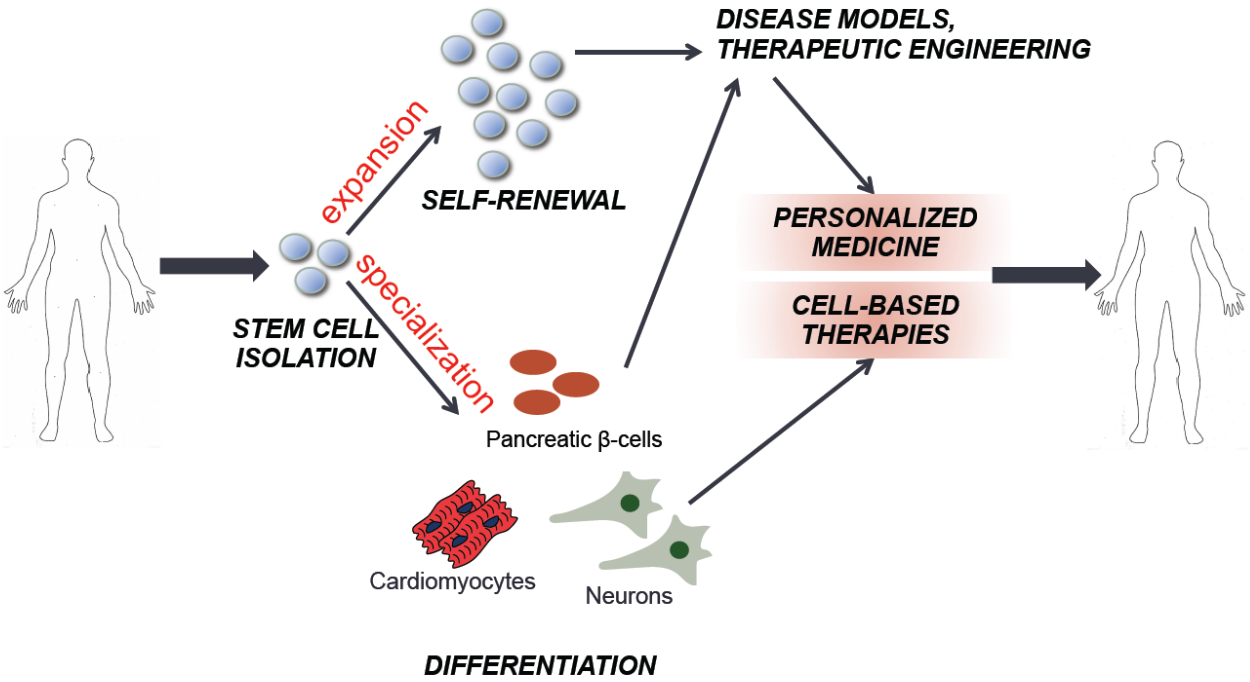


Figure 1.1 Channels for stem cell research

1.1 HUMAN EMBRYONIC STEM CELLS

By definition, a stem cell is a cell that replaces itself through proliferation for prolonged periods of time (property called self-renewal) and gives rise to differentiated cell types in the presence of proper cues (Jones and Thomson, 1999). Human embryonic stem cells (hESCs) are derived from the inner cell mass (ICM) of a blastocyst-stage human embryo. The ICM cells are taken out of their normal embryonic environment and cultured in an *in vitro* setting establishing a cell line for research and therapeutic applications (Mummery *et al.*, 2014). Several decades of research have given rise to a plethora of established protocols for deriving cell types of the major germ layers of human embryonic development from hESCs; namely endoderm (pancreas, lung and liver), mesoderm (heart, blood, vascular and skeletal muscle) and ectoderm (neural and skin cells)

(Murry and Keller, 2008). While significant strides have been made in this field, there are several challenges in maintaining self-renewal and inducing lineage specific differentiation and these will be discussed below.

1.1.1 Self-renewal and its challenges

In the embryo, self-renewal is a transient stage and this is artificially extended indefinitely in an *in vitro* context by mimicking appropriate chemical and mechanical signals (Nichols and Smith, 2012). Proliferation, which is an important feature of the self-renewal stage, provides the potential for obtaining large quantities of cells necessary for regenerative medicine applications and often the proliferative capacity is decreased on differentiation (Molofsky *et al.*, 2004). Consequently, maintaining self-renewal over multiple passages is of therapeutic value but this brings additional constraints that the purity of the state has to be preserved as much as possible. Decades of research have established the types of gene regulatory networks (GRNs) that define the self-renewal state, with the transcription factors (TFs) *OCT4*, *SOX2* and *NANOG* as the core players in these networks (Yeo and Ng, 2013). The expression of these markers, emerging from their interactions in participating GRNs, is taken as a litmus test for self-renewal and the corresponding marker heterogeneity in the population is taken as a test for its purity. As a simple example, hESC cultures are known to show a bimodal distribution of *NANOG* with *NANOG*$^{\text{high}}$ cells in the population in a self-renewal mode and *NANOG*$^{\text{low}}$ cells showing increased propensity to differentiate (Fischer *et al.*, 2010).

Stochastic origin of this transcriptional heterogeneity in the self-renewal state has been the most common theory purported by several experimental and mathematical modeling studies (Torres-Padilla and Chambers, 2014; Wu and Tzanakakis, 2012). Therefore, cell-sorting techniques have become the common standards for obtaining homogeneous cell cultures (Fong *et al.*, 2009; Nicholas *et al.*, 2007). This is however, a less reliable strategy from a control point of view due to the re-establishment of heterogeneity on further culture and associated loss of cell numbers (Torres-Padilla and Chambers, 2014). Heterogeneity also compromises the translational potential of these cells. In 2013, pioneering work from the Vallier and Dalton groups revealed the importance of chemical signaling pathways and cell-cycle state in controlling transcriptional heterogeneity in hESCs (Pauklin and Vallier, 2013; Singh *et al.*, 2013). These features are upstream of the GRNs (though feedback exists) and directly influenced by culture media. This has resulted in the necessity of a mechanistic understanding of hESC signaling and its manipulation by modifying cell culture conditions. Furthermore, these and other experimental studies have eventually led to the realization that hESC fate choice is very complex and in order to improve maintenance of self-renewal, simple analyses using small set of controlling factors are insufficient.

1.1.2 Differentiation and its challenges

Similar to the self-renewal state, GRN programs specific to the induced lineage control hESC differentiation. In general, differentiating hESCs have the choice to select various combinations of genetic markers (called as Epigenetic Landscape of Cell State (Waddington, 1957)). The type of signals present in the cell culture and cell-intrinsic properties guide the final fate choice (referred to as highway guide-rails on an Epigenetic Landscape (Schatten, 2013)). Plasticity of hESCs however plays a dual role; on one hand conferring therapeutic potential but on the other hand increasing contamination of differentiating population by non-desirable lineages. The negative effect is because of the difficulty in recapitulating exactly all the conditions (or constraints) existing *in vivo* (called the micro-environmental niche) in an *in vitro* setting (Schatten, 2013). What these constraints are is still an area of active research. These constraints may be imposed by external factors like basement membranes, extracellular matrices and cell-cell contact, internal factors like composition of cellular receptors, cytoskeleton, chromatin organization etc. (Schatten, 2013).

Evolution of studies in diverse areas of regenerative medicine and tissue engineering has converged on many of the essential factors that mediate the influence of the above factors in this niche (Discher *et al.*, 2009). Growth factors and the signaling pathways that they activate are a major component. Naturally, identifying the mechanisms by which the hESCs regulate signal transduction is an important step in recapitulating the niche and consequently this knowledge can be used to control differentiation potential. Further, chemical stimuli have been the most widely used method for regulating cell fate due to their more defined and relative ease of application as compared to others like mechanical stimuli. These signaling pathways work by transducing the signals that originate from chemical factors in the cellular environment to the nuclear machineries inside the cell via a host of interacting molecules. However, signaling interaction networks have received meager attention compared to GRNs in hESC differentiation field, firstly due to laborious experimentation required to simultaneously measure the activity of multiple signaling molecules with high temporal sampling frequency and secondly, due to the difficulty in interpreting the results without a blueprint of the complete network and kinetics of the embedded reactions. In 2012, a pioneering and thorough experimental work by Dalton group identified several crosstalk interactions that controlled the balance of self-renewal and early differentiation of hESCs indicating additional complexity via crosstalk (Singh *et al.*, 2012a). Thus, this and other experimental studies lead to the realization that “Signaling in PSCs (pluripotent stem cells which hESCs are a part of) is a complex, dynamic process where thresholds, temporal changes and combinatorial effects make important contributions to cell fate outcomes” (quoted text taken verbatim from (Dalton, 2013)).

1.2 RATIONAL CONTROL OF HESC FATE CHOICE

The challenges presented in Section 1.1 show that difficulty in manipulating and maintaining cell fate is the most important concern facing hESC culture and differentiation. The experimental community is taking efforts to address these challenges by identifying better signaling pathway and genomic modulators. Since the early days of hESC derivation, there have been efforts to develop defined chemical media for hESC culture by identifying new chemical factors that can have targeted and well defined effects, thus leading to removal of undefined factors of xeno-genic origin (Xu *et al.*, 2001). Currently, the field is actively investigating small molecules for this purpose (Atkinson *et al.*, 2013).

In general, small molecules are low molecular weight (typically < 500 Da) organic compounds that can rapidly diffuse across cell membranes and reach the intracellular sites of action (Veber *et al.*, 2002). Several research groups are using high throughput screening platforms and large arrayed chemical libraries for discovering appropriate small molecules (Gafni *et al.*, 2013; Zhang *et al.*, 2012b). The focus until now has been on developing and integrating new chemical and functional genomic tools to identify optimal formulations of the additives. Many groups also focus on the molecular mechanism of action of these small molecules (Xu *et al.*, 2008) and on the combinatorial effects of multiple molecules using statistical frameworks (Marinho *et al.*, 2015). While such data driven approaches can identify important molecular interactions, predictive control over such complex interacting pathways is best achieved by mechanistic models. Mechanistic mathematical models are invaluable in gaining such insights into the action of signaling networks.

1.3 MATHEMATICAL MODELING OF HESC SYSTEMS

Mathematical models have been extensively used to understand the behavior of complex biological systems (Murray, 2002). However, the application of mathematical modeling for gaining mechanistic insights is less explored in hESC research. The primary goal of this PhD dissertation is to answer questions (Section 1.4) in hESC signal transduction that are challenging to answer using experiments alone and can be complemented by adequate mathematical treatment. Mathematical models come in different flavors, from small-scale models composed of a few entities and interactions to large-scale network models capturing multiple entities and interactions. Since hESC signal transduction networks belong to the latter category (Dalton, 2013), the discussions in this section will mainly pertain to the advantages, developments and challenges in modeling complex systems.

1.3.1 Typical modeling frameworks for complex interacting systems

1.3.1.1 Mechanism based models

For interacting systems where preliminary information about the nature of interactions is available but the actual values of the rates of these interactions are not available, rules based modeling (RBM) approach is first undertaken. Here, rules are written to describe interactions between the entities of the system based on biological observations. For example, a rule may indicate what level a particular molecule must take when another molecule (same type or different) is nearby or at a certain level. Boolean models and agent-based models belong to this category (Miskov-Zivanov *et al.*, 2013; Ziraldo *et al.*, 2015). These models are advantageous even if partial information about the system is available, since they can incorporate qualitative guesses easily. These models are particularly valuable for building models for systems from scratch and give a rough idea of the behavior of the system, which can be used for hypothesis testing, experimental design and later inquiry using equation based frameworks.

Equation based models (EBMs) are common for mathematical representation of complex physical phenomena in science, engineering and medicine (Aldridge *et al.*, 2006; Daun *et al.*, 2008; Parker and Clermont, 2010). EBMs are particularly suited to represent changes in large number of interacting components with independent variables like time and space. If there is some *a priori* knowledge on the nature of these interactions, solving these equations will allow prediction of how the system evolves in time or space and what happens when the characteristics (parameters/conditions) of the system are varied. The EBM tools are useful when the predictions of the system are not directly obvious and when mechanisms underlying some observations cannot be arrived at using intuition alone. In a typical EBM, time (t) and space ($\{x, y, z\}$) are taken as the independent variables and if these are discrete, then a difference equation results and if continuous, a differential equation results. The entities that undergo changes in time and space, for example, species, molecules, volume etc. are the dependent variables (Y). To write an EBM, a conservation law has to be invoked. Mass and energy balances are common laws used for biological systems. A differential equation with one independent variable (called as ordinary differential equation or ODE) is represented as:

$$\begin{aligned}&\frac{dY}{dt} = \varphi(Y, t, \overline{k}, \overline{U}),\\&Y(t=0) = Y_0\end{aligned} \tag{1.1}$$

Here, the parameter vector, $\overline{k}$, represents physical constants and chemical kinetic parameters and $\overline{U}$ represents external input vector. The parameters, external inputs and the initial condition (Y_0) are to be supplied to solve the ODE. Adding additional spatial structure to the model gives rise to compartmental models and partial differential equations (PDEs). Further, these methods are purely deterministic and give the same outcome for a given set of parameters and initial conditions. In many systems, however, noise effects due to temperature fluctuations, variability in molecular interactions and low numbers of the interacting molecules can modulate the deterministic behavior and these can be described using stochastic differential equations (SDE). In this dissertation, deterministic EBMs will be explored to resolve the large numbers of each signaling molecule in the pathways of importance to hESCs. Furthermore, the experimental techniques commonly employed for hESC signal transduction research employ population averages and a deterministic treatment is sufficient unless otherwise stated in the following chapters. For many signaling pathways, detailed ODE models are available with many of the network connections known and rate parameters calibrated for mammalian systems. Therefore, an EBM framework can be directly applied to signaling in hESCs with the same basic network structure (with modifications allowed based on the context) but with recalibration of the rate parameters.

1.3.1.2 Data driven models

The other category of modeling tools, called data-driven models are useful for teasing out the essential features of the experimental data (or outputs from multiple simulation runs of an

ODE). Availability of big data in signal transduction from high throughput and multiplex strategies require the use of statistical frameworks to make sense of the dense information content (Albeck *et al.*, 2006; Jaqaman and Danuser, 2006). The data available for signal transduction modeling is often in the form of levels, localization and activities of several proteins for multiple timescales and treatment conditions. Commonly used data-driven techniques to analyze such data include clustering for data organization, principal component analysis for data condensation and partial least squares regression for data prediction (Janes and Yaffe, 2006; Vodovotz and An, 2014). In these techniques, the entire data is represented as a matrix or set of matrices and appropriate matrix operations are done to identify and quantify similar and distinct features embedded in the dataset. While both EBMs and data driven models have their own advantages, using techniques from each category and allowing crosstalk between these categories can prove to be very useful for obtaining a fully developed view of the system under consideration (Hua *et al.*, 2006).

1.3.2 Modeling stem cell behavior

Current methods to model stem cell behavior have focused on two aspects of these cells: (1) the behavior of populations of cells and (2) behavior of intracellular signals in a single cell. Population models have been useful to understand the proliferation kinetics of hESCs and influence of growth and differentiation on this process. Intracellular models have focused on mechanisms within a single cell that control the fate choice of hESCs.

1.3.2.1 Population based models

Models capturing the population behavior of stem cells have been explored with great detail in stem cell systems. The earliest application of such models to simulate stem cell proliferation was by physicians Till and McCulloch, who together made the groundbreaking discovery of the existence of multi-potent stem cells in the bone marrow (Till and McCulloch, 1961), which eventually lead to the rise of stem cell research. They applied stochastic models of birth-death processes to explain the enhanced proliferative capacity of such cells and identified a probability distribution of rate parameters that explains the experimental proliferation data (Till *et al.*, 1964). Later models have focused on distinguishing different members of a population based on their differentiation stage. Using a rules based stochastic population model, Task *et al.* estimated the kinetics of differentiation in common endoderm induction conditions and the best sequence of transition stages in the lineage commitment process of hESCs (Task *et al.*, 2012). Another rules based study explored the cell fate transitions in an 3D multi-cellular aggregate of mouse ESCs (mESCs) undergoing differentiation (White *et al.*, 2013). Using an ODE-based population model and sensitivity analysis, Selekman *et al.* estimated the influence of different cell decision rate parameters on overall differentiation yield (Selekman *et al.*, 2013). Such models have been further extended with simulation of cell-cell interactions to capture the spatial distribution of differentiated cells on micropatterned surfaces (Smith *et al.*, 2015).

1.3.2.2 Models of intracellular processes

A large number of studies have explored GRNs controlling fate specification of different types of stem cells (Herberg and Roeder, 2015). Chickarmane *et al.* modeled a bistable switch in a GRN model (utilizing ODE framework) of *OCT4-SOX2-NANOG* pluripotency factors to explain the transition from self-renewal to differentiation (Chickarmane *et al.*, 2006) and refined the model with addition of lineage specific TFs to explain lineage specification (Chickarmane and Peterson, 2008). Bifurcation analysis has been commonly employed to identify the stable states of these networks (Bessonnard *et al.*, 2014). Other groups have explored the influence of stochastic processes on GRNs controlling fate specification during induced and non-induced pluripotency (Glauche *et al.*, 2010; Herberg and Roeder, 2015; MacArthur *et al.*, 2008). These models have focused on a small set of transcriptional mediators. Currently, with availability of high throughput techniques, data-driven methods are being employed for identification of large-scale transcriptional networks of cellular differentiation (Cahan *et al.*, 2014).

Application of mathematical frameworks for signaling pathways in stem cells is relatively sparse. Prudhomme *et al.* applied a multivariate partial least squares regression technique to identify the combinations of intracellular signals that best influence self-renewal and differentiation of mESCs (Prudhomme *et al.*, 2004). Woolf *et al.* used a Bayesian learning algorithm to identify the network structure of signaling molecules and the influence of specific signaling molecules on downstream responses like proliferation and differentiation in mESC system (Woolf *et al.*, 2005). These studies, however, have not explicitly focused on the kinetics of signal transduction during differentiation. In one such study, Mahdavi *et al.* applied a detailed ODE model of the JAK/STAT pathway for identifying an optimal ligand delivery strategy to enhance self-renewal in mESCs (Mahdavi *et al.*, 2007). To our knowledge, mathematical models have not been used to analyze signaling kinetics, crosstalks and signal regulation in hESC system before. But, many ODE based mathematical models exist for the pathways that are of relevance for the hESC system. In order to evaluate the features of signal transduction in hESCs, we adopted these models to the hESC system after making necessary relaxations to the model constraints and adding additional features based on the context. Before utilizing these models for hESCs, certain precautions (discussed in Section 1.4) have to be taken to ensure their predictive capability.

1.4 SPECIFIC AIMS

Our long-term goal of applying mathematical modeling and computational approaches to hESC research is to identify principles that govern fate choice of hESCs. Our hope is that understanding the network motifs and mechanisms by which signals get transduced opens the window to rationally control hESC behavior, reduce culture heterogeneity and enhance efficiency of fate commitment. The overarching goal of this PhD dissertation was to identify signaling mechanisms regulating the process of self-renewal and early differentiation using an integrated experimental and computational workflow. For self-renewal, we chose the system of H1 hESCs cultured in tissue culture plates coated with matrigel. For differentiation, we selected the same system and induced them towards Definitive Endoderm (DE), which is the critical first step towards clinically relevant lineages of pancreas, liver and lungs (Semb, 2008). Here, we focused on identifying signaling mechanisms that are the most critical ones in the cell fate decisions of hESCs. We used experimental techniques that measure population averages of signaling molecules and we recognize this to be the first step in characterizing signaling of a new system like hESC. The PhD project was divided into the three aims listed below (also see Figure 1.2). For each aim, the influence of variability existing in the experimental data as well as the influence of parametric variability on the robustness of model predictions was given special importance.

1.4.1 Specific Aim 1: Identification of robust perturbations that control PI3K/AKT pathway activity in self-renewing hESCs

In self-renewing hESCs, growth factor induced PI3K/AKT pathway has the function of inhibiting differentiation signals and maintaining self-renewal in long-term culture. Currently the common method of activating the pathway is by controlling the external concentration of growth factors. Determination of proper intracellular reactions of this pathway, perturbation of which can enhance signal propagation through this pathway, has not been undertaken before. We developed a detailed ODE based mathematical model of the pathway and by its integration with a meta-model approach and model informed experimental perturbations identified sensitive nodes in the pathway that controlled the activity and variability of key self-renewal molecules. Since the ODE based model of PI3K/AKT pathway was applied to the new system of hESCs, we first identified the parameter intervals that could capture qualitative features of detailed experimental dynamics (Kim *et al.,* 2010). Then a formal study was conducted using a meta-model based Global Sensitivity Analysis (GSA) to identify most sensitive nodes in the pathway (Kent *et al.*, 2013; Kiparissides *et al.*, 2009). The network nodes (or molecules) affected by the most sensitive parameters from this study were experimentally perturbed to ensure that the model structure is adequate to represent the system. The results are detailed in Chapter 2.

1.4.2 Specific Aim 2: Identification of specific combinations of external growth factors that enhance DE differentiation of hESCs

Currently many protocols exist that modulate the activity of one or a few pathways for DE induction from hESCs. The co-operative effect of various endoderm induction pathways, along with its impact on long-term maturation has received less attention. Utilizing a data-driven biclustering + bootstrap approach, we systematically analyzed the combinatorial action of five major signaling pathways and identified robustly co-regulated DE TFs. The results are detailed in Chapter 3.

1.4.3 Specific Aim 3: Quantitative analysis of SMAD signaling in hESCs and modeling crosstalk interactions with AKT

1.4.3.1 Sub aim 3a: Identification of coarse grained network of signaling interactions governing DE differentiation of hESCs

The presence of extensive crosstalk interactions and uncontrolled variability has made signaling data from hESCs difficult to interpret (Dalton, 2013). Rational manipulation of signaling during differentiation has to account for nature of these interactions as well as the time points when they are active. In this aim, we evaluated two common DE induction conditions to identify temporal within-pathway and between-pathway interactions among molecules belonging to the TGF-β/SMAD, PI3K/AKT and MAPK/ERK pathways that drive DE differentiation. The results are detailed in first part of Chapter 4.

1.4.3.2 Sub aim 3b: Mechanistic analysis of activin induced TGF-β/SMAD pathway dynamics and its crosstalk with PI3K/AKT pathway during DE differentiation of hESCs

To obtain a quantitative predictive model of the differentiation process, it is necessary to estimate the kinetics of signal transduction. In this aim, a detailed ODE based model of the TGF-β/SMAD pathway interactions in combination with three literature based crosstalk interactions of SMADs with the molecule AKT from the PI3K pathway were developed. The resulting model was calibrated to experimental time series of signaling during DE differentiation. We utilized a Bayesian Parallel Tempering approach to identify the parameter ensembles that can capture the experimental data. Special care was taken to ensure that the parameters are identifiable, make biological sense and that we get good fits to most of the experimentally measured species (Gutenkunst *et al.*, 2007; Slezak *et al.*, 2010). Theoretical analysis of the parametric ensembles from the calibrated model revealed several differences between the competing crosstalk mechanisms that are currently under experimental investigation. The results are detailed in second part of Chapter 4.

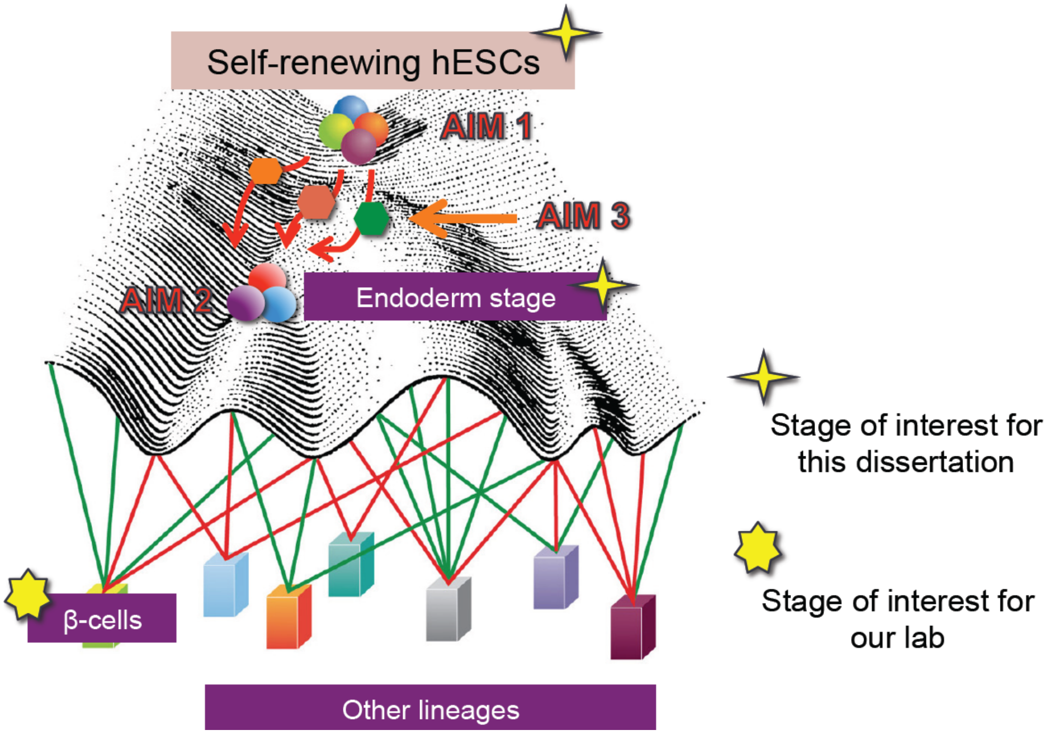


Figure 1.2 Overview of the specific aims of this dissertation

The three aims are overlayed on the differentiation landscape. For this dissertation, self-renewal stage and the first step towards pancreatic beta cells, aka endoderm, is analyzed. Aim 1 concentrated on the self-renewal stage, Aim 2 concentrated on the endpoint of endoderm differentiation but obtained through different pathways and Aim 3 concentrated on the dynamics of transition from self-renewal stage to endoderm using the pathway of minimal modifications.

2.0 REGULATORY INTERACTIONS MAINTAINING SELF-RENEWAL OF HESCS AS REVEALED THROUGH SYSTEMS ANALYSIS OF INSULIN MEDIATED PI3K/AKT PATHWAY

The content of this chapter is taken from Mathew, S., Sankaramanivel, S., Mamiya H. and Banerjee, I., 2014. Regulatory interactions maintaining self-renewal of human embryonic stem cells as revealed through systems analysis of PI3K/AKT pathway. *Bioinformatics* 30(16), 2334-2342

2.1 INTRODUCTION

Long-term maintenance of hESCs in the self-renewal state requires a fine balance of many signaling pathways, including PI3K, TGFβ, WNT and ERK (Singh *et al.*, 2012b). Several earlier studies have reported that the PI3K/AKT pathway plays a central role in balancing self-renewal and differentiation but with limited mechanistic details (McLean *et al.*, 2007b). The pioneering work by Singh *et al.* first recognized the presence of molecular switches controlled by the PI3K/AKT pathway that promotes self-renewal in its active state and strengthens the differentiation signals in its inactive state (Singh *et al.*, 2012b). In this aim, our goal was to evaluate the steady state of the PI3K/AKT pathway in self-renewing hESCs and identify perturbation points in this pathway to improve self-renewal capacity.

Kinase p-AKT is a key effector of the PI3K/AKT pathway and participates in critical functions like survival, metabolism, protein synthesis, cell cycle etc. (Taniguchi *et al.*, 2006). In addition, in hESCs, p-AKT regulates the activity of pluripotency factors like *c-MYC* and controls the levels of differentiation molecules like p-SMAD2/3, p-ERK, p-GSK3β in the self-renewal state (Singh *et al.*, 2012b). Therefore, maintaining high levels of p-AKT is necessary for long-term self-renewal of hESCs.

In spite of the recognized role of p-AKT, there is limited understanding on the maintenance of p-AKT levels in hESCs by the network of regulatory interactions. The PI3K/AKT pathway includes several positive and negative feedback loops and negative regulators like PTP, PTEN and SHIP that together influence its state (Taniguchi *et al.*, 2006). Analysis of such regulatory interactions will be helpful in the design of targeted molecules to support self-renewal. This, however, requires a quantitative systems level approach rather than a restricted study of few interactions. Therefore, in the current work, we have used an integrated experimental and computational approach to identify regulatory interactions maintaining p-AKT levels during hESC self-renewal.

This being the first effort towards modeling the PI3K/AKT pathway dynamics in hESCs, our workflow includes the following critical steps: (i) determine a mathematical structure adequately describing the hESC system, (ii) determine the parameter range over which the model captures hESC behavior, (iii) identify the relative importance of components of the pathway in hESCs, (iv) validate the model predicted sensitive processes in hESCs. These steps result in a validated mathematical representation of the pathway for hESCs. However, any modeling effort of a biological system is incomplete without understanding how parameter variability affects its predictions. We treat this as an important requirement due to the notorious cell-to-cell variability in hESC systems. We, therefore, (v) analyze the propagation of uncertainty in the pathway under various states of the identified sensitive regulators. This will, in turn, allow identification of processes that promote a robust behavior under experimental variability.

To accomplish these objectives, we started with a well-established ODE model of insulin mediated PI3K/AKT pathway developed for adipocytes by Sedaghat *et al.* (Sedaghat *et al.*, 2002). The model is a compendium of accepted knowledge of the pathway, and has been successfully tested in many mammalian systems. We developed a systematic procedure to adopt the Sedaghat model to a system of self-renewing hESCs. We first performed extensive parameter sampling to identify the mechanisms relevant for hESCs. We next evaluated the most significant contributors to the active levels of key molecules using GSA. We adopted random sampling high dimensional model representation (RS-HDMR) based meta-model approach to overcome the large Monte Carlo (MC) sampling requirements of traditional GSA. The model-predicted sensitive processes were successfully validated by a series of perturbation experiments. Our workflow, thus, demonstrates the application of computationally efficient techniques for mechanism detection in uncertain systems like hESCs.

2.2 SYSTEM AND METHODS

2.2.1 Mathematical model of PI3K/AKT pathway

The insulin-mediated activation of the PI3K/AKT pathway can be divided into two modules: Module 1: insulin receptor activation, internalization and recycling, and Module 2: post-receptor signaling cascade involving PI3K/AKT (Figure 2.1A and B). On stimulation with insulin, insulin receptors are auto-phosphorylated and become available for further signaling. These active receptors then undergo intracellular trafficking as shown in Module 1. The active receptors on the surface propagate the signal to components of PI3K/AKT pathway as shown in Module 2. This includes phosphorylated IRS1 (tyrosine), kinase PI3K and phosphoinositol lipids like PI(3,4,5)P3 (or PIP3 henceforth). The signal then propagates to important kinases like AKT and PKC-ζ. Negative regulators that catalyze dephosphorylation reactions include the following: PTP1B or PTP (dephosphoryate active receptors and active IRS1), PTEN (dephosphorylates PIP3 to PI(4,5)P2) and SHIP (dephosphorylates PIP3 to PI(3,4)P2). The pathway also activates negative feedback loops by serine phosphorylation of IRS1 via kinases like PKC-ζ and a double negative loop from AKT resulting in deactivation of PTP. In this article, we relaxed model assumptions by Sedaghat *et al.* for a more generalized analysis. The details of the ODEs and relaxed assumptions are given in Appendix A (Note: Appendix A contains extra figures and tables for this chapter and the names for these are mentioned with letter A in this chapter). The current version of the model comprises 27 reactions, 20 output species and 31 rate parameters. From the rate parameters, 21 were selected as free inputs for GSA (Table A.1 in Appendix A), and the remaining were functions of these selected inputs. Other input parameters included the concentrations of the molecules PTP, PTEN and SHIP. The output molecules of interest for analysis in this chapter were p-IR, p-IRS1 (Y), p-IRS1 (S) and p-AKT.

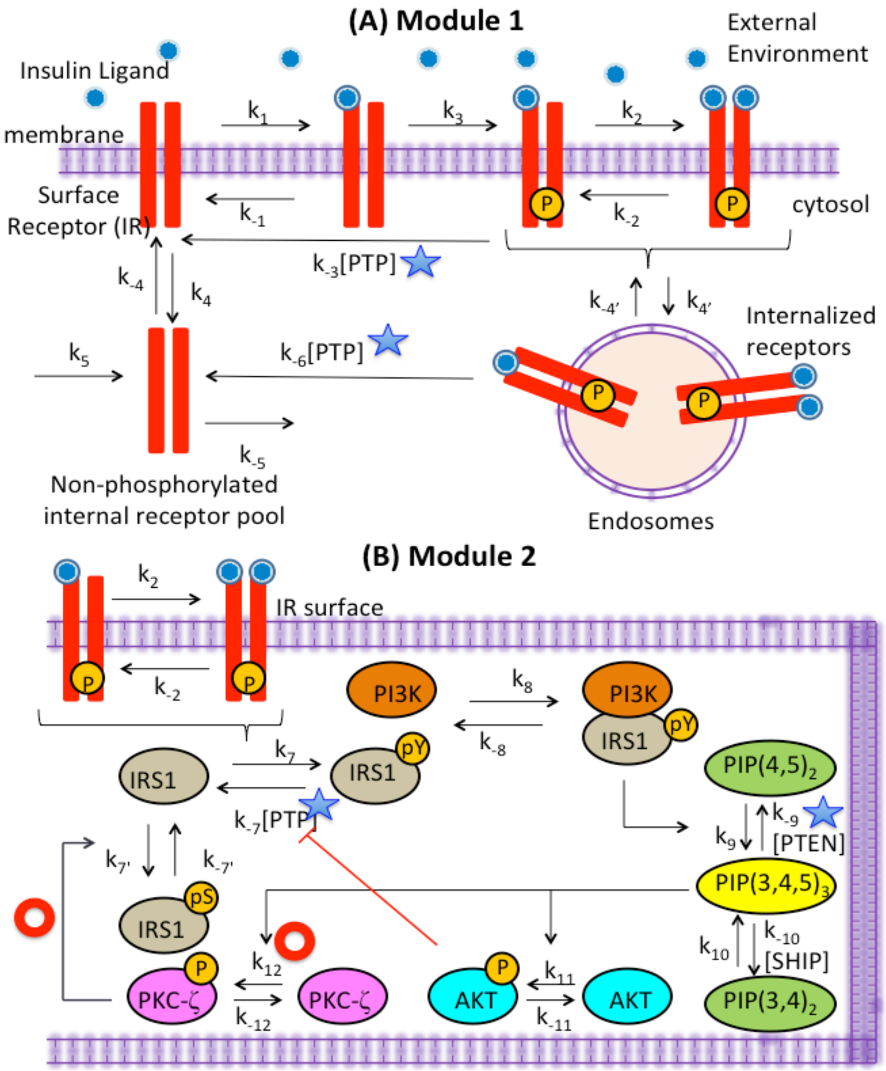


Figure 2.1 Schematic of insulin mediated PI3K/AKT pathway.

(A) Insulin receptor level processes. (B) Intracellular signaling in PI3K/AKT pathway. The reactions marked by red donut (negative feedback) and blue star (PTEN and PTP) are perturbed in experiments mentioned in Figure 2.7.

2.2.2 RS-HDMR based meta-model for analyzing global sensitivity of high dimensional models

Traditional sensitivity analysis techniques are local in nature, and these evaluate the influence of each free parameter in isolation while the remaining parameters are kept constant at their nominal values. This being the first attempt to model hESCs, it was necessary to estimate global sensitivity measures that are applicable in a wide region of the parameter space and capture parameter interactions. The advantages of traditional GSA based on MC methods are, however, challenged by the large number of parameters and the large number of samples required for accurate estimates of the sensitivity indices. To reduce computational cost, we adopted a meta-modeling technique called RS-HDMR developed by Li and Rabitz at Princeton (Li and Rabitz, 2012b).

2.2.2.1 Sample generation

For RS-HDMR, the input parameters must be normalized (say into x_i) so that they lie in the range [0,1]. The normalized variable, x_i, can be converted into its actual value, Ξ_i, in the interval $[a_i, b_i]$ by the transformation: $\Xi_i = a_i + (b_i - a_i) \times x_i$ which is then used to evaluate the ODEs. For the current application, the variable x_i is chosen as a uniform random variable. The model was simulated in FORTRAN R90. Random samples were generated using the ran function in FORTRAN, which generates a uniform random number between 0 and 1 (Teukolski et al., 1989) with a new seed for each MC sample set. Each such sample is denoted by vector $\overline{x}^s = \left(x_1^s, x_2^s, \ldots, x_k^s\right)$, where individual components $x_i \in [0,1]$, k denotes the number of free parameters, and the superscript, s, denotes the sample number. The resulting MC samples are collected in the training matrix, $M_{train} = \begin{bmatrix} \overline{x}^1 \\ \overline{x}^2 \\ - \\ - \\ \overline{x}^N \end{bmatrix} = \begin{bmatrix} x_1^1, x_2^1, \ldots, x_k^1 \\ x_1^2, x_2^2, \ldots, x_k^2 \\ ------- \\ ------- \\ x_1^N, x_2^N, \ldots, x_k^N \end{bmatrix}$, where N is the total number of samples. For the generated matrix, M_{train}, the original ODE model described in the previous section was integrated using the DLSODE solver (Hindmarsh, 1983). The complete ODE integration process for 10^5 samples takes 4 min on INTEL® Core™ 2 Quad CPU (Q8400 @ 2.66GHz). The selection of number of samples for RS-HDMR is discussed in Section 2.3.2.

2.2.2.2 RS-HDMR algorithm for meta-model development

RS-HDMR is an efficient technique for the identification of the nonlinear IO behavior of high dimensional systems (Li *et al.*, 2001a; Li and Rabitz, 2012b). The performance characteristics of the method were verified in high-dimensional ODE models from different fields, namely atmospheric photochemistry (Li *et al.*, 2002b), genetic circuits (Feng *et al.*, 2004), combustion processes (Davis *et al.*, 2011), signaling pathways (Mathew *et al.*, 2014) and in network identification of a biochemical interaction model (Miller *et al.*, 2012). RS-HDMR decomposes the selected output of the ODE, $Y = f(\overline{x})$ (here, steady state levels of molecules like p-AKT), into component functions represented by the following hierarchical expansion in the k input parameters $\overline{x} = (x_1, x_2,...,x_k)$:

$$Y = f(\overline{x}) = f_0 + \sum_{1\le i \le k} f_i(x_i) + \sum_{1\le i<j \le k} f_{ij}(x_i,x_j) + ... + f_{123...k}(x_1,x_2,...,x_k) \qquad (2.1)$$

The individual functions are called component functions and represent the influence of each type of input combination. These functions are evaluated from specific instances of the ODE output as detailed below:

$$f_0 = E(Y)$$

$$f_i(x_i) = E(Y|x_i) - f_0$$

$$f_{ij}(x_i,x_j) = E(Y|x_i,x_j) - f_i - f_j - f_0 \text{ etc.} \qquad (2.2)$$

The component functions are, therefore, higher dimensional integrals defined by:

$$E(Y) = \int_{[0,1]^k} f(\bar{x})d\bar{x}$$

$$E(Y|x_i) = \int_{[0,1]^{k-1}} f(\bar{x})\prod_{m\neq i} dx_m \tag{2.3}$$

Each component function represents the contribution of the corresponding input variable to the output. First order components, $f_i(x_i)$, describe the contribution of a single parameter independently of the others while higher order functions, $f_{ij}(x_i,x_j)$ etc., describe the joint contributions of the parameters. Most importantly, the method rests on the realization that the expansion given in Equation 2.1 often converges rapidly so that higher order interactions (≥ 3) are often negligible. Expansions until the second-order terms are known to sufficiently describe most physical systems (Li *et al.*, 2001b; Li *et al.*, 2002b).

The component functions can be evaluated by polynomial approximations (Li and Rabitz, 2012a; Li *et al.*, 2002a; Li *et al.*, 2010). In this work, we used orthonormal polynomial approximations for 2nd order RS-HDMR given by:

$$f_i(x_i) \approx \sum_{1\leq p\leq u} \alpha_p{}^i \varphi^i{}_p(x_i)$$

$$f_{ij}(x_i,x_j) \approx \sum_{1\leq q\leq v}\sum_{1\leq r\leq w} \beta^{ij}{}_{qr}\varphi^i{}_q(x_i)\varphi^j{}_r(x_j) \tag{2.4}$$

The integers u, v, w are orders of the orthonormal polynomials, ϕ, and are usually ≤ 3 (Li *et al.*, 2010). Based on Li *et al.*, we used Jacobi polynomials in the domain [0,1] as the orthonormal basis functions, ϕ. These polynomials satisfy the following properties of orthonormality, namely,

$$\int_0^1 \varphi_a(x)dx = 0,\ a = 1,2,3$$
$$\int_0^1 \varphi^2{}_b(x)dx = 1,\ b = 1,2,3$$
$$\int_0^1 \varphi_c(x)\varphi_d(x)dx = 0, c \neq d$$

Using these properties, the first three orthonormal polynomials were constructed to be:

$$\begin{aligned} \varphi_1(x) &= \sqrt{3}(2x-1) \\ \varphi_2(x) &= 6\sqrt{5}\left(x^2 - x + \frac{1}{6}\right) \\ \varphi_3(x) &= 20\sqrt{7}\left(x^3 - \frac{3}{2}x^2 + \frac{3}{5}x - \frac{1}{20}\right) \end{aligned} \tag{2.5}$$

See Rice and Do for further description of the construction of Jacobi polynomials (Rice and Do, 2012). By definition, the orthonormality of the basis functions preserve the orthonormality between the RS-HDMR component functions. Hence, the coefficients in Equation 2.4, α_p^i and β_{qr}^{ij} are obtained by:

$$\alpha_p^i = \frac{\int_{[0,1]^k} f(\bar{x})\varphi_p(x_i)d\bar{x}}{\int_{[0,1]} \varphi^2{}_p(x_i)dx_i} \approx \frac{1}{N}\sum_{s=1}^{N} f(\bar{x}^s)\varphi_p(x^s{}_i)$$

$$\beta_{qr}^{ij} = \frac{\int_{[0,1]^k} f(\bar{x})\varphi_q(x_i)\varphi_r(x_j)d\bar{x}}{\int_{[0,1]^2} \varphi^2{}_q(x_i)\varphi^2{}_r(x_j)dx_i dx_j} \approx \frac{1}{N}\sum_{s=1}^{N} f(\bar{x}^s)\varphi_q(x^s{}_i)\varphi_r(x^s{}_j) \tag{2.6}$$

Using MC integration approximation, these coefficients can be estimated by least squares regression on the model output, Y, obtained for the training matrix, M_{train} (Li *et al.*, 2001b). The mean output is obtained by approximating the integrals using summation terms,

$f_0 \approx \frac{1}{N}\sum_{s=1}^{N} f\left(\bar{x}^s\right)$. It is important to note that the entire coefficient in Equation 2.6 can be obtained simultaneously from a single set of MC samples. This leads to a substantial reduction in the computational cost due to the elimination of repeated sampling. As seen from most physical models, the number of samples required for accurate estimation of the coefficients is of the order of 10^3 (Li and Rabitz, 2012a).

Using Equations 2.4-2.6, the terms $\sum_i f_i(x_i) + \sum_{i,j} f_{ij}(x_i, x_j)$ amount to $\tilde{f}(\bar{x})$, which is an estimate predicted by RS-HDMR of $f(\bar{x})$ from Equation 2.1. It is essential to compare how well $\tilde{f}(\bar{x})$ represents $f(\bar{x})$. For this, we generated MC samples and evaluated the model output $Y = f(\bar{x})$ directly from the ODE and from the second order RS-HDMR based prediction, $\tilde{f}(\bar{x})$ and estimated the coefficient of determination (R^2) using the relation:

$$R^2 = 1 - \frac{\sum_{s=1}^{N}\left[f_s(\bar{x}) - \tilde{f}_s(\bar{x})\right]^2}{\sum_{s=1}^{N}\left[f_s(\bar{x}) - f_0\right]^2} \tag{2.7}$$

The numerator of the second term in Equation 2.7 denotes the sum of squares of residuals and the denominator denotes the total sum of squares (hence, proportional to the total variance). In the current application, the variables are chosen to be uniformly distributed, but future applications of RS-HDMR for detailed dynamic modeling will require estimation of the density functions of the rate parameters, which result in weighted component functions.

2.2.2.3 Sobol' sensitivity indices

We used the variance decomposition method to estimate the global sensitivity indices (or Sobol' indices). The Sobol' indices present the sensitivity of the output to specific perturbations of the input parameters. These indices are of various orders based on the number of parameters whose effects are studied. For example, first-order indices show the contribution of individual parameters, the second-order indices show the contributions of pairs of parameters, etc. Each index represents a fraction of the total variance. If σ^2 is the total variance of the model output in $[0,1]^k$, $Y = f(\bar{x})$, the decomposition of total variance assuming that contributions from higher order (≥ 3) are negligible is given by:

$$\sigma^2 = \sum_{1 \leq i \leq k} \sigma_i^2 + \sum_{1 \leq i < j \leq k} \sigma_{ij}^2 \tag{2.8}$$

Here, each term on the right hand side of Equation 2.8 signifies the "independent" contribution of the particular parameter combination. The total variance of the model output in *k*-dimensional space is estimated by the relation,

$$\sigma^2 = \int_{[0,1]^k} \left(f(\bar{x}) - f_0 \right)^2 d\bar{x} \tag{2.9}$$

By definition, the first order and second order Sobol' indices can be related to the variance by the relation:

$$S_i = \frac{\sigma_i^2}{\sigma^2}$$

$$S_{ij} = \frac{\sigma_{ij}^2}{\sigma^2} \tag{2.10}$$

The relations in Equation 2.10 show that the Sobol' index is a fraction of the total variance that is explained by the variance in the selected parameter combination.

Direct MC based evaluation

The estimation of Sobol' indices using direct Monte Carlo integration requires repeated sampling from the parameter space and therefore, is computationally expensive (Feil et al., 2009). For evaluation of Sobol' indices as given in Equation 2.10, the individual variances have to be evaluated first. The individual variances in Equation 2.8 can be obtained by:

$$\sigma_i{}^2 \approx \frac{1}{N-1}\sum_{s=1}^{N}\left(f\left(x_1^s,x_2^s,...,x_k^s\right)\times f\left(x_1^{s'},x_2^{s'},..,x^s{}_i,..,x_k^{s'}\right)\right)-f^2{}_0 \text{ and}$$

$$\sigma_{ij}{}^2 \approx \frac{1}{N-1}\sum_{s=1}^{N}\left(f\left(x_1^s,x_2^s,...,x_k^s\right)\times f\left(x_1^{s'},x_2^{s'},..,x^s{}_i,...,x_j{}^s,..,x_k^{s'}\right)\right)-f^2{}_0 \qquad (2.11)$$

The summation term in Equation 2.11 contains a product of the ODE based outputs obtained for two different MC samples, s and s'. For the second sample, s', all the variables except the ones under study are resampled from another MC matrix. In other words, s and s' are taken from two different random samples but with the same values of the parameters under study for that index. The total variance is calculated from:

$$\sigma^2 \approx \frac{1}{N-1}\sum_{s=1}^{N} f^2\left(\bar{x}^s\right)-f^2{}_0. \qquad (2.12)$$

In Equations 2.10-2.12, the functions, $f(\bar{x}^s)\equiv f(x_1^s,x_2^s,...,x_k^s)\equiv Y$ and are hence, obtained directly from the ODE outputs. These functions are different from the individual component functions ($f_i(x_i), f_{ij}(x_i,x_j)$) of RS-HDMR.

RS-HDMR based evaluation

The RS-HDMR component functions present a convenient way to calculate all the Sobol' indices using a single set of samples (Li *et al.*, 2001b; Li *et al.*, 2002b). The orthogonality of the RS-HDMR component functions allows the estimation of individual variances that are contributed independently (first-order) and jointly (higher order) by the input parameters. Using the component function definitions, the variance decomposition can now be written as:

$$\sigma^2 = \sum_{1 \le i \le k} \int_{[0,1]} f^2{}_i (x_i) dx_i + \sum_{1 \le i < j \le k} \int_{[0,1]^2} f^2{}_{ij} (x_i, x_j) dx_i dx_j \tag{2.13}$$

and consequently, the Sobol' indices can be written as:

$$S_i = \frac{\int_{[0,1]^1} f_i^2 (x_i) dx_i}{f_0^2} \quad \text{and}$$

$$S_{ij} = \frac{\int_{[0,1]^2} f_{ij}^2 (x_i, x_j) dx_i dx_j}{f_0^2} \tag{2.14}$$

Further, using the coefficients obtained described in Section 2.2.2.2, simple relationships can be obtained for the Sobol' indices:

$$S_i = \frac{\sum_{1 \le p \le u} \left(\alpha_p^i \right)^2}{f_0^2} \quad \text{and}$$

$$S_{ij} = \frac{\sum_{1 \le q \le v} \sum_{1 \le r \le w} \left(\beta_{qr}^{ij} \right)^2}{f_0^2} \tag{2.15}$$

As discussed in Section 2.2.2.2, the evaluation of the coefficients in Equation 2.15 requires only one set of MC samples, and by extension, this means that all the Sobol' indices (first and second order) can be obtained simultaneously. This is unlike the case for direct MC based evaluation in Equations 2.11-2.12, where a new set of samples must be generated for each index and hence, is dependent on the total number of parameters and their combinations. Our prime goal is to evaluate the Sobol' indices using RS-HDMR. However, direct MC based indices are also evaluated to compare the accuracy of the RS-HDMR estimates with the direct MC estimates. For direct MC based evaluation of the Sobol' indices, a large number of samples (~ 10^5 -10^6) are required (Feil *et al.*, 2009). For RS-HDMR based evaluation of the indices, lower number of samples (~ 10^3) is sufficient. However, for comparison between the direct MC based indices and RS-HDMR based indices, we chose 10^5 samples in all the plots of this chapter. We ensured that all the indices reached convergence by 10^5 samples and remained unchanged by further increase in sampling size (see Section 2.3.2). The entire process for GSA is represented as a schematic in Figure 2.2.

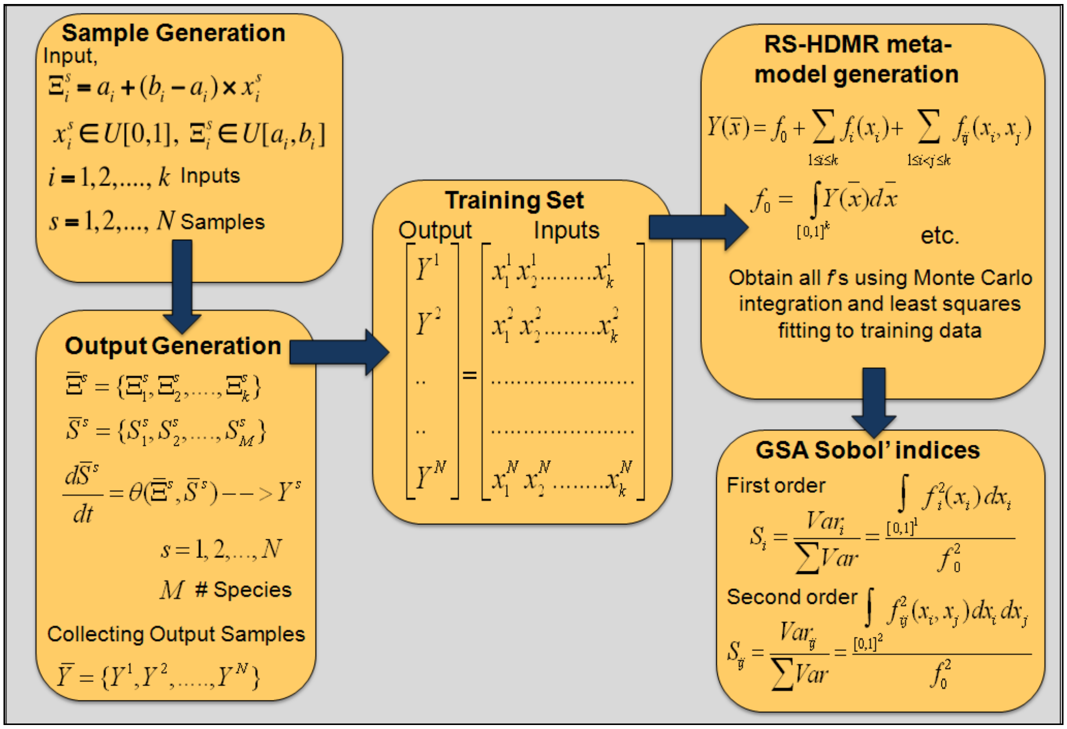

Sample Generation
Input,
$\Xi_i^s = a_i + (b_i - a_i) \times x_i^s$
$x_i^s \in U[0,1],\ \Xi_i^s \in U[a_i, b_i]$
$i = 1,2,....,k$ Inputs
$s = 1,2,...,N$ Samples
Output Generation
$\bar{\Xi}^s = \{\Xi_1^s, \Xi_2^s, ..., \Xi_k^s\}$
$\bar{S}^s = \{S_1^s, S_2^s, ..., S_M^s\}$
$\frac{d\bar{S}^s}{dt} = \theta(\bar{\Xi}^s, \bar{S}^s) --> Y^s$
$s = 1, 2, ..., N$
M # Species
Collecting Output Samples
$\bar{Y} = \{Y^1, Y^2, ..., Y^N\}$
Training Set
Output
Inputs
Y^1 Y^2 .. .. Y^N
=
$x_1^1\, x_2^1 ....x_k^1$
$x_1^2\, x_2^2 ....x_k^2$
$x_1^N\, x_2^N ....x_k^N$
RS-HDMR meta-model generation
$Y(\bar{x}) = f_0 + \sum_{1 \le i \le k} f_i(x_i) + \sum_{1 \le i < j \le k} f_{ij}(x_i, x_j)$
$f_0 = \int_{[0,1]^k} Y(\bar{x}) d\bar{x}$
etc.
Obtain all fs using Monte Carlo integration and least squares fitting to training data
GSA Sobol' indices
First order
$S_i = \frac{Var_i}{\sum Var} = \frac{\int_{[0,1]^1} f_i^2(x_i)\, dx_i}{f_0^2}$
Second order
$S_{ij} = \frac{Var_{ij}}{\sum Var} = \frac{\int_{[0,1]^2} f_{ij}^2(x_i, x_j)\, dx_i\, dx_j}{f_0^2}$

Figure 2.2 Workflow for the entire global sensitivity analysis using RS-HDMR.

We start with selection of input parameters and their intervals. Using several MC samples, the ODE model is simulated to obtain the dynamics of the model variables and the output of interest (like AUC_D). The resulting input-output matrix is utilized to generate the component functions of RS-HDMR followed by the variance based Sobol' indices.

2.2.3 *K*-means clustering

K-means clustering was performed on the dynamic profiles predicted by the model using MATLAB (R2010b, Mathworks, Natick, MA) function *kmeans*. This enabled identification of parameter ranges where the dynamics of the model outputs lie. Before analysis, the dynamic profiles were normalized by the maximum value per simulation. The normalized profiles were clustered using the ‘correlation distance’ as a metric since we are interested in the dynamics. Cluster quality was judged by the Silhouette value (S_i) defined as (Kaufman and Rousseeuw, 2009):

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \tag{2.16}$$

a_i = average distance from the i^{th} profile in the cluster to other profiles in the same cluster**,** b_i = minimum average distance from the i^{th} profile in a cluster to profiles in the other clusters. The minimum denotes that this distance is measured to the closest cluster. The Silhouette value ranges from -1 to 1, with -1 denoting a misplacement of the i^{th} profile and +1 denoting the best placement. The quality of the clustering process was determined by the mean Silhouette value (S_{mean}) (Kaufman and Rousseeuw, 2009). We selected a threshold of 0.6, and determined the number of clusters *k* with S_{mean} values greater than 0.6 (See Table A.2 for variation in S_{mean} with the number of clusters). For our data, three clusters were found to be optimal, beyond which no further improvement was observed in the cluster quality and no new dynamics was observed. Information theoretic based approaches are also commonly used to determine the optimal number of clusters, but comparisons on several datasets have failed to identify any particular advantage in using information theoretic criteria over the Silhouette value (Rendón *et al.*, 2011).

2.2.4 Experimental methods

2.2.4.1 Cell Culture

H1 ES cells were maintained with feeder-free conditions on matrigel-coated plates (hESC-qualified Matrigel, BD Biosciences, San Jose, CA, USA) in mTeSR1 (Stem Cell Technologies, Vancouver, BC, Canada) with the media changed daily. Cells were passaged every 6-7 days by mechanical agitation of the colonies and splitting at a 1:4-1:6 dilution. Cells were examined under the microscope daily and colonies with observable differentiation were removed before the media changes. All experiments were performed on cells in the passages ranging from 52 to 60.

2.2.4.2 Insulin stimulation time course experiments

H1 ES cells were kept in DMEM/F12 (Invitrogen, Carlsbad, CA, USA) and 0.2% Bovine Serum Albumin (BSA, Sigma-Aldrich, St. Louis, MO, USA) for 18h before insulin stimulation was carried out. After 18h, the cells were washed twice with 1x PBS and then insulin (Sigma-Aldrich, St. Louis, MO, USA) was added to fresh DMEM/F12, 0.2% BSA at a concentration of 100 nM. A high insulin concentration that is well in the range of *in vitro* cell culture systems was selected (Kiselyov *et al.*, 2009). The cells were stimulated for 120 min. After stimulation, the cells were analyzed for proteins using Luminex xMAP technology (Luminex, Austin, USA). Further details of the procedure are presented below under Luminex xMAP technology.

2.2.4.3 Experimental modulation of PKCζ, PTEN and PTP levels

GÖ6983 (Calbiochem, Billercia, MA, USA) was used to inhibit PKC-ζ (Zheng *et al.*, 2000). Cells were treated with GÖ6983 at a concentration of 10 µM for 20 minutes in mTeSR1 medium. For PTEN and PTP inhibition, cells were treated with dipotassium bisperoxo (5-hydroxypyridine-2-carboxyl) oxovanadate (V) or bpV(HOpic) (Calbiochem), for 24 hr at 100 nM in mTeSR1 medium. For PTEN + PTP inhibition, same chemical was used at 1 µM for 20 min. Both concentrations were based on previous reports (Schmid *et al.*, 2004). Further details of quantitative analysis are presented below under LiCOR western analysis.

2.2.4.4 Multiplex protein measurements using Luminex xMAP technology

Cell lysis

After the specific treatments, cell lysis was carried out in Cell Extraction Buffer (Invitrogen). The medium was removed and washed twice with 1x PBS. The cells were lysed and then incubated on ice for 30 minutes, with cellular debris removed by centrifugation at 3200 XG for 30 minutes. Protein concentration was measured using the BCA assay kit (Thermo Scientific, Rockford, Illinois, USA). Equal amounts of total proteins (25µg) were used for subsequent analyses.

Luminex analysis

Proteins IR (pY1162/1163), IGF-1R (pYpY1135/1136), IRS-1 (pS312), and Akt (pS473) were simultaneously measured in the same cell lysate using the AKT Pathway Phospho (Catalog no. LHO0001M) and AKT Pathway Total (Catalog no. LHO0002M) magnetic 7-Plex panels (Invitrogen) using MagPix Luminex xMAP technology (Luminex, Austin, USA). The assay was setup with standards and samples according to the manufacturer's instructions. All incubations were performed at room temperature. Total median fluorescence intensity (MFI) and the calibration curves were used to estimate the phosphorylated proteins (in units/ml) and total proteins (in ng/ml). The actual level of phosphorylation was calculated using the ratio of phosphorylated form to total form and expressed as units phosphorylation per ng of total protein. The panel did not contain tyrosine phosphorylated IRS1, and therefore we decided to analyze p-IRS1 (Y) using quantitative western blot in LiCOR (LiCOR Biosciences, Lincoln, NE, USA).

2.2.4.5 Western blot using LiCOR

The proteins were separated using 4-20% pre-cast SDS-PAGE at 100V and then transferred to nitrocellulose membrane at 4^0C overnight at 20V. The membrane was blocked with Odyssey blocking buffer (LiCOR) for 2 h at room temperature. Primary antibody against AKT (pS473) (Cell Signaling, Beverly, MA, USA), PKCζ (pTpT403/410) (Cell Signaling), IRS-1 (pY612) (EMD, Millipore, USA), and GAPDH (Santa Cruz, Dallas, TX, USA) were diluted (1:1000) in Odyssey blocking buffer with 0.1% Tween-20 (Hercules, CA, USA) and incubated overnight at 4^0C on a shaker. Thereafter, the membrane was washed with PBS containing 0.1% Tween-20 for 5 times at 5 minutes each. IR conjugated anti-rabbit secondary antibody diluted (1:20000) in Odyssey blocking buffer with 0.1% Tween-20 was added and incubated for 1 h at room temperature on a shaker. Before scanning in the Odyssey Imager, the membrane was again washed 5 times at 5 minutes each with PBS containing 0.1% Tween-20. Appropriate molecular weight markers were run for each analyzed protein. Densitometric analysis was performed using Image studio (LiCOR) and normalization was done with GAPDH values. Later, the change in the protein phosphorylation was expressed as fold change with respect to untreated cells.

2.3 RESULTS

2.3.1 Selection of parameter ranges for PI3K/AKT pathway in hESCs

Before beginning to analyze the steady state of the PI3K/AKT pathway, we first determined if the structure of Sedaghat model is sufficient to address hESC features of dynamics. In the original Sedaghat model (Sedaghat *et al.*, 2002), several phosphorylation and dephosphorylation rate constants of *Module 2* were fixed by equilibrium ratios. For extension to hESCs, we decoupled these rates by varying them independently in a realistic range. These rate constants include: k_7/IR_p, k_{-7}, k_8, k_{-8}, $k_{9\text{stim}}$ and k_{-9}. To determine their biologically realistic ranges, we first measured the dynamics of key molecules of the pathway under insulin stimulation in H1 hESCs. Representative molecules from different positions of the signaling pathway were selected for analysis: early (p-IR, p-IGF1R), mid (p-IRS1 (Y), p-AKT) and late (p-IRS1 (S)). Figure 2.3 presents the dynamics of the measured phospho-proteins. p-AKT showed an overshoot behavior and it settled at intermediate values of 3-folds by 120 min (Figure 2.3A). The levels of p-IR and p-IGF1R increased rapidly to 2- and 8-fold within 15 minutes and remained at these levels till the end of stimulation (Figure 2.3A-B). Among the IRS1 molecules, tyrosine phosphorylation (p-IRS1 (Y)) showed rapid increase with maximum levels reached by 15 min and then a down-regulation to intermediate values which remained constant upto 120 min (Figure 2.3B). Serine phosphorylation of IRS1 (p-IRS1 (S)) showed an initial dip followed by up-regulation at 30 min and then stabilization to basal levels. A distinct negative correlation was observed between p-IRS1 (Y) and p-IRS1 (S) dynamics.

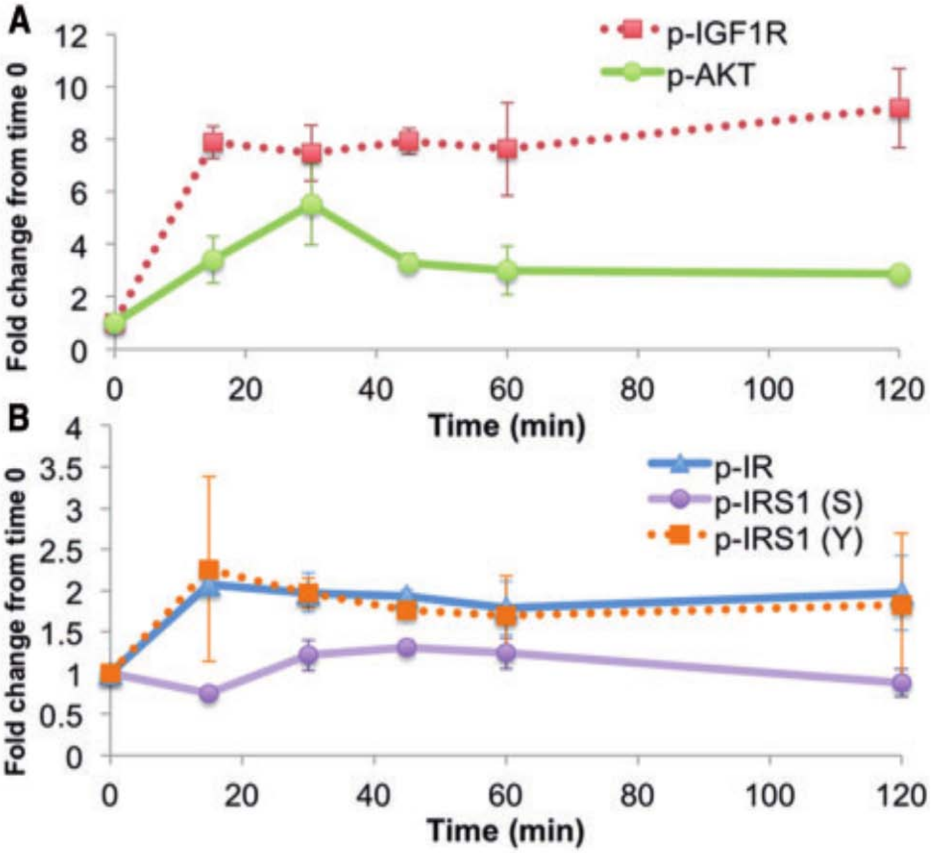


Figure 2.3 Experimental dynamics of pathway molecules in hESCs.

(A) p-IGF1R and p-AKT (B) p-IR, p-IRS1 (Y) and (S). The phosphorylated protein measurements are normalized to corresponding total protein control and to time point zero. Number of repeats for each time point = 3. For p-AKT and p-IRS1(S), the mean values at the peak and steady state (at 120 min) were found to be different with a p-value of <0.01 and <0.05 respectively (calculated using a two-sample student t test with unequal variances). The peak value of p-IRS1 (Y) was not significantly different as compared to the steady state.

Here, we are interested in finding out the parameter ranges that explain the overshoot behavior and intermediate steady state levels of p-AKT. It was observed that there is a delay (~15 min) in p-AKT peak after IRS1 tyrosine phosphorylation. In other cell lines, peak in p-AKT (S473) activation occurs parallel to IRS1 tyrosine phosphorylation. Currently there is no indication on what could biologically result in this delay, although recent work points to a delay in AKT translocation from the cell membrane (due to retention) to the intracellular medium, although the actual molecular players are not identified (Nim *et al.*, 2015). Later in this chapter, we are interested only in the steady state levels of the pathway and a delay due to this mechanism will not affect the long-term steady state levels. Hence, here we are focusing only on capturing the overshoot behavior (See Section 2.5 for additional comments). To select parameter ranges, we first explored the behavior of the relaxed Sedaghat model over a broad parameter range (100-fold around nominal values). The basal levels of *PTP*, *SHIP* and *PTEN* and initial concentration of insulin were also included as additional parameters, resulting in 25 free parameters. 10^5 random samples were drawn from a uniform distribution of 25 parameters and the ODE model was integrated for each sample till 120 min. This being the first modeling effort of the pathway for hESCs, the actual parameter distribution is not known *a priori*. Hence, a uniform distribution was used here. The profiles were normalized to the maximum level and clustered using *k*-means algorithm on p-AKT dynamics. There were three major clusters (Figure 2.4A-C, Table A.2). Among these, clusters 2 and 3 showed the characteristic overshoot behavior, but only cluster 3 showed intermediate steady state values as seen in the experiments in Figure 3A. We also saw good correlation between experimental and clustered profiles for other molecules, p-IR and p-IRS1 (Y).

Cluster 3 being our primary cluster of interest, we wanted to select parameter ranges defining this cluster. Hence, we first analyzed parameters primarily segregating the three clusters. It was observed that of the 25 parameters, only 6 were varying between the 3 clusters: k_7/IR_{p}, k_{-7}, k_8, k_{-8}, $k_{9\mathrm{stim}}$ and k_{-9}. Among these, the activation parameters were high and deactivation parameters were comparatively low in cluster 3. The distribution for de-activation parameter, k_{-7} for each cluster is shown in Figure 2.4D. The distributions for remaining parameters are shown in Figure A1. Based on these distributions, we narrowed the range of the 6 parameters around the peaks for cluster 3 (Figure 2.4D and Figure A.2). The chosen restricted ranges of the parameters are presented in Table A1. We next performed GSA in the restricted parameter ranges to identify the key parameters regulating the self-renewal state of hESCs.

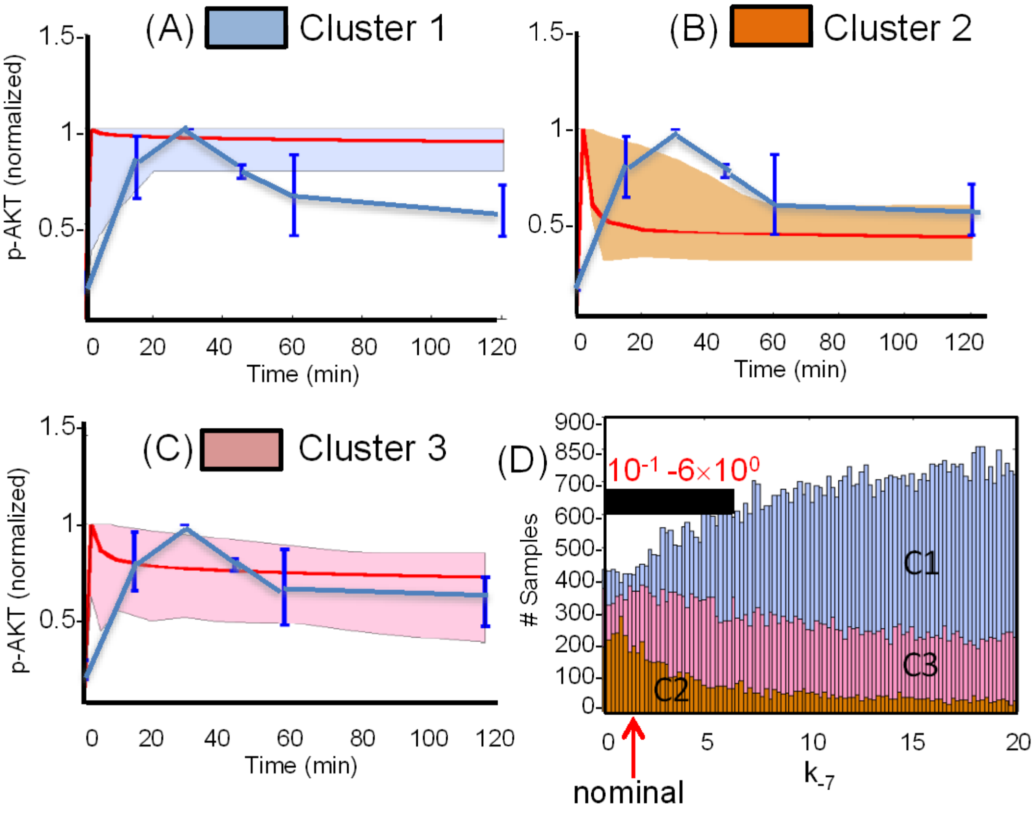


Figure 2.4 *K*-means clusters for p-AKT.

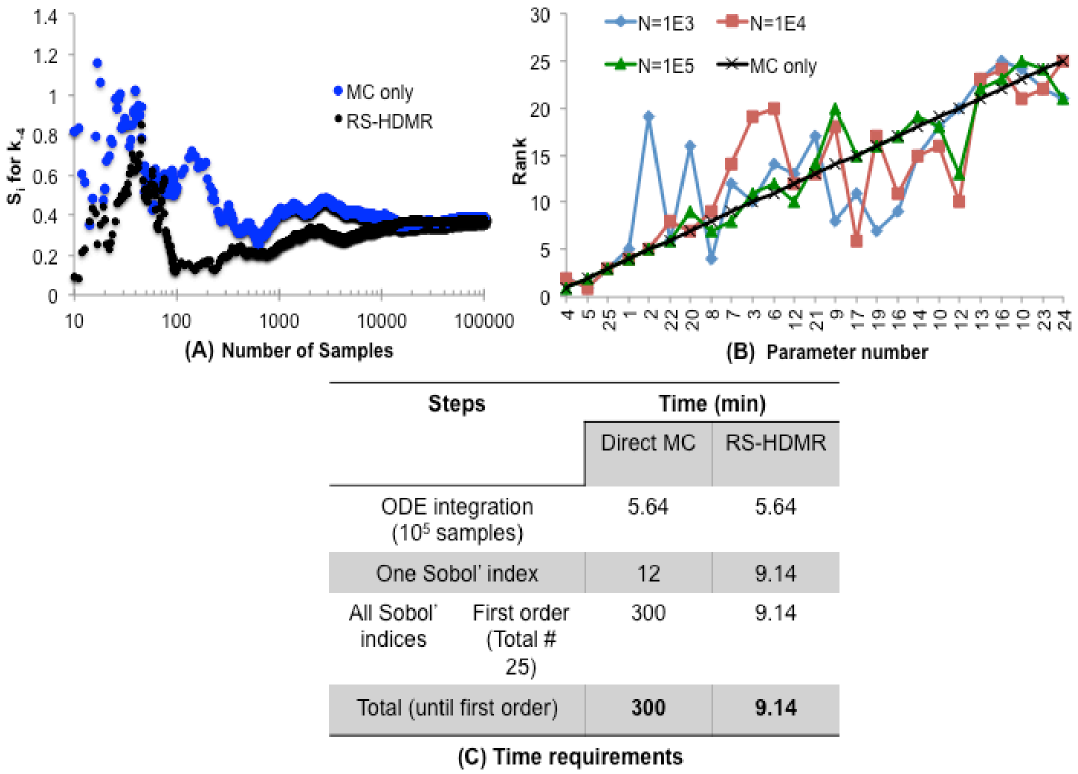
(A-C) Three key dynamic clusters observed in the parameter space. Cluster 3 showed overshoot behavior and intermediate steady state levels. Cluster centroid is shown by red curve and the shaded region shows the cluster extent. Blue lines indicate the overlaid experimental data. Hierarchical clustering analysis gave same type of major clusters (See Figure A.3). (D) Parameter k-7 contained in clusters, C1, C2, C3 from (A-C). The black bar indicates the final selected range for sensitivity analysis and the red arrow shows the location of the nominal value.

2.3.2 Meta-model representation and efficient GSA

To reduce the computational cost associated with GSA, we adopted RS-HDMR (Li and Rabitz, 2012b), to explore the model IO behavior and also to rank the parameters using Sobol' sensitivity indices. The performance characteristics of the method has been validated for diverse physical systems (Li and Rabitz, 2012b). Here, we have applied the technique for a high dimensional, nonlinear signal transduction model. To check the validity of RS-HDMR and its computational efficiency, we compared the Sobol' indices estimated using RS-HDMR with direct MC evaluations for one of the output molecule, p-IR.

First we analyzed the effect of sample size on the sensitivity index evaluated both by direct MC and by RS-HDMR. Direct MC identified the basal rate of receptor recycling, k_{-4}, as the most sensitive parameter with a Sobol' index (S_{i}) of 0.38. Figure 2.5A compares the convergence of Sobol' indices for this parameter by the two methods, for sample size (N) ranging from 10^2 to 10^5. We see that the two methods converge to the same value by a sample size of 10^4. Relative ranking of parameters are often more informative than the actual value of the sensitivity index. Hence, Figure 2.5B represents the ranking of the parameters obtained by direct MC analysis for 10^5 samples and compares it with RS-HDMR predictions at different sample sizes. Overall it is observed that the parameters with higher sensitivity were predicted with great accuracy even at very low sampling of 10^3. Beyond the fourth-ranked parameter for 10^3 samples and eighth-ranked parameter for 10^4, there is considerable deviation from MC analysis. The higher sampling size of 10^5, however, closely predicts the direct MC estimates for most of the parameters throughout the range. Hence, a sample size of 10^5 was chosen for the remaining work.

The primary motivation for adopting RS-HDMR is the computational efficiency of the algorithm. The computational cost of traditional GSA is primarily associated with the need for repeated sampling to evaluate the integrals. In contrast, using polynomial approximations in RS-HDMR, both low and high order Sobol' indices can be estimated simultaneously from one set of MC samples. As an example we have tabulated the typical time requirements for first order indices in Figure 2.5C. Obtaining all the first order Sobol' indices by RS-HDMR requires ~ 9 min while it takes 300 min for direct MC. In RS-HDMR, a critical source of error is the MC integration approximation for the high dimensional integrals. The error of this approximation is inversely proportional to the sample size as $N^{1/2}$ and favorably independent of the dimension (Li and Rabitz, 2012b).



Steps		Time (min)	
		Direct MC	RS-HDMR
ODE integration (10^5 samples)		5.64	5.64
One Sobol' index		12	9.14
All Sobol' indices	First order (Total # 25)	300	9.14
Total (until first order)		**300**	**9.14**

(C) Time requirements

Figure 2.5 Comparison between RS-HDMR and direct MC for p-IR output.

(A) Convergence of the Sobol' index for k-4 with sample size (N). (B) Ranking of parameters with different samples sizes for RS-HDMR as compared to direct MC for 105 samples. (C) Typical computational time requirements for 25 input parameters and first order sensitivity evaluation for both RS-HDMR and direct MC. For RS-HDMR, all the Sobol' indices are evaluated simultaneously.

2.3.3 Sensitive parameters for key molecules of the pathway

Upon confirming the accuracy of second order RS-HDMR for the current model, we next determined the globally sensitive parameters for the steady state of four molecules of the pathway: p-IR, p-IRS1 (Y), p-IRS1 (S) and p-AKT (Figure 2.6A). The resulting output distributions are presented in Figure A.4 and overall performance in Table A3. We see that the sensitivity contribution of each parameter to the different outputs is different. In *Module 1*, rates of recycling of the non-phosphorylated receptors, k_{-4} and active receptor internalization, $k_{4'}$ were important. These parameters primarily changed the steady state levels of the active receptors p-IR and also affected p-IRS1 (Y) and p-AKT levels to a small extent. The initial insulin levels and the binding rate of insulin to the receptors, k_1, were the other parameters affecting p-IR but they did not affect the other molecules. The important parameters from *Module 2* were primarily associated with negative regulators of the pathway. These included many of the deactivation rates: de-phosphorylation of p-IRS1 (Y), k_{-7}, deactivation of PI3K, k_{-8}, and de-phosphorylation of PIP_3 to PI(4,5)P2, k_{-9}. All these parameters were upstream of PIP_3 and affected the p-AKT levels considerably. p-IRS1 (Y) was affected mostly by k_{-7} while p-IRS1 (S) was affected mostly by k_{-8} and k_{-9}. Another important set of parameters was associated with negative feedback by p-PKCζ and subsequent serine phosphorylation of IRS1. These included the Hill Equation parameters, $V_{\max}$, K_d, n and the IRS1 serine phosphorylation rates, $k_{7'}$ and $k_{-7'}$. These parameters significantly affect the p-IRS1 (S) levels followed by p-IRS1 (Y) and p-AKT. Thus, it is seen that for the intracellular molecules, most of the sensitive parameters are 'negative regulators' of molecules upstream of PIP$_3$ or they are associated with ‘negative feedback’. Additionally, for p-AKT there were also important contributions from the second order indices.

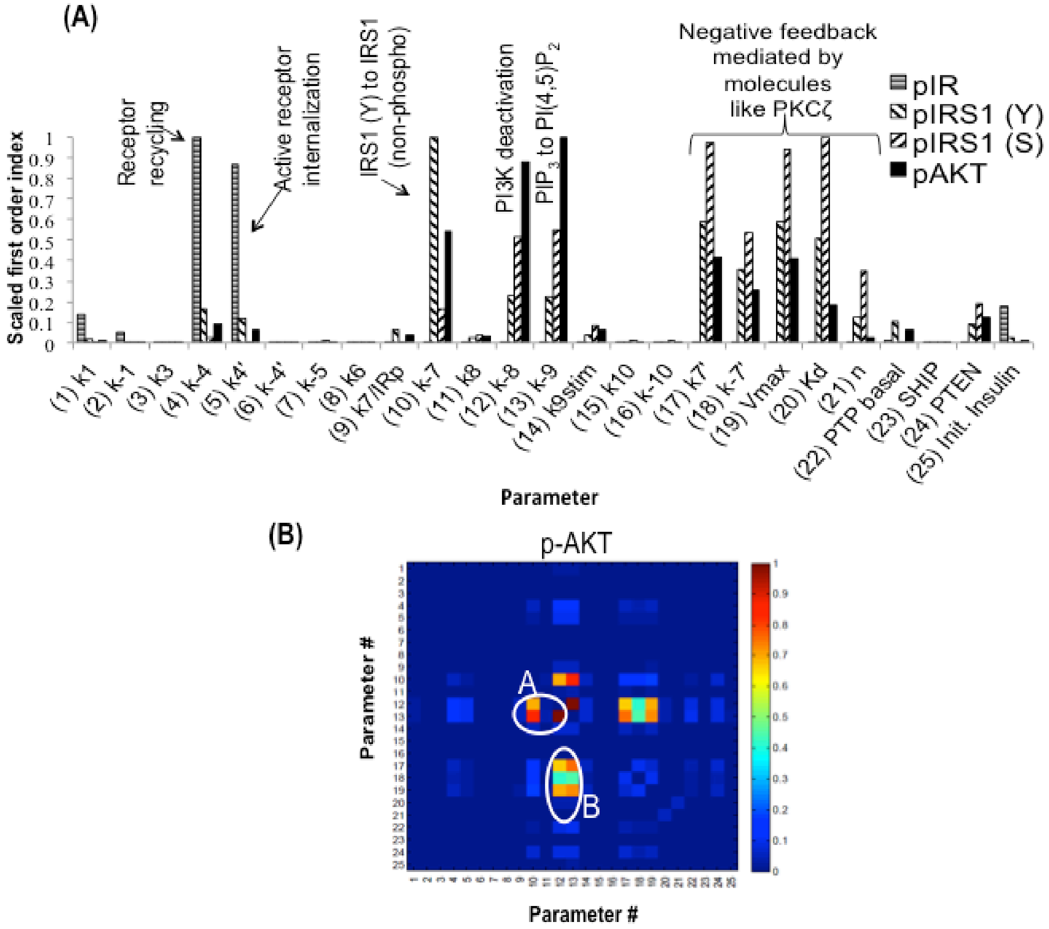


Figure 2.6 Results from second order RS-HDMR analysis.

(A) Scaled first order RS-HDMR based Sobol’ indices. The indices have been scaled by the maximum value for each output to show the relative importance of the parameters. p-IR is only affected by internalization and recycling processes. For the other molecules, negative regulators upstream of PIP3 and negative feedback by serine IRS1 are the most sensitive. (B) Scaled second order Sobol’ indices of p-AKT and the two groups with important interactions.

For p-AKT, the second order indices contributed to 28% of the variance. Figure 2.6B presents the heat-map of the scaled second order Sobol’ indices. The sensitive parameters were found to cluster in two groups: Group A: interaction between negative regulators upstream of PIP$_3$ (~7%) and Group B: interaction between negative feedback parameters with negative regulators from Group A (~6%) (Table A.4). It is important to note that these parameters also had important first order contributions. Second order interactions further increase the sensitivity of these parameters (Figure A.5).

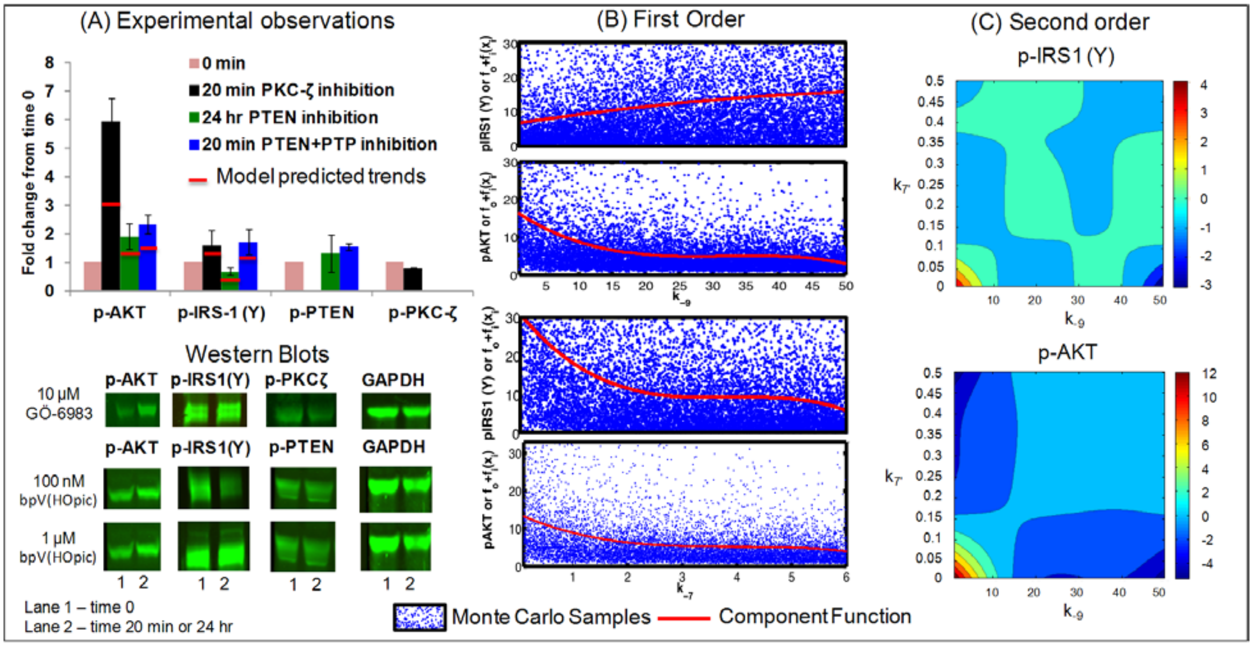


Figure 2.7 Comparison of experimental and model analysis of sensitive processes in self-renewing hESCs. (A) Effect of perturbation of (1) negative feedback by GÖ-6983 at 1 μM, (2) PTEN by bpV(HOpic) at 100 nM and (3) PTEN+PTP by bpV(HOpic) at 1 μM. The top bar graph shows the quantitative analysis of western blots (bottom panel) using LiCOR image analysis software. Note: Missing bars indicate that the proteins were not analyzed in that experiment. ODE model predictions of fold change in p-AKT and p-IRS1 (Y), when p-PKCζ, PTEN and PTP are perturbed by the same amount as the experimental data, are shown by red lines overlaid onto the experimental bars of p-AKT and p-IRS1 (Y). All parameters are at nominal values. (B) First order component functions showing the effect of parameter k_{-9} and k_{-7} on p-IRS1 (Y) and p-AKT. Red curve is the first order component function and blue scatter points are the ODE model Monte Carlo outputs used to construct RS-HDMR functions. The scatter points represent the influence of variability in the other parameters (See Figure A.4 for output density). (C) Second order component functions showing interaction between k-9 and k7' on p-IRS1 (Y) and p-AKT.

2.3.4 Experimental validation of key sensitive processes

As seen from the previous section, the model predicted sensitive perturbations on the system include (1) negative feedback via p-PKCζ and (2) negative regulators that affect the de-phosphorylation of PIP_3 (Summarized in Figure 2.1). To validate the sensitivity of these processes in hESCs, we performed targeted perturbations and we carefully chose perturbations that would result in increase in p-AKT levels to avoid differentiation. For negative feedback, the ideal candidate for perturbation is p-PKCζ and for de-phosphorylation rates upstream of PIP_3, two such candidates exist, namely PTEN and PTP.

2.3.4.1 Influence of negative feedback modulation on self-renewing hESCs

Self-renewing hESCs were subjected to 20 min treatment of 10 µM GÖ6983 (PKC -ζ inhibitor) based on Zheng *et al.* (Zheng *et al.*, 2000). In our experiments, we see large change in p-AKT (6-fold) and moderate change in p-IRS1 (Y) (1.5-fold) levels for small decrease in p-PKCζ level (Figure 2.7A). Similar trends were predicted by model simulations (Figure 2.7A, red lines). Thus, small changes in the strength of negative feedback propagated to large changes in the levels of p-AKT. While sensitivity indices indicate the importance of a parameter on a specific output, the directionality of the effect, positive or negative, cannot be directly deduced from Sobol' indices. The meta-model representation of RS-HDMR is particularly suited for such deduction as this information is contained in the hierarchical component functions of the decomposition. Increasing the strength of negative feedback (increasing $k_{7'}$) decreases p-AKT and p-IRS1 (Y) (Figure A.6). Therefore, we see a positive correlation between p-AKT and p-IRS1 (Y). The sensitivity based on first order indices show smaller increase in p-AKT, but the experiments clearly show a large increase in p-AKT levels. We envisage this to be the effect of nonlinear influence of the other sensitive processes. For example, model simulations show that under complete inhibition of negative feedback, the levels of p- AKT will rise considerably if the influence of negative regulators like PTEN are low to begin with (low k_{-9}) or the levels of positive regulators like PIP_3 are high (Figure A.7).

2.3.4.2 Influence of PIP3 dephosphorylation on self-renewing hESCs under basal PTP levels

Using a direct inhibitor of PTEN, bpV(HOpic), we studied its effects on p-AKT and p-IRS1 (Y) (Schmid *et al.*, 2004). At a low concentration of 100 nM for 24 h, the compound is known to suppress active PTEN, thereby increasing inactive p-PTEN. Our experimental data shows that small change in PTEN resulted in 2-fold increase in p-AKT (Figure 2.7A). The levels of p–IRS1 (Y) showed a 1.5-fold decrease. Similar trends were predicted by model simulations (Figure 2.7A, red lines) and by RS-HDMR (Figure 2.7B). This effect is primarily because of the strengthening of negative feedback leading to indirect inhibition of p-IRS1 (Y). This also points to the fact that p-AKT is more sensitive to PIP_3 levels as compared to p-IRS1 (Y).

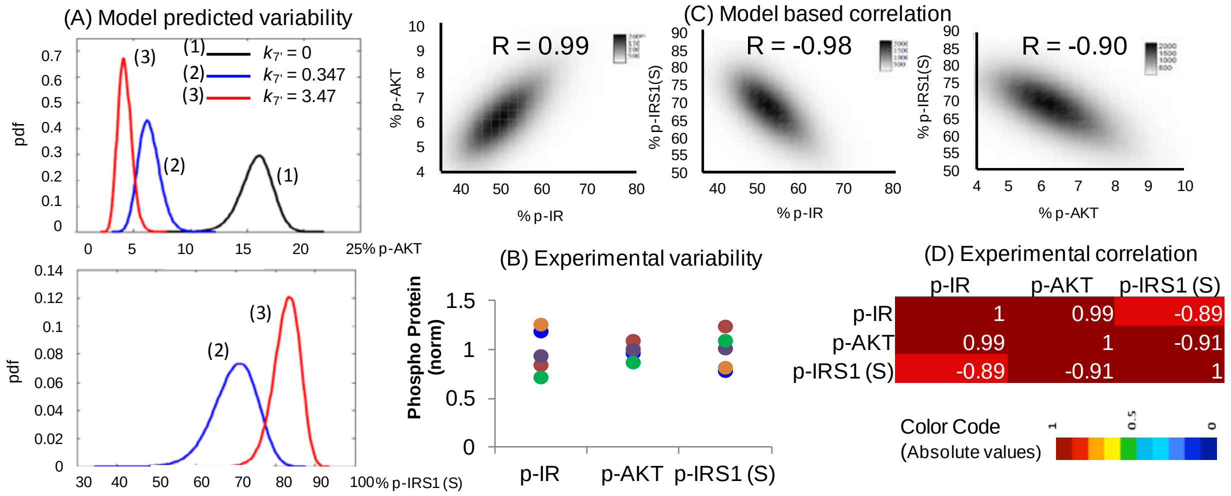
2.3.4.3 Influence of PIP3 dephosphorylation on self-renewing hESCs under PTP inhibition

Next we validated the effect of combined PTEN and PTP inhibition to check if PTP inhibition increases p-IRS1 (Y) when PTEN is still inhibited. At higher concentrations, the same complex bpV(HOpic) can inhibit both PTEN and PTP. We treated hESCs with 1 μM inhibitor for 20 min following Schmid *et al.* (Schmid *et al.*, 2004). Our experimental results show a proportional increase in all the three molecules, p-AKT, p-IRS1 (Y) and p-PTEN (Figure 2.7A). Similar trends were predicted by model simulations (Figure 2.7A, red lines) and RS-HDMR (Figure 2.7B). The increase in p-IRS1 (Y) is primarily due to PTP inhibition (decrease in k_{-7})

since PTEN inhibition alone resulted in a decrease in p-IRS1 (Y). Thus, under PTEN + PTP inhibition, p-IRS1 (Y) overcomes the effect of increase in downstream negative feedback. In our experimental studies we observed that the largest increase in p-AKT levels was brought about by the inhibition of negative feedback. This was also predicted by the second order RS-HDMR component functions for p-AKT. The second order component functions for one such combination, k_{-9} and $k_{7'}$ is presented as a heat-map in Figure 2.7C for p-AKT and p-IRS1 (Y). For low k_{-9} and low $k_{7'}$ there was significant positive contribution to the p-AKT levels but not for p-IRS1 (Y). Combining this with the first order contribution from k_{-9} and $k_{7'}$ in this regime and together with the mean, f_0, we get the total p-AKT of 47% ($f_0 + f\ (k_{-9}) + f\ (k_{7'}) + f\ (k_{-9}, k_{7'})$). Alternatively, this effect is reduced when the negative feedback is strengthened. For example, in the low k_{-9} and high $k_{7'}$ regime, the same contribution to p-AKT amounts to 23%. Therefore, strong negative feedback can considerably decrease the sensitivity of other reactions involving PTEN and PTP.

2.3.5 Robustness of system behavior under parameter uncertainty

While mathematically representing hESC systems, it is important to consider the effect of variability as observed in different experimental repeats. To test how parameter uncertainty influences the variability in the model output, we chose a biologically realistic log-normal distribution of the parameters centered around the nominal values. From the negative feedback parameters, the value of the most sensitive parameter, $k_{7'}$ was varied as follows: (1) no negative feedback, $k_{7'} = 0$ (2) nominal level of feedback, $k_{7'} = 0.347$ and (3) 10 times stronger feedback, $k_{7'} = 3.47$. Figure 2.8A presents the probability distribution of steady state levels of p-AKT and p-IRS1 (S) respectively when exposed to parametric uncertainty, under these different levels of negative feedback. The model shows that strengthening the feedback parameter reduces variance in the distribution for each molecule. Hence the robustness of the system to input perturbations is enhanced in the presence of a strong negative feedback. In addition it was observed that the distribution for p-AKT was narrower as compared to p-IRS1 (S). In order to verify this observation, we plotted steady state levels of p-IR, p-IRS1 (S) and p-AKT from 5 different experimental repeats in Figure 2.8B. We see a comparatively high variability in p-IR and p-IRS1 (S) levels but interestingly, p-AKT shows a narrow range of variability (see cell-to-cell distribution in Figure A.8). Negative feedback, thus, plays an important role in maintaining robust levels of the important molecules under experimental variability. Additionally, Pearson pairwise correlation between the molecules showed a good agreement between the experiments and model predictions (Figure 2.8C-D).



	p-IR	p-AKT	p-IRS1 (S)
p-IR	1	0.99	-0.89
p-AKT	0.99	1	-0.91
p-IRS1 (S)	-0.89	-0.91	1

Figure 2.8 Influence of sensitive parameters on variability observed in hESCs.

(A) Steady state p-AKT (top) and p-IRS1 (S) (bottom) distributions under varying rate parameter values. The parameters identified to be most sensitive by RS-HDMR were varied assuming a log-normal distribution around the mean and a variance of 10%. Negative feedback strength was varied by keeping the parameter $k_{7'}$ at 0 (no feedback), 0.347 (nominal case), 3.47 (strong feedback). (B) Variability observed in the steady state from insulin stimulation experiments in Figure 2.3. The data was normalized with the mean value across 5 repeats. (C) Model predicted pairwise Pearson correlation between molecules for varying negative feedback. (D) Pairwise Pearson correlation between molecules in experiments from Part B.

2.4 DISCUSSION

2.4.1 Meta-model approximation for GSA of complex biological signal transduction models

In this aim we, for the first time, present a detailed analysis of the regulatory interactions in PI3K/AKT pathway actively maintaining the self-renewal state of hESCs. A key step in our workflow is the analysis of the uncertainty associated with the strength of interactions in the PI3K/AKT pathway of hESCs using a meta-model based GSA. GSA captures the complete nonlinear associations between the model parameters in a sufficiently wide region of the parameter space and is suited for non-linear systems. The traditional methods of GSA are variance decomposition schemes involving exhaustive exploration of the parameter space. This renders the use of a detailed parametric analysis of large-scale ODEs expensive restricting the modeler to fairly simple local analysis. To explore IO relationships efficiently, we adopted meta-model approach called RS-HDMR to obtain accurate information on the sensitive model parameters. RS-HDMR constructs a complex surrogate function to replace the ODE model and evaluates MC integrals of the Sobol' indices efficiently. The method has been proven to perform well in a variety of engineering systems where parameter uncertainty is a norm. In the current work the method was explored for a signal transduction model in a 25 dimensional parameter space. We demonstrated that the method is especially accurate in identifying the most sensitive parameters and their functional relationships to the output. Also, the technique, being independent of the total number of parameters can be applied to larger models common in systems biology. Hence, it is a promising alternative to evaluate global sensitivities with computational efficiency, instead of settling for locally based approximations as commonly done in signal transduction studies.

2.4.2 Implications of GSA for molecules of the PI3K/AKT pathway

Through our systems level analysis, we found that parameters associated with post receptor processes can affect the levels of intracellular molecules of the PI3K/AKT pathway more than the receptor level processes for high insulin concentrations. The high sensitivity of the post receptor processes is in support with previous experimental and modeling analyses that show that the functionality of insulin signaling can be severely affected by mutations associated with post receptor signaling molecules (Nyman *et al.*, 2012). Additionally, on removal of equilibrium relationship between the forward and backward reactions, it was observed that the dephosphorylation reactions of the direct cascade were highly sensitive while the phosphorylation reactions were comparatively insensitive as seen in other systems like the MAPK/ERK pathway (Yoon and Deisboeck, 2009). This is an important relation since many of the de-phosphorylation reactions are dependent on the concentrations or functionality of phosphatases that can vary widely with the cell type and state and are also implicated widely in diseased states (Yoon *et al.*, 2010). Our analysis also shows that due to the competing nature of the reactions in this module of the pathway, there is a considerable nonlinear outcome to simultaneous changes in the sensitive parameters, for example reduction of sensitivity to PTEN and PTP inhibition by strengthening negative feedback. Such nonlinear behavior was seen in an experimental study in hESCs, where inhibition of a direct phosphatase to p-AKT rendered the p-AKT levels insensitive to PI3K inhibition (Yoon *et al.*, 2010). Our current analysis thus highlights the importance of nonlinear interactions in determining the effect of perturbations on the pathway components. This is an important outcome of mechanistic modeling and will prove useful in the design of targeted interventions for many systems.

2.4.3 Modeling self-renewal in hESCs

2.4.3.1 Processes affecting pathway dynamics

Insulin stimulation experiments in self-renewing hESCs showed an overshoot behavior in the dynamics of post receptor molecules. Two possible candidates have been identified to explain such overshoot behavior in other cell types: (1) receptor internalization (2) downstream negative feedback from still unknown regulators of receptor de-phosphorylation (Nyman *et al.*, 2012). Usually, there is combined contribution from both the processes. In our hESC system, however, we do not see substantial overshoot behavior in p-IR dynamics, but surely there is a clear decrease in downstream p-IRS1 (Y) levels and an accompanying increase in p-IRS1 (S) levels. This indicates that the negative feedback acting at the level of IRS1 is responsible for decrease in p-IRS1 (Y). This was also seen from the negative correlation between p-IRS1 (Y) and p-IRS1 (S). Our clustering analysis showed that many of the de-phosphorylation reactions above PIP_3 had to be maintained at low levels and it was necessary to couple this with an existing negative feedback to see an overshoot behavior.

2.4.3.2 Processes affecting p-AKT levels

The central molecule like p-AKT can counterbalance the mechanisms that may lead to differentiation and support mechanisms that can lead to self-renewal (Singh *et al.*, 2012b). Any increase in p-AKT levels has been shown to result in increased stability and self-renewal capacity of hESC cultures. For example, the levels of the active form of self-renewal molecule c-MYC can increase with increase in p-AKT levels (Yoon *et al.*, 2010). Yet there are limited efforts to understand how regulatory mechanisms affect long-term maintenance of self-renewal in hESCs, which was the focus of the current study. Under the current culture conditions the receptor level processes were found to be less sensitive. Therefore, a promising strategy to increase p-AKT levels is inhibition of internal signals that suppress p-AKT. Our results suggest that inhibition of negative feedback via PKC-ζ is one such mechanism. A parallel experimental study has recently demonstrated the positive attribute of PKC inhibition in hESC self-renewal, but did not offer any mechanistic insight (Gafni *et al.*, 2013). Additionally, model analysis shows that any perturbation in the phosphorylation and de-phosphorylation reactions of this pathway (for example, PTEN and PTP inhibition) would still need the removal of negative feedback mechanisms to increase sensitivity to these interventions.

From uncertainty propagation analysis, we show that negative feedback also increases the robustness of p-AKT levels to variations in the levels of upstream molecules. Mechanisms like negative feedback are known to impart robustness in many biological systems. Interestingly, the steady state correlation between molecules of the pathway held under experimental variability. In conclusion, the strength of negative feedback needs to be maintained in a fine balance. Weakening the negative feedback is favorable for self-renewal but is associated with increased variability.

2.5 CONCLUSIONS AND FUTURE EXTENSIONS

2.5.1 Major conclusions

The current work has developed a mathematical structure of the PI3K/AKT pathway, validated by experiments, to describe self-renewing hESCs. Adoption of RS-HDMR, a powerful meta-modeling technique, allowed feasible evaluation of GSA of the complex non-linear pathway. An important conclusion from our study is that the maintenance of p-AKT levels, and hence the self-renewal state of hESCs, is controlled by many of the negative processes of the pathway. Additionally, nonlinear interactions identified by RS-HDMR show that the existing negative feedback plays an important role of desensitizing the pathway to input perturbations and thus, regulates the steady state distribution of molecules in self-renewing hESCs. Inhibition of negative feedback can significantly increase p-AKT levels and support self-renewal, but with a tradeoff associated with increased variability. Such mechanistic analysis of new systems like hESCs is a critical step towards identification of new targets for optimizing cell culture conditions.

2.5.2 Assumptions, potential pitfalls and proposed extensions

The current model of the pathway is sufficient to explain the steady state behavior of the molecules. However, in order to utilize the model to explain early behavior, it will be necessary to model the delayed time to peak of p-AKT accurately. Time to peak is an essential feature (in addition to signal amplitude) controlling time sensitive downstream catalytic responses of p-AKT, for example cell cycle. In the current context, we believe that the steady state behavior drives the self renewal fate. But in future extensions, for example crosstalk between PI3K/AKT and other pathways, time to peak will become important. Therefore, based on Nim *et al.* future extensions of the pathway in hESCs will need modeling of the shuttling of AKT between the membrane and cytoplasmic compartments, along with membrane fractionation experiments to determine the kinetics of this process (Nim *et al.*, 2015).

In the current analysis, we considered the dynamics until 2 h. After continuous stimulation with insulin, additional degradation processes will take over the dynamics. These mainly include, loss of ligand from the medium via direct degradation as well as indirect degradation within the intracellular medium. The results presented in this chapter assume that insulin is continuously present, which is valid for a short-term analysis. Hence, an implicit assumption made here is that the cell culture is continuously replenished with input insulin. For hESC cultures, a 24 h to 48 h replenishment of the medium is undertaken. Supraphysiologic concentrations of insulin considered in this study do not decrease substantially to limit p-AKT levels in a 24 h process (Sedaghat *et al.*, 2002). However, for a 48 h process, we need to take this effect into consideration. On the other hand, a temporally changing insulin stimulation in combination with perturbation strategies discussed in the text will be useful to find the optimal method of increasing p-AKT levels without wastage of growth factors. This could be conveniently done in controlled ligand delivery environment of a microfluidic platform. While doing this, it will be necessary to ensure the robustness of the response, and some facets of this was explored by us for the same pathway and hESC relevant parameter ranges in a parallel study (Mathew and Banerjee, 2014).

Change in composition of the media due to processes like cellular metabolism will need attention when stimulations are long. As a first step, it will be interesting to study how metabolites and other molecules result in activation/deactivation of the insulin pathway. Based on the mechanism of action of these molecules, the relative importance of these secondary effects could be judged by whether they affect a sensitive or non-sensitive node in the pathway. Information from these studies will enable extension of the insulin model to include crosstalk effects.

Active receptors continuously accumulate in the endosomes and only some fraction of it is recycled back (Nyman *et al.*, 2012). This leads to desensitization of the cells on further insulin stimulation. Hence, additional time of rest may be useful to allow the receptors to recycle back from the refractory state. Modeling the refractory behavior will be necessary in order to ensure that the Sedaghat model is valid for longer times. Similar modeling work was recently attempted for the TGF-β receptor system (Vizán *et al.*, 2013).

For sensitivity analysis, we have focused on a global approach due to the large variability in hESC systems. We can characterize the signaling variability further by performing a detailed parameter estimation (as done in Chapter 4) and identifying true parameter distributions (not just the intervals) after accounting for time to peak. This step will become necessary in future when integrating this pathway with other pathways of differentiation, where the entire dynamics will become important. In this work, we have considered only one branch of negative feedback via PKC-ζ. Other loops of negative feedback exist in the pathway via pGSK3beta, pmTOR, p-p70S6K, pERK all of which converge on IRS1 (Taniguchi *et al.*, 2006). In general, presence of several negative feedbacks makes the system robust to failure in network connections. But if the feedbacks span multiple time scales, it may lead to oscillations in the pathway in combination with strong feedbacks (Birtwistle and Kolch, 2011). Further, molecules like IRS1 have multiple phosphorylation sites that regulate the overall activity and localization of the molecules. Such effects are interesting and have not been characterized for the hESC system yet. These may affect properties like distribution of markers and can be studied by future single cell analysis techniques of relevant signaling molecules. Multi-site phosphorylation of molecules will blow up the number of molecular species to be considered in the mathematical model and in such cases; rules-based frameworks will become extremely useful.

Here we have considered only the role of the PI3K/AKT pathway activated by insulin. Additional growth factors like FGF, EGF etc. may also activate the same pathway with different kinetics and the self-renewal growth medium is a complex cocktail of growth factors and other nutrients. The integrated effect of these external factors on the PI3K/AKT pathway will be necessary to completely model the self-renewal state. Further, the crosstalk between PI3K/AKT and other pathways (endogenously activated) like MAPK/ERK, TGF-β/SMAD, WNT/β-catenin, will be necessary to completely define the signal transduction during self-renewal. In Chapter 4, one such crosstalk between the PI3K/AKT and TGF-β/SMAD pathways is considered.

3.0 IDENTIFICATION OF TRANSCRIPTION FACTORS CO-REGULATED BY COMBINATORIAL SIGNALS INDUCING ENDODERM DIFFERENTIATION OF HESCS

The content of this chapter is taken from Mathew, S., Jaramillo, M., Zhang, X., Zhang, L. A., Soto-Gutiérrez, A., Banerjee, I., 2012. Analysis of alternative signaling pathways of endoderm induction of human embryonic stem cells identifies context specific differences. *BMC Systems Biology* 6, 154

3.1 INTRODUCTION

In the previous aim, our focus was on the PI3K/AKT pathway that is to be maintained at high levels to inhibit differentiation signals. In order to induce differentiation, the levels of other signaling pathways must be enhanced in parallel with inhibition of PI3K/AKT pathway. Multiple signaling pathways have been reported to have success in inducing DE differentiation with subsequent maturation to liver, pancreas and lung. While there is some understanding of the pathways induced by these individual signaling molecules, detailed knowledge of transcriptional controls activated through these signaling pathways is largely unknown. Moreover, combinatorial effect of these endoderm induction pathways, along with its impact on later stage maturation has received less attention. In this aim, we have analyzed the DE induction stage of the differentiation process, by identifying co-regulated TFs across different growth factor combinations using an integrated experimental and mathematical approach. In addition to identifying co-regulated TFs, analyzing all possible combinations of the signaling pathways provides the benefit of thoroughly characterizing their co-operative effects and possibly identifying better combinations not explored before in empirical studies.

3.1.1 Pathways for differentiation of hESCs to DE

Activin A (henceforth denoted as activin) has been shown to be effective in inducing DE from hESCs and is a necessary induction factor (D'Amour *et al.*, 2005; D'Amour *et al.*, 2006a). However, many studies have shown that activin alone may not produce homogeneous differentiation and additional factors must be used to modulate supplementary signaling pathways along with the TGF-β/SMAD pathway activated by activin (Payne *et al.*, 2011; Zhang *et al.*, 2009a). We chose several widely used DE induction protocols all of which involve activin with either PI3K inhibition (Zhang *et al.*, 2009c), WNT3A (D'Amour *et al.*, 2005), BMP4 (Phillips *et al.*, 2007) or FGF2 (Basma *et al.*, 2009). The hESCs were differentiated into DE using these molecules alone and in all possible combinations, at the end of which the differentiated cell population was analyzed for endoderm markers. This gives rise to 15 experimental conditions and for each condition, 12 TFs were analyzed giving rise to a 15 x 12 expression matrix with three replicates. Further details of the experimental techniques are given in Section 3.2.

3.1.2 Introduction to the mathematical methods

Our aim is twofold: to identify which growth factor combinations are most effective for efficient DE induction; and to identify TF subsets co-regulated by these induction conditions. We first analyzed the mean expression data using Hierarchical clustering (HC) to identify relationships between the conditions and the TFs, followed by biclustering on the original expression data with replicates to identify the TFs which are co-regulated under subsets of these conditions.

3.1.2.1 Hierarchical clustering

HC is a useful technique to analyze and interpret multivariate data. Each data point here is represented as a vector in the high dimensional space and the distances between these data points are measured using a suitable distance measure (Friedman *et al.*, 2001). The high dimensional space of the dataset is described by pairs of points, one from the condition space (15 dimensional) and one from the TF space (12 dimensional). These are the two major dimensions of the dataset. The HC process links the points in each major dimension together and the result is a hierarchical grouping of the data points separately in each of the dimensions (TFs and conditions in our case). Using HC, we can capture the similarities between different growth factor treatments for DE induction using co-regulated TFs (all of them). HC has been successfully used in a number of bioinformatics applications including microarray data analysis, structure identification of bio-molecules and gene pathway identification (Slonim, 2002).

3.1.2.2 Biclustering

The major disadvantage of HC is that the clustering is performed on each major dimension (TF and condition) separately. In other words, when clustering is performed between the TFs, the information on the condition is homogenized and vice versa. However, same TF may have different functions for different conditions and thus may be regulated differently. Therefore, ideally we seek ‘subsets of TFs’ that are co-regulated under ‘subsets of conditions’. This local information is preserved in a biclustering approach where each data-point is truly treated as a pair in the two major dimensions. A bicluster is defined as a subset of TFs that show coherence in the expression across subset of conditions. Figure 3.1A shows example of a coherent pattern. It is important to mention that this information may be partially obtained from a clustergram of a HC *post-priori* by looking at the clustering information in the 2D heatmap. However, this has limited use, since features like overlapping and non-trivial clustering are difficult to analyze this way.

In 2000, Cheng and Church proposed the use of similarity measure called the mean square residue for identification of coherent biclusters (Cheng and Church, 2000). Since then newer and better algorithms have been developed to identify biclusters with particular characteristic trends like coherence, low overlaps and hierarchical structure (Pontes *et al.*, 2015; Yang *et al.*, 2003). These algorithms perform either one or a combination of iterative row and column clustering, greedy iterative search or exhaustive bicluster enumeration (Madeira and Oliveira, 2004). Bleuler *et al.* proposed an evolutionary algorithm (EA) to determine high quality, partially overlapped biclusters using the Cheng and Church formulation (Bleuler *et al.*, 2004). EAs have the advantage of large search space and are efficient methods for complex optimization problems (Divina and Aguilar-Ruiz, 2006). High quality biclusters should satisfy many criteria; namely they should contain as many genes and conditions as possible, low mean square residue, high row variance and should have low overlapping. Divina *et al.* formulated Sequential Evolutionary Biclustering (SEBI) algorithm to identify such biclusters from the expression data, which has been adopted in the current work to identify important biclusters for the endoderm induction data under different combinations of the growth factors (Divina and Aguilar-Ruiz, 2006). SEBI can find high quality biclusters and has been proved to perform well for large-scale biological datasets. At the same time, it allows the user the flexibility of selecting the degree of overlap of the biclusters.

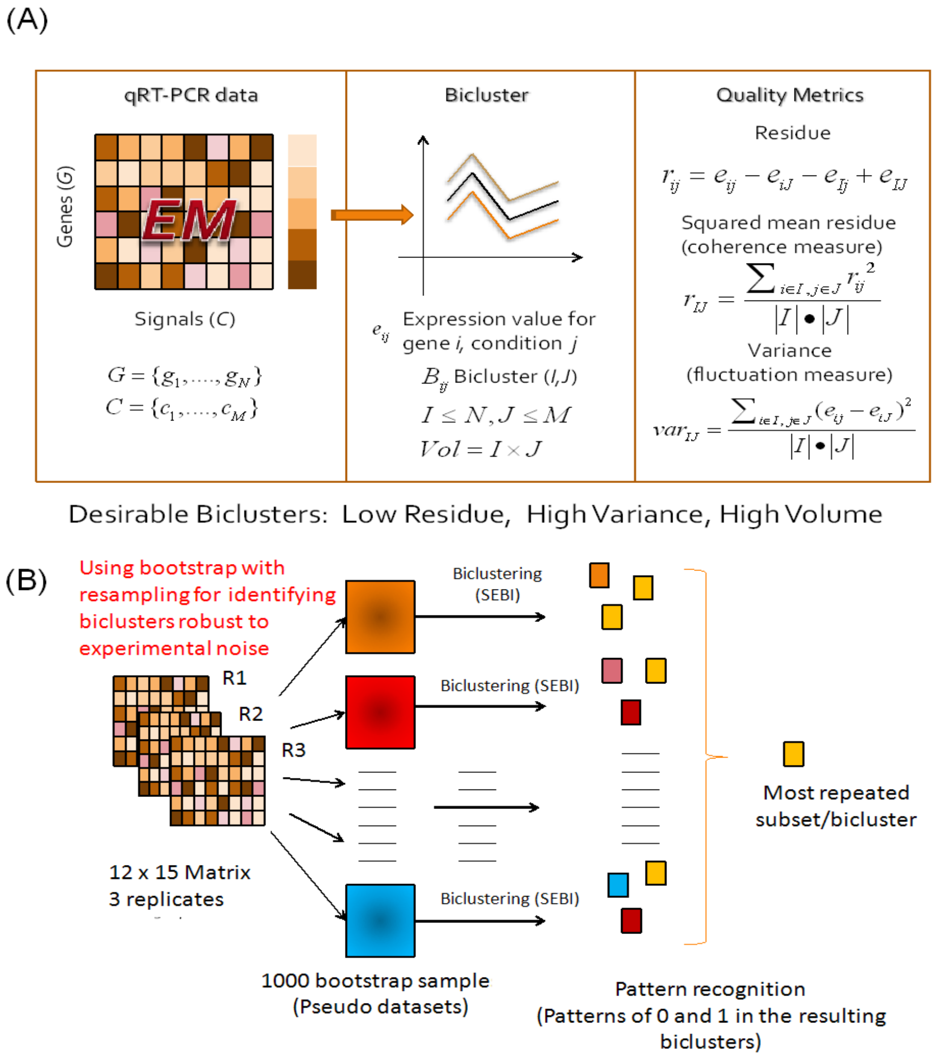

(A)
qRT-PCR data
Genes (G)
EM
Signals (C)
$G = \{g_1, \ldots, g_N\}$
$C = \{c_1, \ldots, c_M\}$
Bicluster
e_{ij} Expression value for gene i, condition j
B_{ij} Bicluster (I,J)
$I \leq N, J \leq M$
$Vol = I \times J$
Quality Metrics
Residue
$r_{ij} = e_{ij} - e_{iJ} - e_{Ij} + e_{IJ}$
Squared mean residue (coherence measure)
$r_{IJ} = \frac{\sum_{i \in I, j \in J} r_{ij}^2}{|I| \bullet |J|}$
Variance (fluctuation measure)
$var_{IJ} = \frac{\sum_{i \in I, j \in J} (e_{ij} - e_{iJ})^2}{|I| \bullet |J|}$
Desirable Biclusters: Low Residue, High Variance, High Volume
(B)
Using bootstrap with resampling for identifying biclusters robust to experimental noise
R1
R2
R3
12 x 15 Matrix
3 replicates
Biclustering (SEBI)
Biclustering (SEBI)
Biclustering (SEBI)
1000 bootstrap sample (Pseudo datasets)
Most repeated subset/bicluster
Pattern recognition (Patterns of 0 and 1 in the resulting biclusters)

Figure 3.1 Biclustering with bootstrap analysis.

(A) Typical bicluster of interest. Expression matrix from qRT-PCR analysis contains many hidden patterns. One such pattern is shown in the middle section, which shows three genes a, b, and c varying coherently when the experimental condition is varied. Several metrics are used to describe coherence, most commonly low residue.

Additionally, high variance and high volume constraints are added to identify non-trivial biclusters. (B) Work-flow for the entire analysis from data collection to identification of robust biclusters. In short, we start with the qRT-PCR data and perform bootstrap with re-sampling from the experimental replicates to obtain 1000 pseudo-datasets. Each of these datasets is subjected to biclustering analysis to obtain the most coherent pattern in each dataset. The resulting biclusters are then analyzed for the most repeated subsets of biclusters.

3.1.2.3 Handling data variability using bootstrapping

The gene expression data obtained for cell culture systems are subjected to noise because of the heterogeneity and stochasticity associated with the system. Differences among the biological replicates may therefore arise due to the inherent heterogeneity of the ES cell population as well as by experimental noise. Therefore, it is essential that the biclustering algorithm be supplemented with additional methods to discover good quality and robust biclusters from noisy gene expression data. One way to do this is to obtain a large number of experimental replicates and perform biclustering over the entire dataset. This is however, expensive and impractical. A mathematical surrogate of this approach is bootstrapping, a concept first presented systematically by Efron *et al.* (Efron and Tibshirani, 1994).

Essentially, bootstrapping generates a pseudo dataset from the small number of experimental replicates by a sampling with replacement technique. The most important assumption in a bootstrap approach is that it relies on empirical distribution presented by the dataset. The advantage of bootstrap lies in estimating statistically significant parameters from a limited number of experimental replicates (Politis and Romano, 1994). Thus, the results from a bootstrap analysis can provide information on the parameter variances and confidence intervals. These bootstrap datasets may be further analyzed by ensemble methods like bagging to identify aggregation of biclusters, referred to as meta-clusters (Hanczar and Nadif, 2011). We have adopted a similar approach to aggregate the individual biclusters identified from the bootstrap datasets. However instead of identifying an ensemble of biclusters, we have concentrated on identifying the most repeated subset of the bicluster, which we denote as robust. This approach is summarized in Figure 3.1B.

3.2 METHODS

3.2.1 Cell culture and treatment

H1 hESCs were placed on hESC certified matrigel coated wells and maintained with mTeSR1 with media change every day. Cells were passaged every 5 to 7 days by incubating in 1 mg/ml dispase for 5 minutes followed by mechanically breaking the colonies and splitting at a 1:3–1:5 dilution. Cells were examined under the microscope every day and colonies with observable differentiation were picked and removed before the media changes. H1 hESCs were allowed to grow to 60-70% confluency before the experiments were started. Once confluency was reached, differentiation was performed by adding DE induction media for 4 days with media change every day. All conditions were prepared in DMEM:F12 supplemented with B27 and 0.2% BSA with 100 ng/ml Activin A. Conditions involved the use of individual and all possible combinations of growth factors and molecules at the following concentrations: basic FGF (F) at 100 ng/ml, BMP4 (B) at 100 ng/ml, WNT3A (W) at 25 ng/ml and Wortmannin (PI3K inhibitor, P) at 1 μM. This leads to 15 different experimental conditions.

3.2.2 Measurement of TF expression

After 4 days of DE induction, cells were lysed and RNA extracted using Nucleospin RNA II kit (Macherey Nagel) according to the manufacturer's instructions. The sample absorbance at 280 nm and 260 nm was measured using a BioRad Smart Spec spectrophotometer to obtain RNA concentration and quality. Reverse transcription was performed using ImProm II Promega reverse transcription kit following the manufacturer's recommendation. qRT-PCR analysis was performed for endoderm and pancreatic markers using the primers listed in Appendix B. A total of 12 transcription factors were studied which included pluripotency marker *OCT4*, mesendoderm marker *BRACHYURY*, DE markers namely, *CXCR4*, *SOX17*, *CER*, *FOXA2* and pancreatic progenitor markers *PTF1α*, *PDX1*, *GATA4*, *HNF1β*, *HNF4α* and *HNF6*. *GAPDH* was selected as the housekeeping gene. Briefly, the fold change was calculated from the cycle times, C_T, after normalization with respect to the control sample and housekeeping gene, *GAPDH* as $2^{-\Delta\Delta C_T}$,

where $\Delta\Delta C_T = \left[\left(C_{T,target} - C_{T,GAPDH}\right)_{sample} - \left(C_{T,target} - C_{T,GAPDH}\right)_{day0}\right]$. The control sample was chosen to be undifferentiated cells at day 0. The TF expression profiles can be grouped together to form an expression matrix with the rows corresponding to the measurements of interest (like the relative mRNA concentrations) and the columns corresponding to the experimental conditions or samples. Thus, each element in the matrix refers to the intensity of the particular measurement in a given sample. A schematic of the experimental data collection is presented in Figure 3.2A.

3.2.3 Hierarchical clustering

Hierarchical clustering partitions the data into clusters through an iterative process, where similarity or dissimilarity between every pair of variables in the data matrix is calculated using an appropriate distance measure followed by grouping the variables in close proximity using a linkage function. We used in-built MATLAB functions to perform the analysis using various distance measures e.g. Euclidean, correlation distance, city block etc., on the mean centered and variance scaled expression matrix. The results were represented as a clustergram i.e. the linkage tree and the corresponding heat map. We tested the tree generated using different linkage measures after normalization of the mean expression matrix and found all the trees to be very similar with the cophenetic correlation coefficient greater than 0.9.

3.2.4 Biclustering using SEBI

Biclustering can be described as two-dimensional clustering, where a subset of genes exhibiting similar trend across a subset of conditions is being identified. Such subsets can be considered to be participating in similar regulatory mechanism, hence constituting a regulatory network. In order to identify sets of TFs expressing coherent trends under specific sets of conditions, we analyzed our TF-condition matrix, X, using the SEBI algorithm developed by Divina *et al.* (Divina and Aguilar-Ruiz, 2006).

3.2.4.1 Biclustering formulation

The SEBI algorithm identifies coherent biclusters sequentially with the help of a number of metrics as described below. For a bicluster $B(I,J) \in X$, containing elements, e_{ij} for $i \in I$, $j \in J$ the residue, r_{ij} of each element in the bicluster is defined as: $r_{ij} = e_{ij} - e_{iJ} - e_{Ij} - e_{IJ}$. The gene base is defined as $e_{iJ} = \frac{\sum_{j \in J} e_{ij}}{|J|}$, with I and J representing the total number of genes and conditions respectively in the bicluster B. The condition base is defined as $e_{Ij} = \frac{\sum_{i \in I} e_{ij}}{|I|}$. The base of the bicluster is the mean of all entries in the bicluster, i.e., $e_{IJ} = \frac{\sum_{i \in I, j \in J} e_{ij}}{|I| \times |J|}$. The residue, therefore, indicates the degree of coherence of the element with other elements in the bicluster. Further, the squared mean residue of all the elements in the bicluster is defined as $r_{IJ} = \frac{\sum_{i \in I, j \in J} r^2{}_{ij}}{|I| \times |J|}$. It is possible to have biclusters having constant expression values and hence have low residue value. To avoid such trivial biclusters, the variance metric is introduced. The variance, var_{IJ}, of a bicluster is defined as, $var_{IJ} = \frac{\sum_{i \in I, j \in J} \left(e_{ij} - e_{iJ}\right)^2{}_{ij}}{|I| \times |J|}$. Hence, the variance captures fluctuating trends. Finally, we would be interested in biclusters with as many genes and conditions as possible i.e. having large volume. The basic premise of the analysis is that the genes belonging to a bicluster are under the influence of a common regulatory pathway and hence show coherence in their expression trends. However it is possible for the genes to participate in multiple regulatory pathways, to capture which we allow certain degree of overlapping amongst the biclusters discovered sequentially by the SEBI algorithm using a penalty term. Thus, our final goal is to find biclusters of maximum size, with mean squared residue lower than a given threshold (δ), with relatively high row variance, and a low level of overlapping among the biclusters. We represent this as an optimization problem with objective function defined as:

$$\min_{(i,j)\ of\ X} F(B) = \frac{m\_residue(B)}{\delta} + \frac{1}{row\_\mathrm{var}iance(B)} + w_d + penalty$$

In this function, $B(I,J)$ is an individual solution, δ is the mean squared residue of the bicluster B, $row\_\mathrm{var}iance$ is the row variance of B, $penalty = \sum_{i\in I, j\in J} w_p\left(e_{ij}\right)$, where w_p is defined as:

$$w_p\left(e_{ij}\right) = \begin{cases} 0 & if \left|Cov(e_{ij})\right| = 0 \\ \dfrac{\sum_{n\in N, m\in M}\left|Cov(e_{nm})\right|}{e^{-\left|Cov(e_{ij})\right|}} & if \left|Cov(e_{ij})\right| > 0 \end{cases}$$

Where N, M are the number of rows and columns of the expression matrix, respectively and $\left|Cov(e_{ij})\right|$ is the number of previous biclusters containing e_{ij}. The use of the penalty term biases the search against members which already have appeared in the previous biclusters, thus reducing the overlapping amongst the biclusters.

w_d is defined as $\left(w_r \bullet \dfrac{\delta}{row_B} + w_c \bullet \dfrac{\delta}{column_B} \right)$ and δ is the threshold mean squared residue and biclusters with mean squared residue above δ are discarded.

3.2.4.2 Solution procedure

The current optimization formulation has been identified to be NP-hard and has been shown to be effectively handled by evolutionary techniques like Genetic Algorithm (GA) (Divina and Aguilar-Ruiz, 2006). GA is an iterative search process which looks for the fittest member of a population (candidate solutions) using the biological principle of evolution under mutation and natural selection (Golberg, 1989; Tanay *et al.*, 2005). In a typical GA, the optimization variables are encoded as a sequence of binary bits and these sequences are concatenated to form the chromosome. Thus, for the present formulation, each chromosome consists of I binary bits for genes and J binary bits for conditions forming the $I + J$ binary bits of the chromosome. The binary variables, 0 and 1 represent the absence or presence of a gene (or condition) respectively. Thus, a GA population is made of chromosomes with each chromosome representing a candidate bicluster.

Each chromosome has a metric associated with it called the fitness which we wish to maximize. The GA algorithm is initiated by randomly initializing a population of chromosomes (i.e. biclusters). The population is continuously evolved in every generation by the operators: reproduction, crossover and mutation. At the end of every generation, individuals for the next one are selected on the basis of their fitness values. This cycle of evolution is continued until a predetermined termination criterion is reached. For the present case, we continued the simulations for a maximum number of generations until no further change in the population was observed. The biclustering formulation was coded in FORTRAN R90 and the Genetic Algorithm (version 1.7a) driver obtained from David Carroll, CU Aerospace, Urbana, IL. Computations were performed on INTEL (R) Core (TM) 2 Quad CPU (Q8400 @ 2.66 GHz).

3.2.4.3 Identifying robust biclusters

The inherent noise in biological systems makes it difficult to draw meaningful conclusions from a deterministic analysis. The formulation proposed above is based on the mean gene expression data which possibly reduces confidence in the identified bicluster. Here we have adopted the bootstrap technique to obtain robust biclusters from noisy experimental data. Bootstrap is a statistical technique to generate large data set from a small number of experimental replicates, using sampling with replacement technique. The present formulation systematically re-samples the original experimental data set using Monte Carlo algorithm to generate the artificial data set. The optimization formulation of the biclustering problem is then solved at each of the bootstrap data points to generate a family of alternate biclusters. The final goal will be to identify the most repeated biclusters in the entire array, based on the justification that such a bicluster will be relatively insensitive to experimental noise and hence is robust. To this end, the number of repeats of a particular gene-condition combination is analyzed using the quicksort algorithm (N log N). Our analysis showed that the complete bicluster was typically not repeated significantly; instead only subsets of the biclusters were repeated sufficient number of times. For identification of robust biclusters, we set the threshold frequency of repeats as 500 out of every 1000 alternate biclusters. The most repeated subsets are thereby concluded to be robust under experimental noise. The workflow for the entire analysis is depicted in Figure 3.1B.

3.3 RESULTS

The focus of this work is to understand the mechanism of endoderm induction using different growth factors, acting alone and in combination, from an integrated experimental and computational approach. The H1 human embryonic stem cells were induced towards endoderm lineage using activin along with alternate growth factors, namely FGF2, BMP4, PI3KI, WNT3A, added in 15 combinations. The cells differentiated thereof were analyzed in detail for their gene expression levels, specifically concentrating on a broad range of endoderm markers along with representative pancreatic endoderm markers.

3.3.1 Experimental analysis of endoderm differentiation using combinations of major pathways

Figure 3.2A shows the mean expression data plotted as fold changes in 12 genes across the 15 experimental conditions. At this stage, the fold change data showed interesting trends for the different conditions. When using only one factor other than activin, PI3KI along with activin was found to give the highest expression of most of the DE markers while BMP4 and activin in combination was found to give the lowest expression among the four conditions. Interestingly, BMP4 was found to perform better in combination with another factor like WNT3A or FGF2. Also, FGF2 containing conditions were found to favor *CER* while BMP4 containing conditions to favor *HNF4α*. Among the 4 conditions which contain 3 factors other than activin, combinations of FGF2, BMP4 and PI3KI perform well. Using all the factors together was not particularly useful since all the TFs maintained expressions in the same range as other combinations. Figure 3.2B shows the range of variation observed in each of the transcriptional markers across the 15 experimental conditions along-with the experimental replicates. The levels of DE markers *CER*, *FOXA2*, *CXCR4* and late endoderm markers *HNF4α*, *HNF1β* and *GATA4* change substantially when the induction conditions are changed. This level of analysis, however, makes it difficult to draw mechanistic insights from the dataset. Hence, we performed a more rigorous mathematical analysis to separate out the TF trends and associate them with the appropriate conditions. Because of the inherent differences in expression level of different genes, it is essential to normalize the data to avoid bias. For the mathematical analysis, the data presented in Figure 3.2A was normalized by mean centering and variance scaling so that every TF has a mean expression value of zero and standard deviation of one.

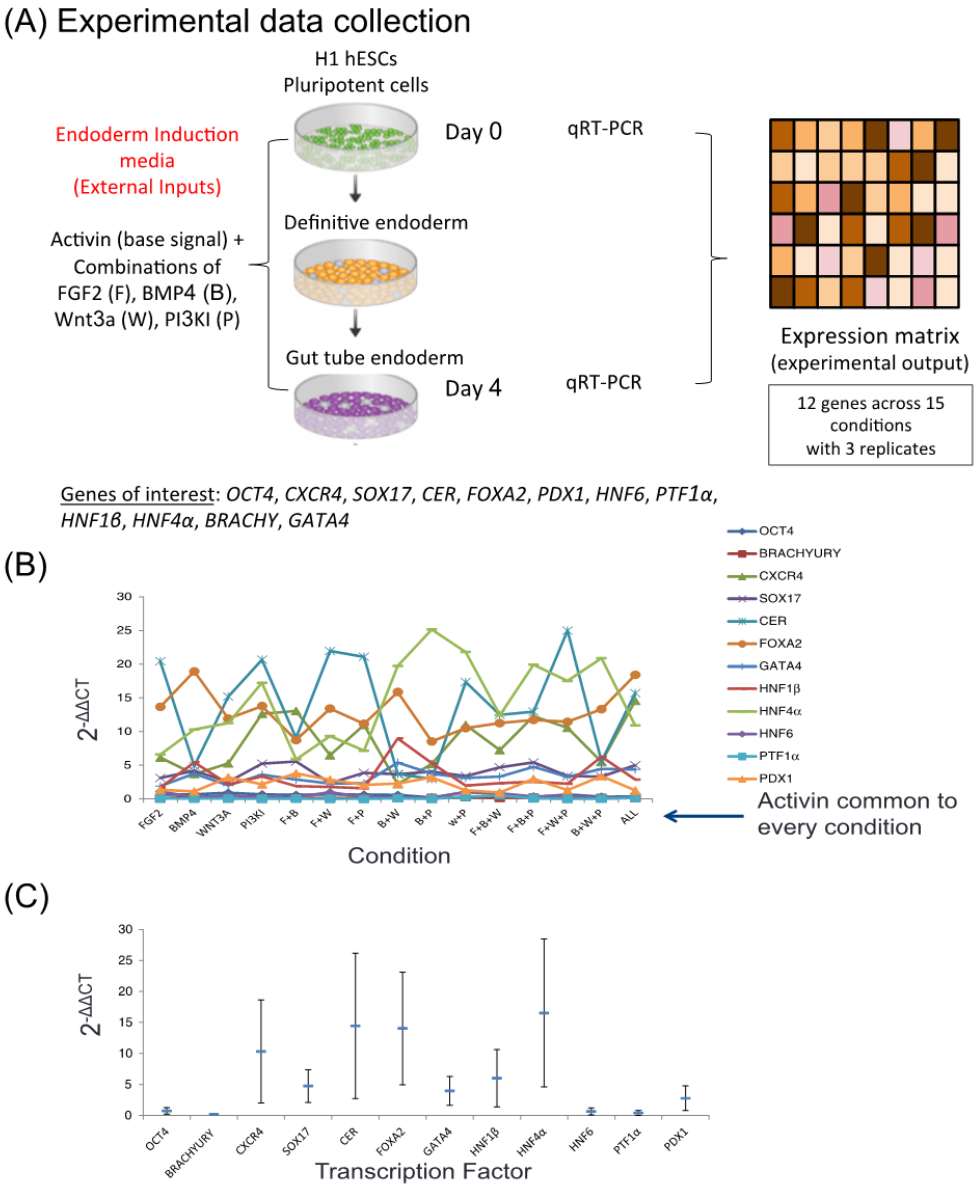


Figure 3.2 Experimental data used for biclustering + bootstrap analysis

(A) Experimental variables and outputs. (B) The fold change calculated from the mean expression data from qRT-PCR on day 4 of the differentiation process is plotted from the expression matrix, *X*, constructed using rows as the TFs and columns as the experimental conditions. (C) Variation observed in the 12 transcriptional markers with changes in the signaling pathways presented as mean ± SE. All the major DE markers *CER*, *CXCR4*, *FOXA2*, *SOX17* and the later endoderm markers *HNF4α*, *HNF1β* and *GATA4* show significant changes with the nature of DE induction.

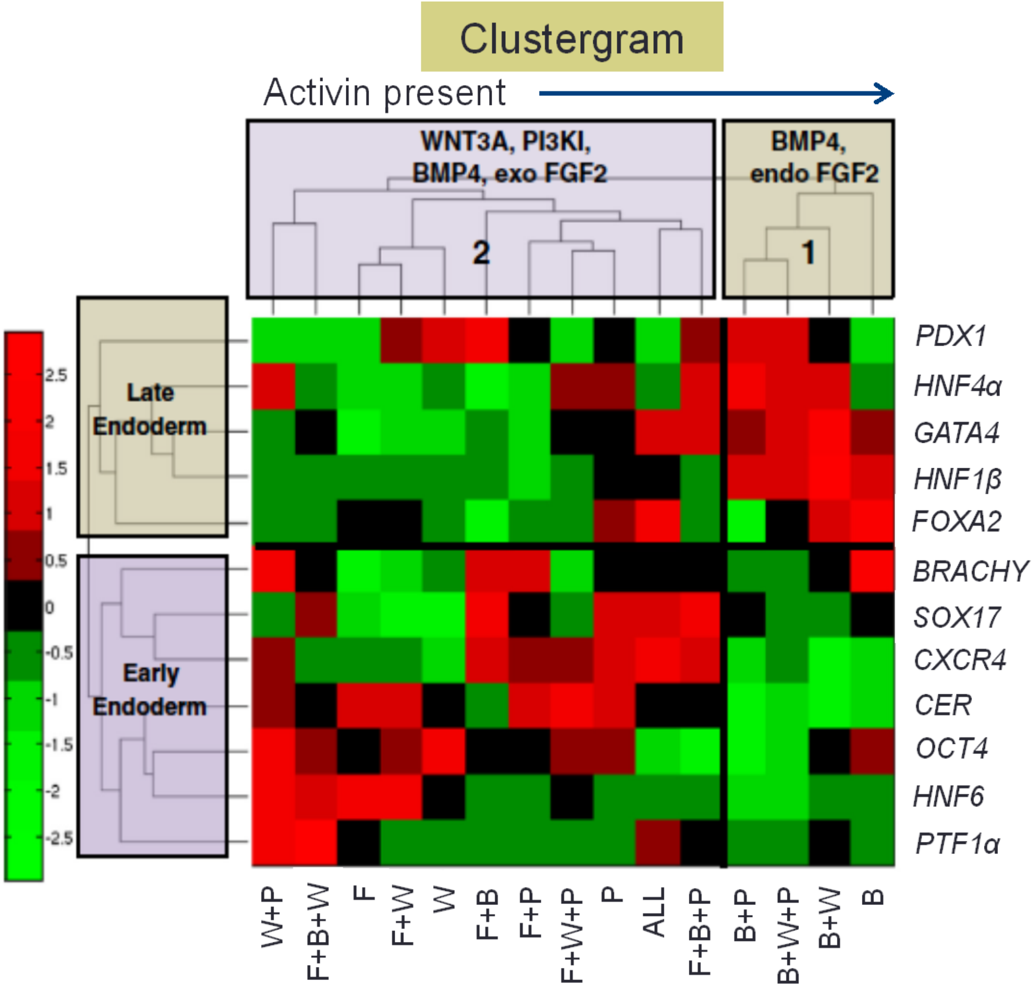


Figure 3.3 Hierarchical clustering on mean expression data

The conditions cluster into two major groups, one containing BMP4 in the absence of exogenous FGF2 and the other containing all the other treatments and BMP4 in combination with exogenous FGF2. Activin A is common among all the treatments. The TFs cluster into two groups, the late and early endoderm markers.

3.3.2 Hierarchical clustering of the mean expression data

The mean experimental data matrix was first analyzed using hierarchical clustering which clusters the TFs and conditions separately, as shown in Figure 3.3. Among the conditions, two major branches were observed: the first cluster contains BMP4 dominant conditions (B, B + W, B + P, B + W + P) and the second cluster contains the remaining conditions which also includes BMP4 but interestingly only in combination with FGF2. The TFs also segregate into two branches; the first branch contains the late endoderm markers and one of the DE markers (*HNF4α, HNF1β, GATA4, PDX1, FOXA2*), the second branch contains the early DE and late endoderm markers (*OCT4, BRACHYURY, CER, HNF6, CXCR4, SOX17, PTF1α*). The first group of markers is particularly high in BMP4 dominant conditions and low in the other conditions. The second group of markers is low in the BMP4 dominant conditions and high in the presence of PI3KI, WNT3A and BMP4 and high FGF2. Thus our results point to differences in activin and BMP4 induced endoderm in the presence and absence of exogenous FGF2.

The clusters identified by the hierarchical algorithm reflect our biological understanding of the induction conditions as seen from the previous studies. A major difference between the two clusters of conditions was the context dependent function of BMP4. In the presence of FGF2 and high activin, BMP4 was found to favor the endodermal lineage which was seen in several recent studies (Bernardo *et al.*, 2011; Xu *et al.*, 2011) and was also on par with PI3KI dominant conditions which gave the best endoderm in our experiments. Also, in our BMP4 dominant conditions, the late stage markers showed very high expression while the major DE markers were low indicating that the resulting endoderm may already be mature. Among the second group of conditions, PI3KI and high activin resulted in high expression of three major DE markers *SOX17, CXCR4 and CER* which is supported by a number of earlier studies (McLean *et* *al.*, 2007a; Singh *et al.*, 2012b). Using all the factors together does not improve upon the endoderm derived by PI3KI treatment. The second group of conditions also contains FGF2 as a major factor along with WNT3A. It is found that both pluripotency (*OCT4*) and the endoderm factors (*CER* and *HNF6*) are relatively favored by conditions involving FGF2 and WNT3A as the major contributor. In fact, FGF2 has been found to be sufficient to maintain the hESCs in the pluripotent state and has also been used for endoderm induction in several differentiation protocols (Shiraki *et al.*, 2008). Thus, FGF2 can potentially favor both pluripotency as well as endoderm differentiation depending on associated conditions (for example level of activin).

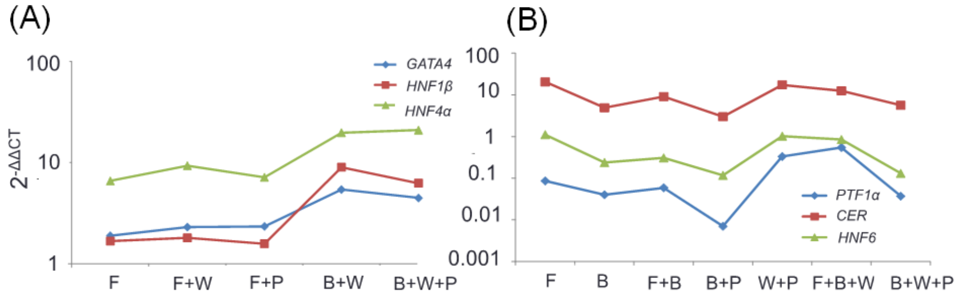


Figure 3.4 Biclusters obtained from normalized mean expression data

(A) Optimal Biclusters. The bicluster contains 3 genes across 5 conditions. (B) Subsequent optimal bicluster containing 3 genes and 7 conditions. The bicluster parameters selected were δ = 1.5, Wc, Wr = 1.

3.3.3 Identification of co-regulated transcription factors by biclustering

While hierarchical clustering enables a fast and simplistic analysis of the experimental data sets, it does not provide information on which subsets of TFs are co- regulated across subsets of conditions. Identifying such co-clusters will be beneficial, since the governing signaling pathways change with the induction condition and the same TFs may not be co-regulated. The technique of biclustering serves to mine subgroups of such TFs exhibiting similar trends in their expression level under sub- sets of conditions. Hence TFs appearing in the same bicluster can be inferred to be co-regulated and constituents of a similar network architecture. The experimental data matrix, *I*, constituting the mean expression data across all the growth factor conditions is analyzed using the algorithm elaborated in Methods section. Here, the biclustering approach is formulated as an optimization problem solved using genetic algorithm (GA) and the quality of every candidate bicluster is assessed by a fitness function. The fitness function has a number of free parameters associated with it that can be tuned in order to identify certain desired trends. The detailed procedure on the selection of the optimum parameters is outlined in the Appendix B.

The developed optimization based bicluster identification algorithm was applied to the mean expression data with the above mentioned parameters, which resulted in a 3-gene 5-condition bicluster as illustrated in Figure 3.4A. However, to identify additional biclusters, possibly with overlaps, the SEBI algorithm was subsequently run by penalizing the identified biclusters. One such bicluster is presented in Figure 3.4B. Although, the SEBI algorithm allows some degree of overlapping amongst the subsequent biclusters, the current mean dataset did not result in any overlaps.

Recently, a new method was proposed by Banka *et al.* called as Fuzzy Possibilistic Biclustering which assigns a membership value to each gene-condition pair in the expression matrix and therefore, allows varying degree of overlapping amongst the biclusters (Filippone *et al.*, 2006; Mitra *et al.*, 2007). However, though the method has been proven to provide very large biclusters with acceptable residue, the selection of the degree of fuzziness often depends upon the question that the biologists have set to answer (Nosova *et al.*, 2011). In our case, we are interested in analyzing the well identified markers of endoderm induction under necessary signaling pathways. Since, our aim is to discover subtle differences in the gene regulation when the induction conditions are changed, a traditional crisp method like SEBI will be more useful for identifying the best induction condition.

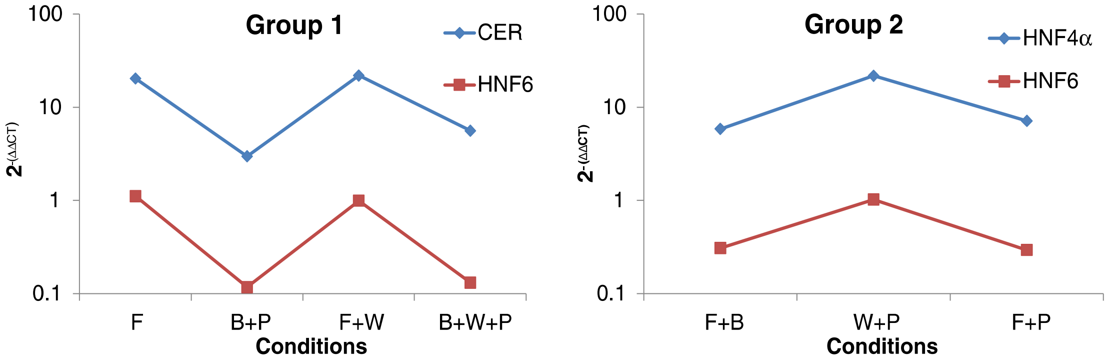


Figure 3.5 Robust biclusters identified from 1000 bootstrap datasets

Robust biclusters are the most repeated subsets (>500). The bicluster parameters selected were δ = 1.5, Wc, Wr = 1. Note: Group 1 contains five subsets only one of which is shown.

3.3.4 Robust biclusters identify WNT3A treatment to favor both early and late endoderm

The above identified biclusters were for the mean dataset, and hence does not explicitly take into account the experimental variations. In general biological datasets are known for their noise and uncertainty, and in particular stem cells have inherent heterogeneity and stochasticity. In order to increase confidence in the identified bicluster we undertook bootstrap analysis on the experimental data to generate 1000 pseudo-datasets. Each of these datasets were treated as an experimental repeat and subjected to the entire biclustering analysis. In order to identify somewhat overlapped biclusters, we ran the biclustering algorithm five times at each data point by subsequently penalizing previously identified biclusters.

The next task was to determine a robust bicluster from this array of alternate biclusters. We hypothesize that the robust bicluster will not be significantly affected by the experimental noise, and hence will appear a large number of times in the bootstrapped-bicluster data set. However, a thorough search of the entire array of alternate biclusters for frequency of repeats did not yield any satisfactory outcome. Thus we could not find a single bicluster that was significantly repeated in its entirety across the data set. Instead, we realized subsets of genes and conditions of the bicluster were being repeated with very high frequency instead of the entire bicluster. Hence, we focused on identifying such subsets from the family of bootstrap + bicluster solutions. Setting a minimum threshold of 50% repeats across the bootstrap samples, we identified 6 such subsets. First five of these contained different combinations of the same two markers and four conditions. Hence we collected them together into a single group. The profiles of the repeated subsets are presented in Figure 3.5. These subsets are of two kinds: Group 1 contains (*CER*, *HNF6* | F, F+W, B+W+P, B+P) and Group 2 contains (*HNF6, HNF4α*| F+B, F+P, W+P). It is important to note that the robust biclusters were different from the biclusters obtained for the mean expression data. For example, the biclusters in Figure 3.4 show that *HNF4α* clusters closer to *HNF1β* (and *GATA4*) rather than *CER*. This is also evident from our hierarchical clusters in Figure 3.3. The fact that they do not appear together in the robust biclusters is interesting and shows that analysis from mean datasets can be risky for stem cell systems when there is inherent variability among the replicates. Supportively, the *HNF4α*, *HNF1β* (and GATA4) combination occurs in subsets with less than 300 repeats (data not shown). Figure 3.6 shows a summary of the robust biclusters represented as a bipartite graph of genes and conditions. The identified biclusters are biologically relevant to the development stages *in vivo*. Group 1 contains endoderm markers *CER* and *HNF6* under FGF2/WNT3A and BMP4/WNT3A/PI3KI. *CER* is an important early marker for the DE stage rising after the formation of the primitive streak during development while *HNF6* is a marker for a more primitive foregut stage in pancreas development (D'Amour *et al.*, 2005). Thus, Group 1 is similar to the foregut development stage in vivo (Zorn and Wells, 2009). In addition, the conditions in Group 1 contain FGF2 and WNT3A but not BMP4 and as seen from Figure 3.5, *CER* and *HNF6* decrease under BMP4 dominance. Thus, the biclustering analysis shows that the early marker *CER* and a late endoderm marker *HNF6* are controlled by the FGF2, WNT3A pathway and are relatively down-regulated under BMP4 and PI3KI. Group 2 contains another primitive foregut stage marker *HNF4α* alongwith *HNF6*. Interestingly here, the biclustering results show that pancreatic endodermal transcriptional machinery may not be favored at the DE stage by the FGF2 + BMP4 combination although in our hierarchical clustering results FGF2 + BMP4 combination clustered with the other conditions that gave a better DE signature. We also note that WNT3A and PI3KI combination with high activin increased the expression of *HNF4α* and *HNF6* and these conditions also gave a successful DE signature as seen from the hierarchical clustering. Thus our results indicate that WNT3A pathway can favor both early and late markers like *CER*, *HNF4α* and *HNF6*. Also, WNT3A + PI3KI induced DE cells may be more capable of developing into later pancreatic lineages. While WNT3A and PI3KI have been used for DE induction towards pancreatic maturation (D'Amour *et al.*, 2006a; Zhang *et al.*, 2009c), the effect of co-induction has not been explored yet. However, direct modulation of molecules from these two pathways was undertaken by Singh *et al.* and found to lead to better endoderm differentiation and pancreatic maturation as compared to one of these alone (Singh *et al.*, 2012b).

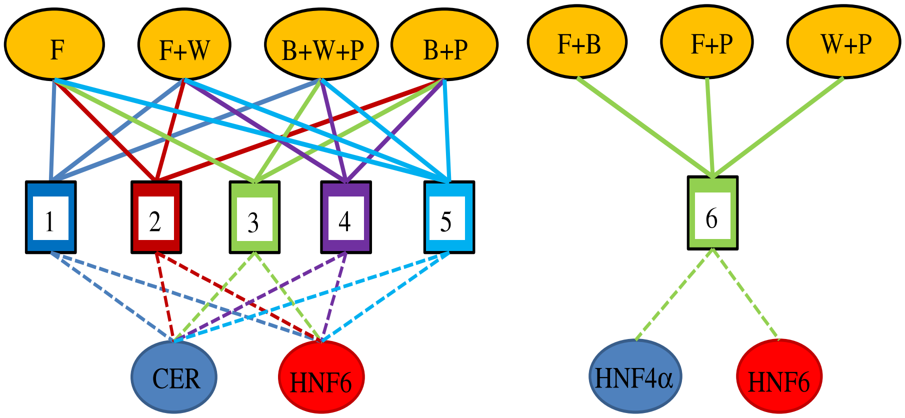


Figure 3.6 Robust subsets of co-regulated TFs presented as a bipartite graph.

We have identified high Activin along with PI3K inhibition or activin in combination with WNT3A to work the best to co-regulate early endoderm marker *CER* and late endoderm markers *HNF6*. The Group 2 TFs *HNF4α* and *HNF6* are part of the network inducing NGN3 and PDX1, reminiscent of the pancreatic genotype and are favored by high activin with PI3KI and WNT3A.

3.4 DISCUSSION

The differentiation of hESCs into the endoderm lineages is carried out by the activation of different signaling pathways mimicking *in vivo* development. However, there is no consensus on which induction method is the most desirable and whether combination of these could result in an endoderm with the best signature. Here, we have used a combination of experimental and mathematical techniques to shed light on these concerns.

3.4.1 The DE signature differs under exogenous activation of different signaling pathways participating in endoderm commitment

Our experiments with different DE inducing conditions show that the DE potential of the differentiating hESCs is highly dependent on the method of DE induction. The major DE markers (*CER*, *CXCR4*, *FOXA2*, *SOX17*) showed considerable variation when some of the pathways were activated above their basal levels.

All the pathways studied here have been known to be important at the earlier stages of *in vivo* endoderm differentiation and has also been documented as necessary for *in vitro* differentiation. The common denominator in our studies is activin which is an essential inducer of DE. This is primarily because activin, being a member of the TGFβ family, mimics nodal signaling which is proven to be necessary for endoderm development (Payne *et al.*, 2011). Activin has been shown to maintain pluripotency at low concentrations and to induce mesoderm and endoderm at high concentrations (Singh *et al.*, 2012b). However, activin alone may not result in efficient endoderm induction (Zhang *et al.*, 2009a). Low PI3K signaling was essential for efficient induction of DE from hESCs (McLean *et al.*, 2007a). Our hierarchical clusters show that Activin and PI3K inhibition in combination favor the up-regulation of a number of DE markers and form the most minimal signaling pathways to be modulated for efficient DE induction. In fact a number of recent studies have identified the interplay between PI3K/AKT and Activin/SMAD2,3 pathways and the resulting regulation of the gene transcription events necessary for early DE induction (Singh *et al.*, 2012b).

Among the DE markers, *CER* showed up-regulation on differentiation, and the highest up-regulation was achieved in the presence of FGF2, WNT and PI3KI treatments. Katoh *et al.* recently identified the binding domains of several key signaling effectors of the activin and WNT pathways on the promoter regions of *CER* in hESCs (Katoh and Katoh, 2006). According to their results, the key nodal effectors SMAD3/SMAD4 as well as the WNT effectors beta-catenin and TCF/LEF transcriptional complex regulate the expression of the *CER* gene. In addition to high activin and WNT signaling, PI3K inhibition may be necessary to enhance the effect of nodal signaling as SMAD3/SMAD4 complex is negatively regulated by AKT (Singh *et al.*, 2012b). Exogenous FGF2 simultaneously activates the ERK pathway and maintains the expression of other key regulators of differentiation (Mfopou *et al.*, 2010). However, BMP4 effectors SMAD1/3 may compete with the activin pathway and thus reduce the up-regulation of *CER*, as substantiated by the consistent grouping of the BMP4 dominant conditions in the hierarchical clustering with low *CER* as a common marker.

The response to the BMP4 pathway, however, was highly dependent on the context, namely the presence and absence of FGF2 which was a striking feature of the hierarchical clustering on the 15 conditions. BMP4 is typically known as an activin antagonist and high concentrations of BMP4 in the culture with high activin results in mesoderm fate (Poulain *et al.*, 2006; Sulzbacher *et al.*, 2009; Sumi *et al.*, 2008). At the same time, BMP4 alone results in the extra-embryonic lineages (Xu *et al.*, 2002). The presence of FGF2 with BMP4 modulates the net response to the mesendoderm fate, which is an intermediate stage that can result in DE and mesoderm. Several recent studies have demonstrated the use of this combination to promote endoderm formation (Xu *et al.*, 2011; Yu *et al.*, 2011). FGF2 sustains the expression of *NANOG* (a pluripotency marker) and this sustained *NANOG* expression is found to shift the outcome of BMP4 induced differentiation of hESCs towards mesendoderm (Yu *et al.*, 2011). However, prolonged use of FGF2 and BMP4 together may be detrimental for pancreatic differentiation, since this combination has been shown to induce hepatic differentiation after the DE stage (Zorn and Wells, 2009). Also, BMP4 dominant clusters showed high expression of late endoderm markers *HNF4α, HNF1β* and *GATA4*. This may indicate that BMP4 accelerates the differentiation to the mesendoderm phase and therefore, the overall dynamics may be faster for the BMP4 dominant case. But, it was striking to note that the expression of *HNF6*, another important marker for late endoderm was still lower in the BMP4 dominant case. Hence, hierarchical clustering alone was not sufficient to answer if BMP4 addition could be useful for late endoderm differentiation. Importantly, BMP4 dominant conditions gave low expression of markers from the robust biclusters. Thus the current analysis shows that BMP4 may not be a suitable choice for endoderm induction. WNT3A/β-catenin signaling has been shown to be important both for maintenance of pluripotency as well as induction of differentiation (Zorn and Wells, 2009). The WNT pathway is also found to be important in the formation of primitive streak due to which it is often used in the very early stages of in vitro differentiation until the formation of mesendoderm (D'Amour *et al.*, 2005). Stabilization of β-catenin by canonical WNT signaling is found to be responsible for differentiation by epithelial-mesenchymal transition; however presence of WNT after this stage supports mesoderm (Sumi *et al.*, 2008). Also, FGF2 is found to synergistically influence the WNT pathway (Katoh and Katoh, 2006). WNT alongwith PI3KI was commonly present in both the groups identified by our hierarchical clustering. WNT was consistently found to be supportive to the activin + FGF2 signaling assessed by the up-regulation of DE markers. Hence, WNT and PI3KI may be the essential pathway modulators necessary for endoderm differentiation.

3.4.2 Robust biclusters identify the necessary pathways for efficient endoderm differentiation to the pancreatic lineage

The robust biclusters identified by the biclustering + bootstrap analysis show the most important trends preserved under experimental variations. Supportively, *CER, HNF6* and *HNF4α* belonged to the robust clusters. As mentioned earlier, *CER* is an important target of the activin and WNT signaling pathways and *HNF6* is a very early pancreatic progenitor marker taking part in the transcriptional network activating pancreatic progenitors. As seen from the Group 1 bicluster, FGF2 + WNT3A conditions favor *CER* and *HNF6* while BMP4 limits their up-regulation. It is also found that the stability of β–catenin is partly enhanced by PI3K signaling (activated by FGF2) (Voskas *et al.*, 2010) and hence this combination of high activin + FGF2 + WNT3A may work to control the expression of some endoderm markers like *CER* and *HNF6*. At the same time, CER protein is a negative regulator of the TGF-β (activin, BMP4) pathway and upregulation of CER is necessary to limit the activation of these pathways, since inhibition of the TGF-β pathway was found to be necessary for efficient differentiation to the pancreatic progenitors after PDX1 and HNF6 expression (Nostro *et al.*, 2011). However, external addition of WNT3A may still be necessary since CER negatively regulates the WNT pathway (Katoh and Katoh, 2006).

Alternatively, the markers *HNF4α* and *HNF6* which occur in Group 2 are co-regulated under FGF2 + BMP4, FGF2 + WNT3A + PI3KI action. These markers also occur in the MODY network for induction of Neurogenin expressing cells which represents mature pancreatic lineage (Wilding and Gannon, 2004). *HNF6* occupies a predominant position in regulating the expression of *HNF4α* and other genes prior to *PDX1* induction. A key result identified by the bicluster was the consistent up-regulation of the late pancreatic markers *HNF4α* and *HNF6* under

WNT3A + PI3KI dominant conditions and studies by Nostro *et al.* have indicated the necessity of WNT3A for induction of pancreatic progenitors (Nostro *et al.*, 2011). *CER*, *HNF6* combination was also upregulated under WNT3A conditions and thus WNT3A addition was found to favor both DE markers as well as late pancreatic endoderm markers supposedly showing similarity with in vivo pancreatic organogenesis. The presence of FGF2 and BMP4 lowers the expression of these markers and is consistent with the inhibition of FGF2 and BMP4 at the later stages for inhibition of a hepatic fate and efficient pancreatic lineage selection (Nostro *et al.*, 2011). The key signaling pathway interactions from the robust biclusters are summarized in Figure 3.7.

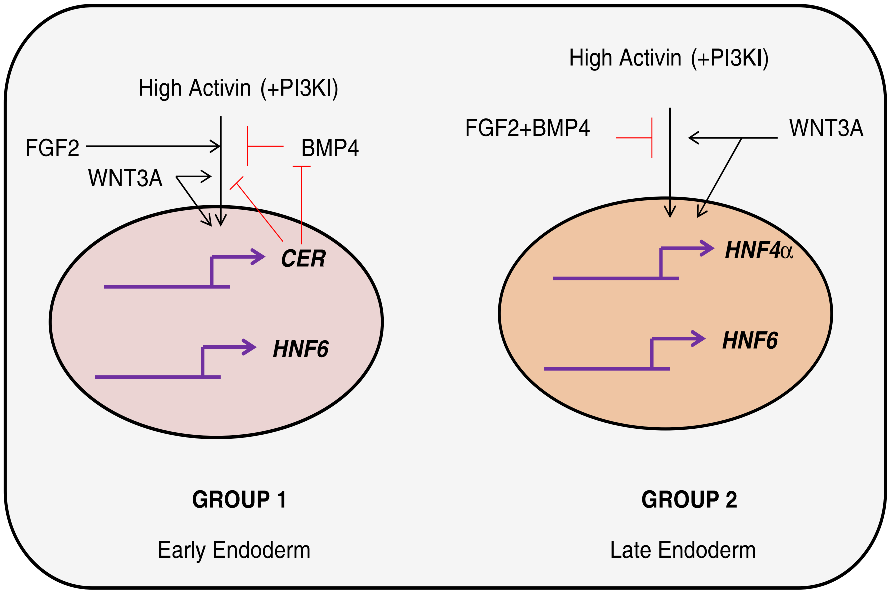


Figure 3.7 Functional dependence of the coregulated genes on the active signaling pathways of endoderm induction.

CER and *HNF6* are favored by High activin and WNT3A, FGF2 while *HNF4α* and *HNF6* are favored by High activin, WNT3A and PI3KI. Combining the early and late stages, high activin with PI3KI and WNT3A together is an effective strategy for endoderm differentiation.

3.5 CONCLUSIONS AND FUTURE DEVELOPMENTS

3.5.1 Major conclusions

The focus of the current work was to achieve insights into the *in vitro* differentiation process of human embryonic stem cells to the endoderm stage using both experimental and mathematical approaches. Our work has identified the differences between the different protocols for endoderm induction. Essentially, high activin A and PI3K inhibition or high activin A with FGF2 or WNT3A serve well as early DE inducer. Additionally, biclustering shows that the early and late endoderm markers are co-regulated under high activin and WNT3A. Thus, overall high activin with PI3KI and WNT3A together may serve better for *in vitro* differentiation of hESCs to the definitive endoderm and pancreatic endoderm lineages. Work by Dalton *et al.* indicate that components of the pathways activated by these conditions are necessary for effective DE lineage specification (Singh *et al.*, 2012b). This condition is currently being used in regular DE differentiation protocols in our lab.

3.5.2 Assumptions, potential pitfalls and proposed extensions

In the current analysis, only a snapshot of the differentiation process was used, since the gene expression differences were evaluated only on day 4. While this is sufficient to analyze differentiation at the final stage of definitive endoderm, it does not give information on the intermediate stages unless the final set of markers selected are such that they have correlations to the markers preceding them. This was taken care of in the current analysis by selecting appropriate markers. However, in future if we would like to expand the network and ensure coexpression of multiple markers (which is a better indicator of the cell state) instead of 2-3 markers in the biclusters identified in this aim, it will be necessary to consider multiple time steps. In such a situation, techniques like tri-clustering would be necessary where the time ordering information is preserved (Gutiérrez-Avilés *et al.*, 2014; Tchagang *et al.*, 2012).

4.0 QUANTITATIVE NATURE OF SMAD SIGNALING IN HESCS AND MODELING CROSSTALK INTERACTIONS WITH AKT

The DBN content of this chapter is taken from Mathew, S., Sundararaj, S. and Banerjee, I., 2015. Network analysis identifies crosstalk interactions governing TGF-β signaling dynamics during endoderm differentiation of human embryonic stem cells. *Processes* 3, 286-308.

4.1 INTRODUCTION

The fate choice of hESCs is controlled by complex signaling milieu synthesized by diverse chemical factors in the growth media. Prevalence of crosstalks and interactions between parallel pathways renders any analysis probing the dynamics of fate choice elusive. Although, some key interactions within major signaling pathways as well as interactions between parallel pathways in differentiating hESCs was recently characterized (Singh *et al.*, 2012a), experimental studies have not focused on the dynamics of signal transduction in differentiating hESCs. Further, the nature of interactions present in the signaling network and the sequence of signal propagation events are cumulatively captured in the dynamics of key molecules in a signaling pathway (Heinrich *et al.*, 2002). In the previous chapter, our focus was on one time snapshot of the endoderm differentiation process. Here, we are focusing on the entire early dynamics of signaling molecules, but using two major minimal pathways for a thorough analysis. In this aim we (i) determined the crosstalk network interactions and (ii) developed the mechanistic model of these interactions from the experimental dynamics of signaling molecules. The quantitative and experimental approaches applied in this aim are new to the hESC system and were particularly selected to address high variability in the dynamic signaling data of hESCs.

The first step in the process is network identification. Previous studies have shown that the effectors of TGF-β pathway (SMAD molecules) play a major role in fate choice of hESCs (Singh *et al.*, 2012a). The Activin mediated TGF-β/SMAD pathway is a major pathway associated with many functions of organ development including proliferation, differentiation, migration and cell death. This pathway is activated by superfamily of cytokine ligands (TGF-β, Activin, Inhibin, Nodal and Lefty) that activate serine/threonine signaling (Clarke and Liu, 2008; Hagos and Dougan, 2007; Schier, 2009). Among these ligands, Nodal is the primary ligand during embryonic development and its functions are mimicked by Activin ligand in *in vitro* cultures (Schier, 2009). TGF-β is the ligand that is commonly associated with inflammation, tissue homeostasis and cancer cell signaling (Clarke and Liu, 2008; Massagué, 2012). However, all of these ligands activate the same pathway and include the molecules SMAD2 and SMAD3 as their primary effectors. However, the context of survival pathways like PI3K/AKT and mitogen activated pathways like MAPK/ERK ultimately decides whether active SMAD complexes support self-renewal or differentiation of hESCs. This is because of critical crosstalk interactions between TGF-β/SMAD and these other pathways (Dalton, 2013). The efficiency of endoderm differentiation is consequently diminished without appropriate removal of negative interactions with parallel pathways. Until now, there has not been a thorough mathematical and network level analysis of the existing interactions, which is the focus of this aim. Due to the high variability associated with hESC systems, it is also necessary to infer robust connections from noisy data. Bayesian models provide a natural framework to investigate the causal dependence between nodes in a network and derive probabilistic relationships that most likely explain experimental observations (Needham *et al.*, 2007). These models have proven successful in network reconstruction from noisy signal transduction data (Woolf *et al.*, 2005; Zielinski *et al.*, 2009). Among the different Bayesian models, Dynamic Bayesian Networks (DBNs) provide the best representation of the adaptive nature of signal transduction networks (Murphy, 2002). DBNs provide information on the conditional dependencies between participating molecules from their measured time series. As a first step for network identification, a multiplex measurement platform was used to measure detailed dynamics of multiple signaling molecules of the TGFβ pathway along with key crosstalk molecules. The measurements were made under Activin induction condition along with a perturbed case where PI3K pathway was simultaneously inhibited. We observed divergent dynamics of SMAD signaling molecules between these two conditions. DBN inference results conducted on the entire time series of key signaling molecules identified molecule from PI3K/AKT pathway (p-AKT) as a major molecule of crosstalk with the TGF-β/SMAD pathway.

The DBN inference does not give information on the kinetics of signal transduction, for which a more detailed reaction rate based approach like ODE is necessary. Therefore, a detailed ODE based mathematical model of the activin stimulated TGF-β/SMAD pathway was developed and the model was calibrated to the experimental data in hESCs. Using the model, we explored the reason for the divergent dynamics of SMAD molecules in hESCs. Using the most important crosstalk molecule p-AKT, different scenarios were tested for the actual mechanism of AKT and SMAD interactions. We utilized an ensemble parametric estimation process to check which of these mechanisms explains the experimental observations. We identified differences between the competing mechanisms that can be experimentally tested. This resulted in a comprehensive model of the TGF-β/SMAD pathway for hESCs differentiating to endoderm with the most valid p-AKT mediated crosstalk interactions. Future developments of the model by incorporating the entire PI3K/AKT pathway developed in Chapter 2 will enable rational control of the differentiation process.

4.2 METHODS (DBN INFERENCE)

4.2.1 Experimental treatments

H1 hESCs were placed on hESC certified Matrigel (BD Biosciences, Billerica, MA, USA)-coated tissue culture plate for 5–7 days in mTESR1 (Stemcell Technologies, Vancouver, BC, Canada) at 37 °C and 5% CO2 before passaging. Cells were examined under the microscope every day and colonies with observable differentiation were picked and removed before the media changes. The maintenance protocol was adopted from our previous studies (Jaramillo *et al.*, 2014; Mathew *et al.*, 2012; Richardson *et al.*, 2014).

hESCs were allowed to grow to 60%–70% confluency before experiments were started. Once confluency was reached, endoderm differentiation was induced by adding 100 ng/mL Activin A (R & D Systems, Minneapolis, MN, USA) in the presence or absence of 1 μM Wortmannin (PI3K inhibitor; Sigma-Aldrich, St. Louis, MO, USA) for 24 h (or otherwise indicated). In the remainder of this chapter, these conditions are called as high and low PI3K respectively. The differentiation media were made using DMEM/F12 (Life Technologies, Grand Island, NE, USA), supplemented with 0.2% bovine serum albumin (BSA; Sigma-Aldrich, St.

Louis, MO, USA) and 1xB27 (Life Technologies, Grand Island, NE, USA). The induction protocol for endoderm was adopted from our previous study (Jaramillo *et al.*, 2014; Mathew *et al.*, 2012).

4.2.2 Experimental time series data

Intracellular expression of signaling proteins were measured by MagPix analysis using the TGFβ Signaling Pathway Magnetic Bead 6-Plex Cell Signaling Multiplex Assay (EMD Millipore, Catalog no.: 48-614MAG) according to manufacturer's instructions. The detailed protocol for MagPix is described in Section 2.2. Mean fluorescence intensity (MFI) was measured using the xMAP (Luminex, Madison, WI, USA) instrument. Measurements were obtained for 6 analytes, namely total TGFβ receptor 2 (t-TGFβRII), total SMAD4 (t-SMAD4), phosphorylated SMAD2 (p-SMAD2 Ser465/Ser467), p-SMAD3 (Ser423/Ser425), p-AKT (Ser473) and p-ERK (Thr185/Tyr187). The time points selected for analysis were: 0, 0.5, 1, 1.5, 2, 3, 6, 12, 18 and 24 h (10 time points, each from a different well of tissue culture plate). Three repeats were conducted per experimental condition and quantitative analysis was performed on each repeat separately. Total protein content of the sample was measured using BCA total protein kit (Thermo Scientific, Grand Island, NE, USA), according to manufacturer's instructions.

The entire data (6 molecules) were used for DBN inference. To apply the algorithm to the high and low PI3K data, the data were preprocessed by normalizing the raw MFI values of each protein by its maximum MFI value for the given time series. We tested other types of normalization like mean centering and variance scaling and found that this did not change the most important results. A common concern with biological datasets is the inherent variability arising from batch-to-batch and well-to-well variability. This is further enhanced in hESC systems, used in the current work, due to inherent variations in differentiation, which cannot be conveniently controlled in the current experimental setting. However, even though the individual repeats elicited high variability in measured MFI values, many features of the overall protein dynamics was largely conserved. Therefore, DBN was repeated separately on each experimental repeat and the commonly repeated connections were collected together in a consensus graph.

4.2.3 Identification of network interactions

Bayesian networks are probabilistic graphical models that relate nodes via directed edges, with the direction showing the causal relationship between the nodes (Needham *et al.*, 2007). These relationships are stronger as compared to correlative methods. Graphical models have nodes that represent entities that can interact (here molecules) and edges show how the nodes influence each other. The node where the edge originates is commonly called a parent node and the node where the edge ends is called a child node. Each node in the network is described by conditional probabilities as tables or functions. In continuous space, the relationship is represented by conditional probability distributions, and Gaussian distributions are commonly used to model the relationships (Koller and Friedman, 2009; Needham *et al.*, 2007). Bayesian networks however cannot represent cyclic loops like feedbacks that are common in signal transduction networks. The problem of cyclic loops can be overcome by use of a generalization of Bayesian networks via DBNs (Grzegorczyk and Husmeier, 2011b).

4.2.3.1 Details of DBN algorithm

DBNs relate variables between adjacent time points such that a child node at a given time point is related to the parent nodes at a previous time point, thereby expanding the network in time.

Based on the system and the dynamics, the relationship can go back one or several time steps. A common approach to construct DBN is by using score equivalence criterion (Koller and Friedman, 2009). Here, a scoring metric (for example, maximum likelihood (ML) estimate in combination with regularization strategies) is used to evaluate how well a graph reconstructs the experimental data. Although DBNs provide good representation of biological networks, they are computationally expensive. Grzegorczyk *et al.* developed a computationally efficient algorithm to identify non-stationary DBNs (Grzegorczyk and Husmeier, 2011b). Specifically, in non-stationary DBNs, the network structure is kept constant between different time points, but the model parameters are allowed to vary between different time segments. The method has been successful in discovering biologically relevant interactions from diverse biological data sets including times series of gene expression using qRT-PCR and MagPix protein concentrations across species (Aerts *et al.*, 2014; Azhar *et al.*, 2013; Dojer *et al.*, 2006; Emr *et al.*, 2014; Grzegorczyk and Husmeier, 2011a). The model systems are diverse, including circadian rhythms in *A. thaliana*, morphogenesis in *D. melanogaster*, synthetic metabolic networks in *S. cerevisiae*, serum inflammatory cytokine mediators in pediatric acute liver injury etc. Full details of the algorithm are presented in the manuscript and supplementary material of Grzegorczyk et al. (Grzegorczyk and Husmeier, 2011b). A brief discussion of the algorithm based on the original manuscript is presented below.

Consider a set of N interacting nodes of a signaling network represented by $X_1, X_2, ..., X_n$ and a directed graph structure G. An edge pointing from X_i to X_j in a DBN with time lag equal to one time step shows that the realization of X_j at time step t is dependent on the realization of its parent X_i at time step $t-1$. It is commonly assumed that a time lag equal to one time step is sufficient to represent the relationship, indicating that the data have to be sampled at the right time intervals for the dynamics to be represented correctly. The parent node set, π_j, of a node X_j is the set of all nodes from which an edge points to X_j in G. Grzegorczyk *et al.* proposed a non-stationary generalization of the Bayesian Gaussian with score equivalence model (called BGe), and it is a node-specific mixture of BGe models (Grzegorczyk and Husmeier, 2011b). The non-stationary DBN is based on the following Markov chain expansion:

$$P\left(D\middle|G,\underline{V},\underline{K},\underline{\theta}\right)=\prod_{n=1}^{N}\prod_{t=2}^{m}\prod_{k=1}^{\kappa_n}\psi\left(D_n^{\pi_n}\left[t,\underline{\theta}^k{}_n\right]\right)^{\delta_{V_n(t),k}} \qquad (4.1)$$

$$\psi\left(D_n^{\pi_n}\left[t,\underline{\theta}^k{}_n\right]\right)^{\delta_{V_n(t),k}} = P\left(X_n(t)=D_{n,t}\middle|\pi_n(t-1)=D_{\pi_n,t-1},\underline{\theta}^k{}_n\right) \qquad (4.2)$$

where, D is the time course data, $\delta_{V_n(t),k}$ is the Kronecker delta, $\underline{V}$ is a matrix of latent variables that indicate which BGe mixture component generates a data point, $\underline{K}=\left(\kappa_1,\kappa_2,..,\kappa_n\right)$ is a vector of mixture components, m is the total number of time points. Vectors and matrices are denoted by single underbars in the symbols of all the equations of this manuscript. Each column of matrix $\underline{V}$ is the vector $\underline{V}_n$, which divides the time series for a node into different time segments. The endpoints of these time segments are called as change-points. Each time segment between change-points is a different BGe model with parameters $\theta^k{}_n$, which includes the mean and covariance matrix of the conditional dependences for the mixture component. The allocation scheme in Equation 4.1 provides representation of a nonlinear regulatory process by a piecewise linear process. From Equation 4.1, the marginal likelihood conditional on the latent variables is given by:

$$P\left(D\middle|G,\underline{V},\underline{K}\right)=\int P\left(D\middle|G,\underline{V},\underline{K},\underline{\theta}\right)P\left(\underline{\theta}\right)d\theta=\prod_{n=1}^{N}\psi^*\left(D_n^{\pi_n}\left[\kappa_n,\underline{V}_n\right]\right) \qquad (4.3)$$

$$\psi^*\left(D_n^{\pi_n}\left[\kappa_n, \underline{V}_n\right]\right) = \prod_{k=1}^{\kappa_n} \psi\left(D_n^{\pi_n}\left[k, \underline{V}_n\right]\right) \tag{4.4}$$

$$\psi\left(D_n^{\pi_n}\left[k, \underline{V}_n\right]\right) = \int \prod_{t=2}^{m} P\left(X_n(t) = D_{n,t} \middle| \pi_n(t-1) = D_{(\pi_n, t-1)}, \underline{\theta}^k{}_n\right) P\left(\underline{\theta}^k{}_n \middle| \pi_n\right) d\,\underline{\theta} \tag{4.5}$$

Equation 4.4 is the local change-point BGe score (called as cpBGe) for node n. In this work, a Gibbs MCMC sampling scheme was followed to sample from the local posterior distributions. Although, the location of change-points is inferred, the actual values of the parameters are not directly obtained since they are integrated out as seen from Equation 4.3. In the algorithm, the change-points were sampled from a point process prior using dynamic programming and the graphs were sampled by sampling parent node set (restricted to 3 parents per node) from a Boltzmann posterior distribution using the cpBGe score. Additional details of the sampling procedure are given in (Grzegorczyk and Husmeier, 2011b). Algorithm and code developed by Azhar *et al.* was used in this work which was based on the work by Grzegorczyk *et al.* (Azhar *et al.*, 2013; Grzegorczyk and Husmeier, 2011b). The sampling parameters were kept at nominal values suggested by the authors. All simulations were performed in MATLAB® (Natick, MA, USA) on Linux 64-bit platform and single core of INTEL® (Santa Clara, CA, USA) CoreTM 2 Quad CPU (Q8400 @ 2.66 GHz).

4.2.3.2 Constructing the DBNs

The DBN inference was performed on each of the two experimental conditions separately to identify the network interactions that exist in each condition. The marginal edge probability was monitored for each Gibbs sampling step. The marginal edge probability for a given edge denotes the fraction of the graphs in which that edge was present. Each Gibbs sampling step represents an instance of the network that can best explain the experimental time series. In the early phases of the simulation, the network is not yet stabilized and hence, the likelihood scores and the marginal edge probabilities fluctuate. The marginal edge probabilities of the final network were calculated after a burn-in phase when the distributions have stabilized. The marginal edge probability scores from networks obtained for the three experimental repeats were averaged to obtain a consensus network for a given condition or time zone. Finally, only those edges that were present in more than 50% of the sampled graphs were kept in the consensus DBN, a criterion used in the study by Azhar *et al.* (Azhar *et al.*, 2013). Any value less than 50% indicates that the number of samples in which the associated edge was absent is more than the number of samples in which it is present.

4.3 RESULTS AND DISCUSSION (DBN INFERENCE)

4.3.1 Experimental dynamics of signaling molecules

Figure 4.1 shows the dynamics of six signaling molecules after activin addition in the presence (shown by blue dashed line) and absence (shown by red continuous line) of PI3K inhibitor, called as low and high PI3K conditions respectively. The original data was normalized by time 0 values to obtain the fold change. The mean levels and standard deviation from 3 experimental repeats are plotted here. The time points selected for the study include: 0, 0.5, 1, 1.5, 2, 3, 6, 12, 18 and 24 h. The high PI3K condition represents the differentiation protocol where only the TGF-β/SMAD2,3 pathway is externally activated while the PI3K/AKT pathway is left unperturbed. In this condition, p-AKT levels are maintained near the basal levels, only slightly lower (Figure 4.1A). For the purpose of this manuscript, the basal levels are defined as the protein levels at time 0. It is seen that the mean levels of p-AKT fluctuate in the early time points (< 6 h). Levels of t-TGFβRII (Figure 4.1B) also remain close to basal levels under high PI3K signaling. For p-SMAD2 (Figure 4.1C), an overshoot behavior is seen with levels reaching the maximum within 2-3 hours and settling at intermediate levels by 6 h. For p-SMAD3 (Figure 4.1D), the dynamics shows a different behavior than p-SMAD2 even though both are activated by the same ligand-receptor complex. In general, the dynamics shows a continuous increase instead of the overshoot behavior seen for p-SMAD2. t-SMAD4 (Figure 4.1E) is maintained near the basal levels for this condition. p-ERK shows a minimal and delayed increase (Figure 4.1F) under high PI3K.

The low PI3K condition represents a modulation over the high PI3K condition with the PI3K/AKT pathway externally inhibited in addition to activation of TGF-β/SMAD2,3 pathway. In this condition, we see a considerable decrease in p-AKT levels since it is a downstream effector of PI3K signal (Figure 4.1A). However, interestingly this decrease is short-lived. Even after continued inhibition of PI3K, the levels of p-AKT start increasing from 3 hours with the levels reaching near basal levels by 12 h. The levels of t-TGFβRII in this condition are lower than high PI3K condition at time points from 6 h (Figure 4.1B). The dynamics of p-SMAD2 is similar to the high PI3K condition (Figure 4.1C) with slightly higher fold-change at early time points. The fold-change in p-SMAD3 is higher compared to high PI3K signaling and it also shows substantial increase at later time points. t-SMAD4 (Figure 4.1E) shows fluctuations at early time points and a slight reduction at later time points (from 6 h). p-ERK (Figure 4.1F) shows a slow rise as compared to p-SMAD2,3 and the increase is substantial as compared to high PI3K signaling. Thus, overall, the low PI3K condition results in higher fold-changes in levels of phosphorylated SMAD3 and ERK than high PI3K condition. The low PI3K condition favors differentiation to endoderm over a 4-day differentiation process (Figure C.1).

The dynamics shown in Figure 4.1 is the first detailed study of signaling dynamics obtained for hESCs under endoderm induction conditions. Two unique features are observed for hESCs, namely the rise in p-AKT levels under continued PI3K inhibition and the divergent dynamics of p-SMAD2 and p-SMAD3 under high Activin levels. Further, as is typical for hESC system, there is high degree of variability in the levels of most molecules and the degree of variability is different at different time points. The variability is higher for low PI3K condition, a possible effect resulting from high degree of cell death observed in this condition since PI3K is an important cell survival pathway. The differences in the levels and dynamics of molecules between high and low PI3K conditions indicate existence of crosstalk interactions between the TGF-β/SMAD2,3, PI3K/AKT and MAPK/ERK pathways. Previous reports from the Dalton group has indicated interactions between these pathways using static end-point analysis (Singh *et al.*, 2012a). Here, we use a computational framework to identify all possible interactions from the information contained in the signaling dynamics.

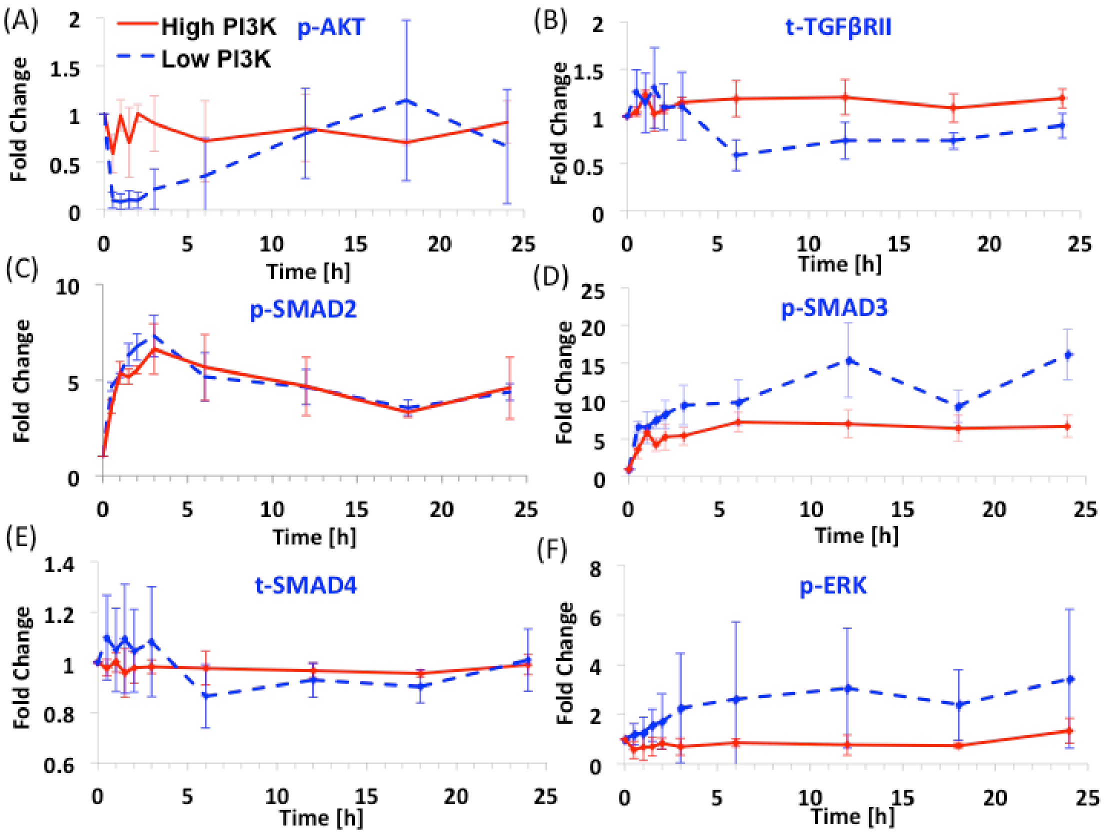


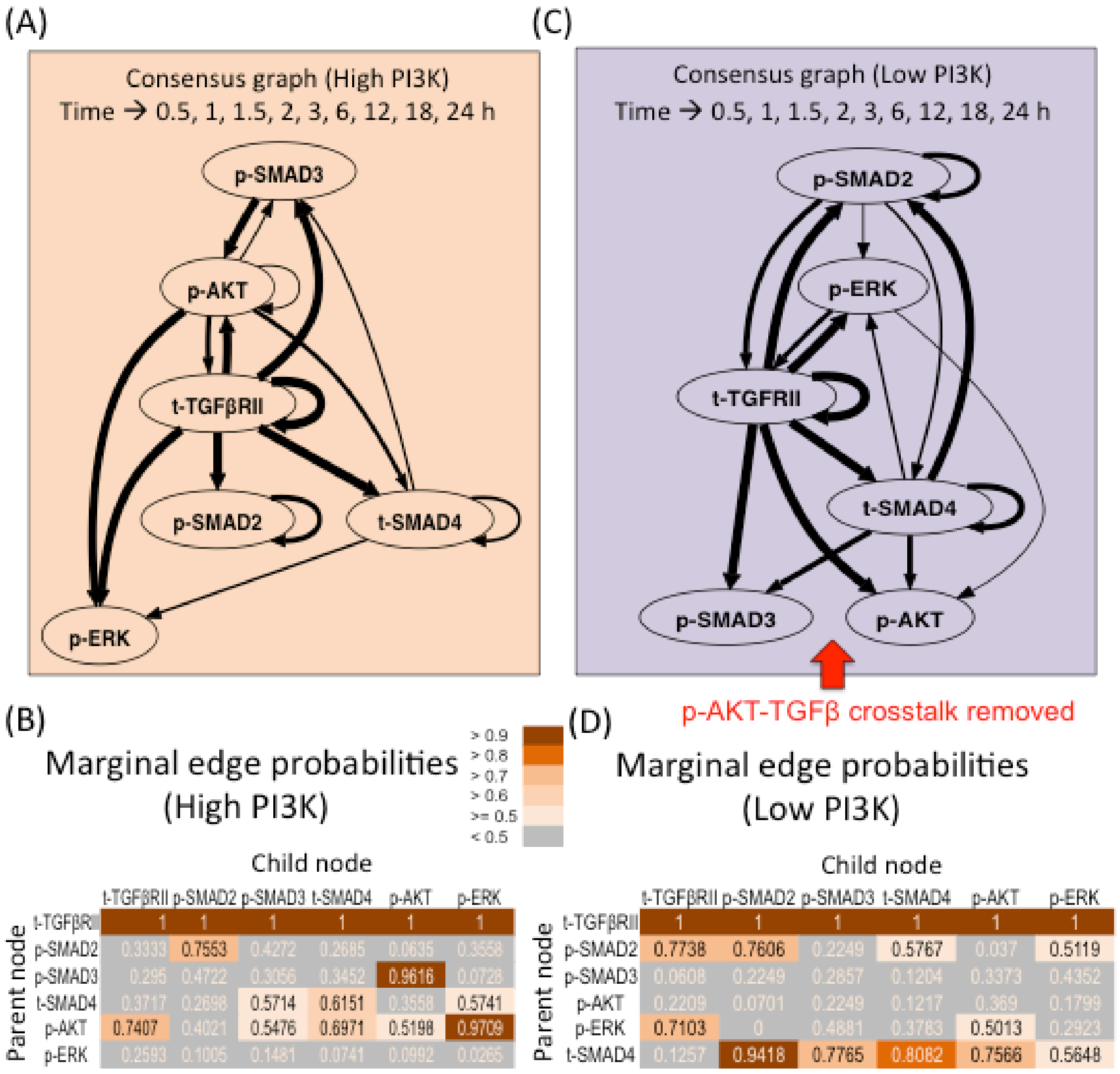
Figure 4.1 Dynamics of key molecules from the TGF-β/SMAD, PI3K/AKT and MAPK/ERK pathways for two endoderm induction conditions.

4.3.2 Predictions of network interactions by DBN inference

We next applied DBN inference on the entire time series data for each repeat separately. This gives rise to three DBNs for each condition. The connections identified in each repeat are compiled together to form a consensus graph.

4.3.2.1 Consensus graph

Figure 4.2A and Figure 4.2C shows the consensus digraphs for the high and low PI3K data. The convergence diagnostics for the DBNs for each repeat is presented in Figure C.2. The log likelihood score stabilized very early in the sampling runs for both conditions (see Figure C.2A-B). For the current data, it was found that 250 Gibbs sampling steps were sufficient to converge to the marginal edge posterior distribution (see Figure C.2C-D). This was confirmed over independent sampling runs, due to stochastic nature of the algorithm. Then, 500 sampling steps were performed to obtain enough samples in the converged region to calculate the marginal edge probabilities. At the end of 500 Gibbs sampling steps, the final marginal edge probabilities were calculated using the later half of the 500 samples (the early half belongs to the burn-in phase of the simulation). The mean marginal edge probabilities from the three samples are presented in Figure 4.2B and Figure 4.2D. Any edge, which was present in less than 50% of the samples, was removed from the consensus graph. Note that the DBN for each sample represents the network that can explain the entire time series of that sample, with only network parameters allowed to vary between time segments. We also tested the robustness of the connections by increasing the number of time points (by interpolation of the data). The number of time points was increased until 20 and no changes in the major connections of the dataset were observed.


(A)
Consensus graph (High PI3K)
Time → 0.5, 1, 1.5, 2, 3, 6, 12, 18, 24 h
p-SMAD3
p-AKT
t-TGFβRII
p-SMAD2
t-SMAD4
p-ERK
(C)
Consensus graph (Low PI3K)
Time → 0.5, 1, 1.5, 2, 3, 6, 12, 18, 24 h
p-SMAD2
p-ERK
t-TGFRII
t-SMAD4
p-SMAD3
p-AKT
p-AKT-TGFβ crosstalk removed
(B)
Marginal edge probabilities
(High PI3K)
> 0.9
> 0.8
> 0.7
> 0.6
>= 0.5
< 0.5
(D)
Marginal edge probabilities
(Low PI3K)
Child node
Parent node

Parent node	t-TGFβRII	p-SMAD2	p-SMAD3	t-SMAD4	p-AKT	p-ERK
t-TGFβRI	1	1	1	1	1	1
p-SMAD2	0.3333	0.7553	0.4272	0.2685	0.0635	0.3558
p-SMAD3	0.295	0.4722	0.3056	0.3452	0.9616	0.0728
t-SMAD4	0.3717	0.2698	0.5714	0.6151	0.3558	0.5741
p-AKT	0.7407	0.4021	0.5476	0.6971	0.5198	0.9709
p-ERK	0.2593	0.1005	0.1481	0.0741	0.0992	0.0265

Parent node	t-TGFβRII	p-SMAD2	p-SMAD3	t-SMAD4	p-AKT	p-ERK
t-TGFβRII	1	1	1	1	1	1
p-SMAD2	0.7738	0.7606	0.2249	0.5767	0.037	0.5119
p-SMAD3	0.0608	0.2249	0.2857	0.1204	0.3373	0.4352
p-AKT	0.2209	0.0701	0.2249	0.1217	0.369	0.1799
p-ERK	0.7103	0	0.4881	0.3783	0.5013	0.2923
t-SMAD4	0.1257	0.9418	0.7765	0.8082	0.7566	0.5648

Figure 4.2 Dynamic Bayesian Networks inferred for endoderm induction conditions.

(A) Consensus graph for high PI3K data. The thickness of the edges reflects the value of edge probabilities (>=0.5). (B) Marginal edge probability table for high PI3K data. The parent node is the node whose value at time (t-1) affects the value of child node at time t. (C) Consensus graph for low PI3K data. (D) Marginal edge probability table for low PI3K data.

4.3.2.2 High PI3K condition

The consensus graph shows the average interactions that are present for the given experimental condition across the three samples. As seen from Figure 4.2A, the dynamics of the receptor influences all the other molecules in the network, both molecules of the TGF-β pathway (p-SMAD2,3, SMAD4) and molecules of parallel pathways (p-AKT and p-ERK). The receptor is also self-regulated. The mean marginal edge probabilities in Figure 4.2B show that these edges are present in 100% of the sampled graphs (t-TGFβRII as the parent node and all the other molecules including the receptor being the child node). Next common edges include p-ERK regulation by p-AKT and p-AKT regulation by p-SMAD3 present in 97 and 96% of the graphs respectively. Remaining possible interactions include: regulation of the receptor levels by p-AKT (74%), t-SMAD4 by p-AKT and t-SMAD4 (60-70%), p-SMAD3 by p-AKT (55%), p-AKT self-regulation (52%), p-SMAD3 and p-ERK by t-SMAD4 (57%). The graphs for the individual repeats are presented in Figure C.3.

4.3.2.3 Low PI3K condition

For the low PI3K condition, the edges originating from the receptor are similar to the high PI3K case and are also reflected in 100% of the graphs (Figure 4.2C-D). Next highly represented edges include p-SMAD2 regulation by t-SMAD4 (94%) and t-SMAD4 self-regulation (81%). Remaining possible interactions include: t-SMAD4 as a parent node for p-SMAD3 (77%), p-AKT (76%), p-ERK (56%), p-ERK as the parent node for t-TGFβRII (71%) and p-AKT (50%), p-SMAD2 as a parent node for t-TGFβRII (77%), p-SMAD2 (76%), t-SMAD4 (58%) and p-ERK (51%). The graphs for the individual repeats are presented in Figure C.4.

4.3.2.4 Comparison between digraphs of high and low PI3K conditions

Influence of total receptor levels

The DBN inference identified several similarities and differences in the interactions present in the two conditions. Firstly, the dynamics of the total receptor levels affect the downstream molecules in both the conditions. This influence of total receptor levels is reflected in all the individual samples across both the conditions (Figure C.3-C.4). This indicates that the changes in the receptor levels are important in influencing the downstream molecules during endoderm induction.

Interactions between intracellular molecules

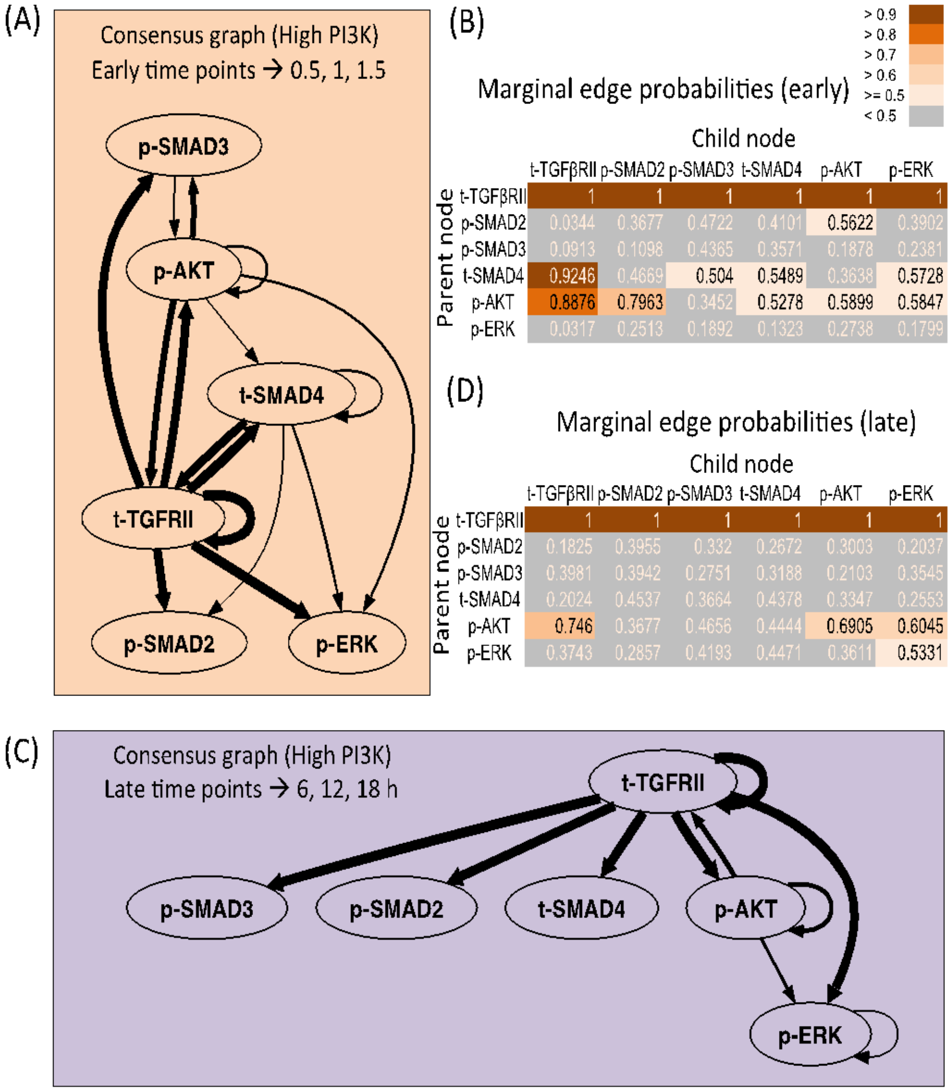
Among the TGF-β pathway molecules, p-SMAD2 has increased regulatory interactions in the low PI3K condition, especially influencing the receptor levels. Further, p-SMAD2 shows influence on p-SMAD3 and p-ERK in sample 1 of high PI3K (Figure C.3). p-SMAD3 shows interactions with p-AKT in the high PI3K condition. This interaction is removed in the low PI3K condition. The low PI3K condition also shows increased role for t-SMAD4 in influencing the p-SMAD2 and p-SMAD3 dynamics. p-AKT shows striking differences in the connections between the two conditions. For example, p-AKT interacts with and regulates majority of the nodes in the high PI3K condition. However, in the low PI3K condition, p-AKT does not regulate other nodes, but instead acts as a child node for all of its interactions. This is an interesting prediction, because the levels of p-AKT increased back in spite of continued inhibition in the low PI3K condition. The current analysis indicates that a short-term decrease in p-AKT levels is sufficient to remove the influence of p-AKT on TGF-β pathway molecules. Next important difference is in the regulatory role of p-ERK. p-ERK is not regulating any of the other nodes in the high PI3K condition. This is also reflected in each of the repeats in high PI3K condition (Figure C.3). Interestingly in the low PI3K condition, p-ERK takes an important role in regulating the receptors. p-ERK shows increased regulatory role on p-SMAD3 and t-SMAD4 in one of the samples (Figure C.4).

4.3.3 Changes in regulatory structure across time zones

Next we selected two time segments (0.5, 1, 1.5 h, called as early) and (6, 12, 18 h, called as late) to check if the regulatory interactions existing in the early and late zones of the dynamics is the same. This is necessary to check if the crosstalk interactions exist throughout the 24 hr time series, or only in certain time zones. DBN inference was done on each zone separately. It is important to note that each of the resulting networks is particular to the time segment of interest since the algorithm has not seen data from the other zone. Nevertheless, the regulatory structure identified in each segment will confirm if these segments contain similar information as any other portion of the dynamics. We increased the sampling frequency in each segment by interpolation of the data, so that there are 10 time points in each of the time zones.

4.3.3.1 High PI3K condition

Figure 4.3A-B presents the consensus graph and marginal edge probabilities respectively for the early time points. The network is very similar to the network obtained using the entire time series of high PI3K condition (Figure 4.2A), with some minor differences. The key regulations by the receptor as well as supplementary crosstalk interactions are identified from the early time points. Figure 4.3C-D presents the consensus graph and marginal edge probabilities respectively for the late time points. The network obtained only contains regulation by the receptor and some repeats contain the regulation by p-AKT on the receptor and p-ERK levels.



Marginal edge probabilities (early)

Child node

Parent node	t-TGFβRII	p-SMAD2	p-SMAD3	t-SMAD4	p-AKT	p-ERK
t-TGFβRII	1	1	1	1	1	1
p-SMAD2	0.0344	0.3677	0.4722	0.4101	0.5622	0.3902
p-SMAD3	0.0913	0.1098	0.4365	0.3571	0.1878	0.2381
t-SMAD4	0.9246	0.4669	0.504	0.5489	0.3638	0.5728
p-AKT	0.8876	0.7963	0.3452	0.5278	0.5899	0.5847
p-ERK	0.0317	0.2513	0.1892	0.1323	0.2738	0.1799

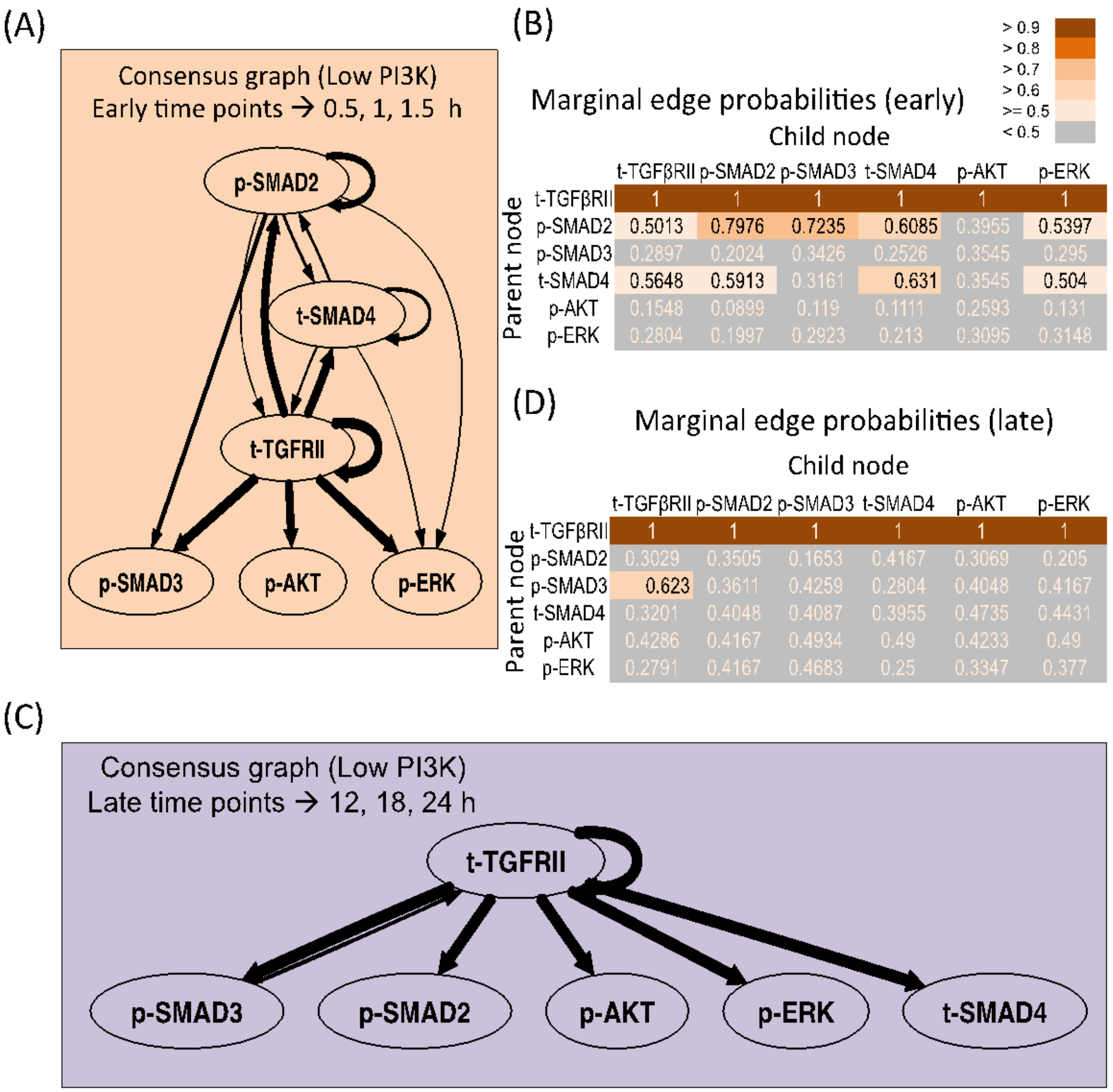
Marginal edge probabilities (late)

Child node

Parent node	t-TGFβRII	p-SMAD2	p-SMAD3	t-SMAD4	p-AKT	p-ERK
t-TGFβRII	1	1	1	1	1	1
p-SMAD2	0.1825	0.3955	0.332	0.2672	0.3003	0.2037
p-SMAD3	0.3981	0.3942	0.2751	0.3188	0.2103	0.3545
t-SMAD4	0.2024	0.4537	0.3664	0.4378	0.3347	0.2553
p-AKT	0.746	0.3677	0.4656	0.4444	0.6905	0.6045
p-ERK	0.3743	0.2857	0.4193	0.4471	0.3611	0.5331

Figure 4.3 Dynamic Bayesian Network inferred for endoderm induction conditions under different time zones and high PI3K.

(A) Consensus graph for high PI3K data, early dynamics (t = 0.5, 1, 1.5 h). (B) Marginal edge probability table for high PI3K data, early dynamics. (C) Consensus graph for high PI3K data, late dynamics (t = 6, 12, 18 h). (D) Marginal edge probability table for high PI3K data, late dynamics.



Marginal edge probabilities (early)

Child node

Parent node	t-TGFβRII	p-SMAD2	p-SMAD3	t-SMAD4	p-AKT	p-ERK
t-TGFβRII	1	1	1	1	1	1
p-SMAD2	0.5013	0.7976	0.7235	0.6085	0.3955	0.5397
p-SMAD3	0.2897	0.2024	0.3426	0.2526	0.3545	0.295
t-SMAD4	0.5648	0.5913	0.3161	0.631	0.3545	0.504
p-AKT	0.1548	0.0899	0.119	0.1111	0.2593	0.131
p-ERK	0.2804	0.1997	0.2923	0.213	0.3095	0.3148

Marginal edge probabilities (late)

Child node

Parent node	t-TGFβRII	p-SMAD2	p-SMAD3	t-SMAD4	p-AKT	p-ERK
t-TGFβRII	1	1	1	1	1	1
p-SMAD2	0.3029	0.3505	0.1653	0.4167	0.3069	0.205
p-SMAD3	0.623	0.3611	0.4259	0.2804	0.4048	0.4167
t-SMAD4	0.3201	0.4048	0.4087	0.3955	0.4735	0.4431
p-AKT	0.4286	0.4167	0.4934	0.49	0.4233	0.49
p-ERK	0.2791	0.4167	0.4683	0.25	0.3347	0.377

Figure 4.4 Dynamic Bayesian Network inferred for endoderm induction conditions under different time zones and low PI3K.

(A) Consensus graph for low PI3K data, early dynamics (t = 0.5, 1, 1.5 h). (B) Marginal edge probability table for low PI3K data, early dynamics. (C) Consensus graph for low PI3K data, late dynamics (t = 12, 18, 24 h). (D) Marginal edge probability table for low PI3K data, late dynamics.

4.3.3.2 Low PI3K condition

Figure 4.4A-B presents the consensus graph and marginal edge probabilities respectively for the early time points, averaged over repeats 1 and 3. The network is very similar to the network obtained using the entire time series of low PI3K condition (Figure 4.2B), with some minor differences. The key regulations by the receptor as well as supplementary crosstalk interactions are identified from the early time points. Figure 4.4C-D presents the consensus graph and marginal edge probabilities respectively for the late time points. It is seen that only the receptor-mediated regulation is identified in this region with no additional crosstalk interactions identified.

4.4 CONCLUSIONS (DBN INFERENCE)

4.4.1 Major conclusions

This work is the first account in identifying specific signaling interactions governing endoderm differentiation of hESCs using network analysis tools. The DBNs inferred for the high and low PI3K data accomplished two major tasks: (1) They identified molecular interactions within the TGF-β pathway along-with crosstalk interactions with parallel pathways and (2) They identified distinct pathway regulations during the early and late phases of the signaling dynamics. One key prediction from the entire analysis is the influence of receptor levels on downstream molecules including SMAD, AKT and ERK. In the canonical pathway, TGFβRII is known to activate SMAD molecules after formation of the ligand-receptor complex (Guo and Wang, 2008). TGFβ signaling also participates in several non-canonical signaling leading to activation of PI3K/AKT and MAPK/ERK pathways in many cell lines (Guo and Wang, 2008; Zhang, 2008). Our analysis indicates that the levels of the receptors (TGFβRII) are still in the regime where they are limiting and any change in their level is reflected downstream.

Several important interactions from p-AKT were identified indicating the existence of p-AKT mediated crosstalk in high PI3K condition and its removal under low PI3K. Ours is the first systematic study to identify these crosstalk interactions in differentiating hESCs. The regulation of p-SMAD3 by p-AKT is well known in other cell lines, mainly inhibition of p-SMAD3 phosphorylation by mTORC1 and sequestration of non-phospho SMAD3 by p-AKT (Conery *et al.*, 2004; Danielpour and Song, 2006; Remy *et al.*, 2004; Song *et al.*, 2006; Zhang *et al.*, 2013). The regulation of p-SMAD2 by p-AKT is observed only in one sample of the high PI3K condition (Figure C.3). Literature shows that most negative regulation of p-AKT is on p-SMAD3 and not p-SMAD2 (Song *et al.*, 2006), however some reports indicate negative regulation of both p-SMAD2 and p-SMAD3 by p-AKT in neuroblastoma and CHO cell lines (Qiao *et al.*, 2006; Sun *et al.*, 2006). Therefore, it is possible that influence on p-SMAD2 is weak and therefore, not identified amongst the other strong interactions. The removal of crosstalk interactions with p-AKT in the low PI3K condition is interesting although the actual mechanism needs further study. The regulation of the receptors and t-SMAD4 by p-AKT was also seen but these interactions are not as widely studied as those of p-AKT and p-SMADs.

The DBNs showed regulation of p-ERK by p-AKT in the high PI3K condition. It is well reported that p-ERK is inhibited by p-AKT and many of its downstream effectors (via mTORC1) in multiple cell lines (Aksamitiene *et al.*, 2012). Previous study has shown the interaction between AKT1 and cRAF in hESCs leading to inhibition of RAF/MEK/ERK signaling (Singh *et al.*, 2012a). Our experiments show that the levels of p-ERK are higher in low PI3K condition and the influence of p-AKT on p-ERK is also absent from the low PI3K DBNs. This indicates that this interaction negatively influences endoderm induction. Although not identified with any significance, possibility of p-ERK mediated interactions under low PI3K signaling is interesting. It is known that p-ERK has additional roles in linker phosphorylation of SMAD molecules which can affect the nucleo-cytoplasmic shuttling and ultimately their dynamics as modeled by Liu *et al*. (Liu *et al.*, 2014). This could be the reason for seeing increasing p-ERK influence on SMAD molecules under low PI3K condition in some samples. But, since this was observed in only some samples of low PI3K that used the entire time series information, additional investigation needs to be done. Future studies of long-term p-ERK dynamics (> 24 hr) and perturbation experiments will enable further exploration of this portion of the network. Overall, the identified DBNs demonstrate significant biologically relevant interactions. Such agreement with literature observations along with prediction of additional interactions prove the applicability of quantitative methods like ODEs for teasing out the network level properties of complex systems like hESCs.

4.4.2 Assumptions, potential pitfalls and proposed extensions

One main assumption for the DBN applied here was that the dependences are valid between two adjacent time points. However, often it may span multiple time steps. This requires highly (1) resolved data in the time dimension or (2) addition of intermediate molecular species that help provide edges between two otherwise unconnected nodes, or (3) higher order DBNs, but currently available methods are computationally very expensive for higher order DBNs.

The DBN inference does not provide direct outputs of the type of interactions (positive or negative). These interactions can be tested by direct experimental perturbations. On the other hand, correlation and regression analysis of the original experimental data using a selected node as the output and all its parent nodes as the input can be done. If proper normalization of the data is undertaken, the coefficients of the regression model will indicate the relative strength and sign of the influence of the parent nodes. See Figure C.5 for a preliminary analysis of the DBNs using Pearson correlation metric. It is seen that for the DBNs predicted in previous sections, the direction of correlation for most pairs of molecules follows the literature observations. However, such analysis may be flawed by the assumption of linearity of the model. To overcome this, simple ODE based models, using mass action or other non-linear kinetics depicting the connections in the DBN, may be generated. Simulation of such phenomenological models with suitable selection of rate parameters followed by comparison with the experimental time series will be another check. In the remaining sections, an ODE based model with the direction of interaction based on literature observations is developed and the same experimental dataset is used to check if the model is able to capture the data well.

4.5 METHODS (KINETIC MODELING)

The results from previous section show that p-AKT is the main crosstalk molecule influencing the SMAD levels. DBN, however, does not provide information on the strength of these interactions and the kinetics of the crosstalk process. Further, the mechanism of crosstalk is still not characterized, for example which reaction in the main pathway is influenced by p-AKT? In the remaining sections, we explicitly modeled possible interactions and evaluated how well the competing mechanisms recapture the experimental data from Section 4.3.1. We used an ODE based framework to describe the kinetics of signal transduction in the SMAD pathway followed by incorporation of AKT mediated interactions.

4.5.1 Basal Activin induced TGF-β/SMAD pathway

Due to the importance of activin mediated signaling for DE differentiation, this chapter will focus on activation of SMAD pathway by this ligand alone. Most mechanistic mathematical models available in the literature have been developed for the TGF-β ligand case in cancer cells (Clarke *et al.,* 2006; Schmierer *et al.*, 2008; Vilar *et al.*, 2006; Zi *et al.*, 2011), and therefore, certain aspects of the model have to be re-calibrated for activin case and these will be discussed in the following sections. The entire SMAD pathway can be separated into two modules: (1) Receptor activation, trafficking and regulation, (2) Intracellular SMAD activation and shuttling. A more detailed description of each module is given below and a schematic is presented in Figure 4.5.

4.5.1.1 Receptor activation, trafficking and regulation

Cell surface has two types of activin receptors, ActRI and ActRII (R1 and R2 respectively). These receptors are activated by the same sequence of reactions as TGF-β receptor (Attisano *et al.*, 1996). The Activin ligand in the medium complexes with R2 and activates the receptor. This active receptor phosphorylates receptor R1 and forms a ligand receptor complex (LRC) on the surface. R1 and R2 in the basal and active condition are susceptible to receptor internalization and recycling. Each receptor undergoes degradation in the endosome. LRC also internalize into the endosomes where it forms the active signaling complex that catalyzes phosphorylation of cytoplasmic substrates like R-SMADs (SMAD2 and SMAD3). LRC undergoes dissociation in the endosomal environment leading to release of R1 and R2. These receptors are recycled back and become available for further complexation with activin on the surface while the activin in the endosomes undergoes degradation. LRC levels on the surface are under negative regulation of the pathway via the complex SMAD7/Smurf2/SIK (Kang *et al.*, 2009). SMAD7 molecule in this complex is transcribed by the active SMAD2,3 complexes in the nucleus from the intracellular module. Current models in the literature consider a very crude incorporation of SMAD7 mediated negative feedback, for example the feedback is considered a function of the nuclear level of p-SMAD2-SMAD4 complex. Since the strength of this connection has not been estimated for the hESC case, we incorporated details of SMAD7 transcription and translation into the model and modeled the negative feedback explicitly as a function of SMAD7 molecule levels.

4.5.1.2 Intracellular SMAD activation and shuttling

Non-phosphorylated SMAD2 and SMAD3 (as monomers) undergo continuous nuclear import and export. The initial levels of SMAD2 and SMAD3 in the cytoplasm and nucleus reflect the dynamic equilibrium between these two shuttling processes. Experiments in cancer cells have shown that SMAD2,3 are more abundant in the cytoplasm than the nucleus while SMAD4 is equally distributed between the two compartments (Clarke *et al.*, 2006). Once LRCs are formed in the endosomes, they activate SMAD2 and SMAD3 via C-terminal phosphorylation. These phospho SMADs then undergo complex formation with Co-SMADs like SMAD4. Phosphorylation of SMAD2,3 is essential for complex formation. Various types of complexes are formed after activation; for example, homomeric SMAD complexes like p-SMAD2-p-SMAD2 and p-SMAD3-p-SMAD3 and heteromeric SMAD complexes like p-SMAD2-SMAD4 and p-SMAD3-SMAD4. These complexes also undergo nuclear import. Many mathematical models have shown that complexed SMADs may have a lower nuclear export rate than non-complexed SMADs. In fact, most mathematical models have kept nuclear export rate of complexed SMAD as zero. Export of these nuclear SMADs happen only after they decomplex and convert to monomeric SMADs (p-SMAD or SMAD). The dephosphorylation of p-SMADs is catalyzed by a variety of phosphatases when in the monomeric form and these phosphatases are mainly nuclear in location.

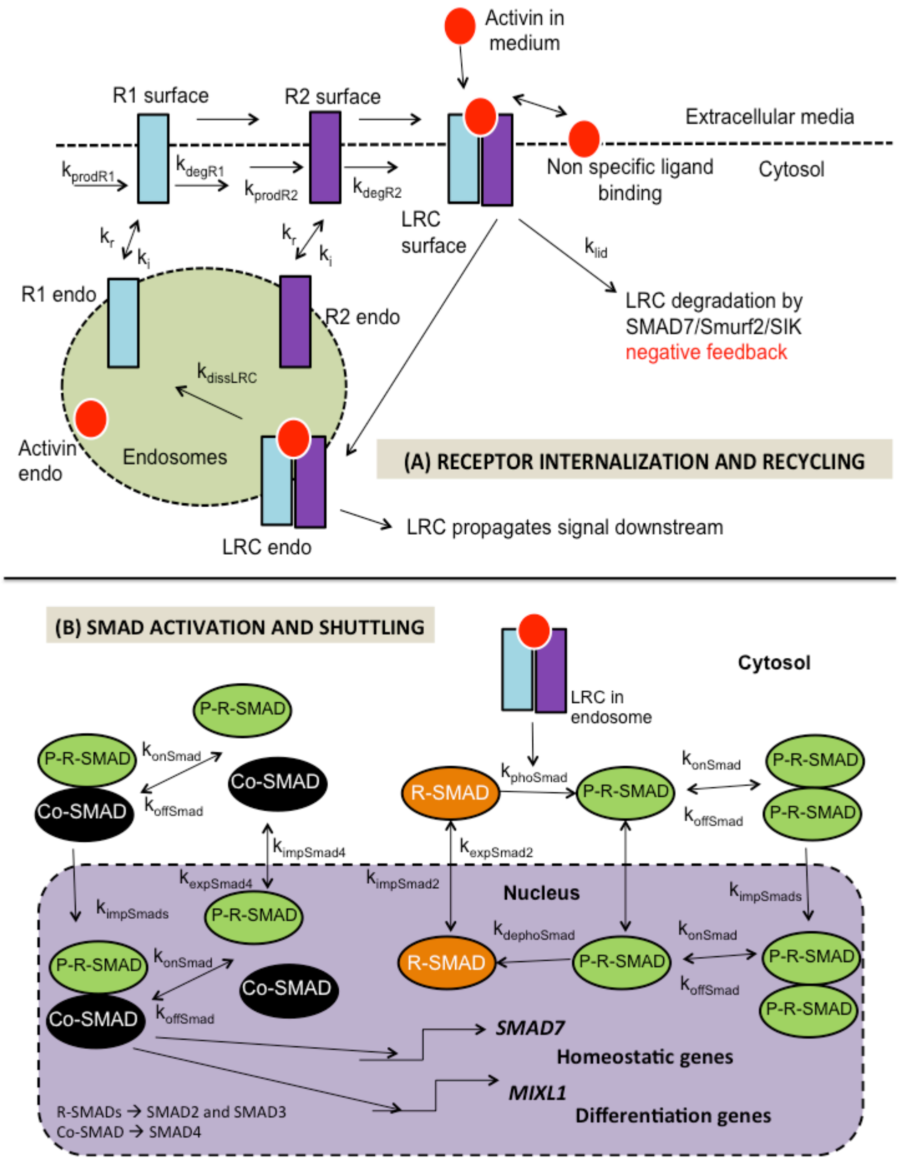


Figure 4.5 Schematic of Receptor level and SMAD level interactions

4.5.2 Mathematical model of basal pathway

The basal ODE model used in this aim was developed for TGF-β/SMAD2 signaling by a number of groups over a span of a decade (Chung *et al.*, 2009; Clarke *et al.*, 2006; Schmierer *et al.*, 2008; Vilar *et al.*, 2006; Vizán *et al.*, 2013; Wegner *et al.*, 2012; Zi *et al.*, 2011). We used the final form of the model published by Zi *et al.*, which is the most comprehensive model to date, and this model incorporates all the essential features of each module (Zi *et al.*, 2011). However, many changes were necessary to the model for the hESC case, the major additions being distinct SMAD2 and SMAD3 related reactions and SMAD7 mediated negative feedback. The following sub-sections describe the species and their reactions, model assumptions, initial conditions and rate parameters.

4.5.2.1 Activin levels, compartment volumes and cell number

The total activin level in the medium is 100 ng/ml. All concentrations in the model are converted to nM. Activin used for our cell culture is a 25.6 kDa homodimer (R & D Systems, Minneapolis, MN, USA). We use 1 mL of media containing 100 ng/ml activin per well of a 6-well tissue culture plate. Hence the concentration of activin per well is 3.91 nM. Each well of this culture plate contains ~10^6 cells; value obtained by manual cell counting using a hemocytometer. The total volume of a mammalian cell is taken as 3.3 pM (Schmierer *et al.*, 2008). For a 24 h differentiation process, there are no substantial changes in the morphology of the cells for the conditions explored here (experimental observation). This assumption may be invalid for measurements made over a typical 4-day differentiation process, but all the data used for current scenario are for a 24 h process. To calculate the nuclear and cytoplasmic compartment volumes, the pixel area of nucleus and cytoplasm were measured using image analysis. Cytoskeletal actin and nuclear DAPI stained undifferentiated hESCs were used for the measurement of cytoplasmic and nuclear areas respectively. Here, an assumption was made that the cells are cylindrical discs with small thickness in the z-direction. Therefore the nuclear to cytoplasmic area ratio is very close to the volume ratios. The average N to C ratio ($\rho_{N/C}$) for undifferentiated cells from different regions of a colony came out to be 0.43 ± 0.2. This ratio was used to obtain average nuclear and cytoplasmic volumes of the cell using the following equations:

$$\rho_{N/C} = \frac{V_{nuc}}{V_{cyt}},$$

$$V_{tot} = V_{cyt} + V_{nuc},$$

$$V_{cyt} = \frac{V_{tot}}{(\rho_{N/C} + 1)}$$

$$V_{nuc} = \frac{V_{tot}\,\rho_{N/C}}{(\rho_{N/C} + 1)} \tag{4.6-4.9}$$

Total cell number in a well is assumed constant during the 24 h study period. From the modeling results it was observed that any decrease in cell number (for example in the low PI3K condition), did not lead to any substantial increase in downstream signaling (via increased ligand availability per cell) indicating that activin levels are under saturation in our cell culture medium. In the literature, most studies are conducted at lower ligand concentrations than those explored in the hESC differentiation protocols and even at these levels the downstream SMAD levels are found to be not in the ligand-limited regime. At concentrations explored in this study (3.9 nM), the decrease in activin levels in a 24 h period is minimal, with model predicted minimum values (from different simulation runs) of 3.3 nM at 24 h. Activin in the medium is lost due to cellular uptake during LRC formation and by non-specific binding and unbinding on the cell surface, modeled in the following form:

$$V_{med}\frac{dAct_{med}}{dt} = -V_{cyt}k_{aLRC}Act_{med}R1_{surf}\,R2_{surf} - V_{med}\left(k_{onns}Act_{med} - k_{offns}Act_{ns}\right) \tag{4.10}$$

Dissociation rate of activin binding to non-specific location on the cell is fixed by the relation:

$$k_{offns} = k_{D\_ns} \times k_{onns} \tag{4.11}$$

4.5.2.2 Receptors

Dynamics of R1 on the surface is dependent on the formation of R1, internalization of R1 into the endosomes, recycling of R1 and LRC complex formation. Identical reactions can be written for R2 with R2 related rate parameters.

$$\begin{aligned} V_{cyt}\frac{dR1_{surf}}{dt} &= V_{cyt}k_{prodR1} - V_{cyt}k_iR1_{surf} + V_{cyt}k_rR1_{endo} \\ &\quad - V_{cyt}k_{aLRC}Act_{med}R1_{surf}\,R2_{surf} \end{aligned} \tag{4.12}$$

$$\begin{aligned} V_{cyt}\frac{dR2_{surf}}{dt} &= V_{cyt}k_{prodR2} - V_{cyt}k_iR2_{surf} + V_{cyt}k_rR2_{endo} \\ &\quad - V_{cyt}k_{aLRC}Act_{med}R1_{surf}\,R2_{surf} \end{aligned} \tag{4.13}$$

The dynamics of endosomal receptors (both R1 and R2) is dependent on the receptor degradation in the low pH environment of the endosomes, internalization and recycling and LRC dissociation.

$$\begin{aligned} V_{cyt}\frac{dR1_{endo}}{dt} &= -V_{cyt}k_{\deg R1}R1_{endo} + V_{cyt}k_iR1_{surf} - V_{cyt}k_rR1_{endo} \\ &\quad + V_{cyt}k_{dissLRC}LRC_{endo} \end{aligned} \tag{4.14}$$

$$V_{cyt}\frac{dR2_{endo}}{dt} = -V_{cyt}k_{\deg R2}R2_{endo} + V_{cyt}k_iR2_{surf} - V_{cyt}k_rR2_{endo} + V_{cyt}k_{dissLRC}LRC_{endo} \tag{4.15}$$

The initial conditions for these species are given as follows (by solving algebraic equations obtained after equating the right hand side to zero):

$$R1_{surf}[t=0] = \frac{k_{prodR1}\left(k_r + k_{\deg R1}\right)}{k_{\deg R1} \bullet k_i} \tag{4.16}$$

$$R2_{surf}[t=0] = \frac{k_{prodR2}\left(k_r + k_{\deg R2}\right)}{k_{\deg R2} \bullet k_i} \tag{4.17}$$

$$R1_{endo}[t=0] = \frac{k_{prodR1}}{k_{\deg R1}} \tag{4.18}$$

$$R2_{endo}[t=0] = \frac{k_{prodR2}}{k_{\deg R2}} \tag{4.19}$$

4.5.2.3 LRC levels

LRC is the catalyst for activation of SMADs in the endosomal compartment. LRC is distributed between the cell surface and endosomes. The levels of LRC on the surface are controlled by its rate of formation via complexation between activin, R1 and R2 followed by internalization into endosomes and negative feedback mediated degradation via SMAD7/Smurf2/SIK complex. The levels of endosomal LRC are controlled by internalization, degradation and dissociation. The initial condition for each of these species is zero.

$$\begin{aligned} V_{cyt}\frac{dLRC_{surf}}{dt} &= V_{cyt}k_{aLRC}Act_{med}R1_{surf}\,R2_{surf} - V_{cyt}k_{iLRC}LRC_{surf} \\ &\quad - V_{cyt}k_{lid}LRC_{surf}SMAD7_{prot} \end{aligned} \tag{4.20}$$

$$\begin{aligned} V_{cyt}\frac{dLRC_{endo}}{dt} &= V_{cyt}k_{iLRC}LRC_{surf} - V_{cyt}k_{\deg LRC}LRC_{endo} \\ &\quad - V_{dissLRC}LRC_{endo} \end{aligned} \tag{4.21}$$

4.5.2.4 Non-phospho SMADs

Intracellular SMAD4 undergoes nucleo-cytoplasmic shuttling both in the absence and presence of the ligand. Further it forms heteromeric complexes with pSMAD2 and pSMAD3. The levels of total SMAD4 in a cell are abundant and assumed to be constant and kept fixed at 10^5 molecules/cell (Clarke *et al.*, 2006).

$$\begin{aligned} V_{cyt}\frac{dSMAD4_{cyt}}{dt} &= -V_{cyt}k_{impS4}SMAD4_{cyt} + V_{nuc}k_{expS4}SMAD4_{nuc} \\ &\quad - V_{cyt}\left(k_{onSmads}SMAD4_{cyt}\,pSMAD2_{cyt} - k_{offSmads}\,pSMAD2SMAD4_{cyt}\right) \\ &\quad - V_{cyt}\left(k_{onSmads}SMAD3_{cyt}\,pSMAD2_{cyt} - k_{offSmads}\,pSMAD3SMAD4_{cyt}\right) \end{aligned} \tag{4.22}$$

$$\begin{aligned} V_{nuc}\frac{dSMAD4_{nuc}}{dt} &= V_{cyt}k_{impS4}SMAD4_{cyt} - V_{nuc}k_{expS4}SMAD4_{nuc} \\ &\quad - V_{nuc}\left(k_{onSmads}SMAD4_{nuc}\,pSMAD2_{nuc} - k_{offSmads}\,pSMAD2SMAD4_{nuc}\right) \\ &\quad - V_{nuc}\left(k_{onSmads}SMAD3_{nuc}\,pSMAD2_{nuc} - k_{offSmads}\,pSMAD3SMAD4_{nuc}\right) \end{aligned} \tag{4.23}$$

The initial conditions for cytoplasmic and nuclear SMAD4 are given by:

$$SMAD4_{cyt}[t=0] = \frac{10^5}{N_{av}\times 10^{-9}}\frac{1}{V_{cyt}}\frac{1}{\frac{k_{impS4}}{k_{\exp S4}}+1} \tag{4.24}$$

$$SMAD4_{nuc}[t=0] = SMAD4_{cyt}[t=0]\frac{V_{cyt}}{V_{nuc}}\frac{k_{impS4}}{k_{\exp S4}} \tag{4.25}$$

Intracellular SMAD2 and SMAD3 also undergo nucleo-cytoplasmic shuttling in the absence and presence of the ligand. Further, SMAD2 and SMAD3 form homomeric and heterometric complexes with themselves and with SMAD4 respectively after their phosphorylation. Various other combinations of complexes are also possible, but these have been less experimentally characterized and are found to be less abundant under current conditions (Schmierer *et al.*, 2008). Note that phosphorylation occurs in the cytoplasm since we assume the LRC complex is localized in the cytoplasm and dephosphorylation occurs in the nucleus.

$$\begin{aligned} V_{cyt}\frac{dSMAD2_{cyt}}{dt} &= -V_{cyt}k_{impS2}SMAD2_{cyt} + V_{nuc}k_{expS2}SMAD2_{nuc} \\ &\quad -V_{cyt}k_{phoS2}SMAD2_{cyt}LRC_{endo} \end{aligned} \tag{4.26}$$

$$V_{nuc}\frac{dSMAD2_{nuc}}{dt} = V_{cyt}k_{impS2}SMAD2_{cyt} - V_{nuc}k_{expS2}SMAD2_{nuc} + V_{nuc}k_{dephoS2}pSMAD2_{nuc} \tag{4.27}$$

$$\begin{aligned} V_{cyt}\frac{dSMAD3_{cyt}}{dt} &= -V_{cyt}k_{impS3}SMAD3_{cyt} + V_{nuc}k_{expS3}SMAD3_{nuc} \\ &\quad -V_{cyt}k_{phoS3}SMAD3_{cyt}LRC_{endo} \end{aligned} \tag{4.28}$$

$$V_{nuc}\frac{dSMAD3_{cyt}}{dt} = V_{cyt}k_{impS3}SMAD3_{cyt} + V_{nuc}k_{expS3}SMAD3_{nuc} + V_{nuc}k_{dephoS3}pSMAD3_{nuc} \tag{4.29}$$

The initial conditions for non-phospho SMADs are given by:

$$SMAD2_{cyt}[t=0] = \frac{10^5}{N_{av} \times 10^{-9}} \frac{1}{V_{cyt}} \frac{1}{\frac{k_{impS2}}{k_{\exp S2}} + 1} \tag{4.30}$$

$$SMAD2_{nuc}[t=0] = SMAD2_{cyt}[t=0] \frac{V_{cyt}}{V_{nuc}} \frac{k_{impS2}}{k_{\exp S2}} \tag{4.31}$$

$$SMAD3_{cyt}[t=0] = \frac{10^4}{N_{av} \times 10^{-9}} \frac{1}{V_{cyt}} \frac{1}{\frac{k_{impS3}}{k_{\exp S3}} + 1} \tag{4.32}$$

$$SMAD3_{nuc}[t=0] = SMAD3_{cyt}[t=0] \frac{V_{cyt}}{V_{nuc}} \frac{k_{impS3}}{k_{\exp S3}} \tag{4.33}$$

4.5.2.5 Monomeric phospho SMADs

Phospho-SMAD concentration changes due to phosphorylation, dephosphorylation, nuclear import and export and formation and dissociation of higher order complexes like pSMAD-SMAD4 and pSMAD-pSMAD. The initial conditions of all monomeric phospho SMADs are zero.

$$\begin{aligned} V_{cyt} \frac{dpSMAD2_{cyt}}{dt} &= V_{cyt} k_{phoS2} SMAD2_{cyt} \, LRC_{endo} - V_{cyt} k_{impS2} \, pSMAD2_{cyt} \\ &+ V_{nuc} k_{expS2} \, pSMAD2_{nuc} \\ &- V_{cyt} \left(k_{onSmads} SMAD4_{cyt} \, pSMAD2_{cyt} + k_{offSmads} \, pSMAD2SMAD4_{cyt} \right) \\ &- 2V_{cyt} \left(k_{onSmads} \, pSMAD2_{cyt} \, pSMAD2_{cyt} + k_{offSmads} \, pSMAD2 \, pSMAD2_{cyt} \right) \end{aligned} \tag{4.34}$$

$$\begin{aligned} V_{nuc}\frac{dpSMAD2_{nuc}}{dt} &= V_{cyt}k_{impS2}pSMAD2_{cyt} - V_{nuc}k_{expS2}pSMAD2_{nuc} \\ &- V_{nuc}k_{dephoS2}pSMAD2_{nuc} \\ &- V_{nuc}\left(k_{onSmads}SMAD4_{nuc}\,pSMAD2_{nuc} + k_{offSmads}pSMAD2SMAD4_{nuc}\right) \\ &- 2V_{nuc}\left(k_{onSmads}pSMAD2_{nuc}\,pSMAD2_{nuc} + k_{offSmads}pSMAD2\,pSMAD2_{nuc}\right) \end{aligned}$$

(4.35)

$$\begin{aligned} V_{cyt}\frac{dpSMAD3_{cyt}}{dt} &= V_{cyt}k_{phoS3}SMAD3_{cyt}LRC_{endo} - V_{cyt}k_{impS3}pSMAD3_{cyt} + V_{nuc}k_{expS3}pSMAD3_{nuc} \\ &- V_{cyt}\left(k_{onSmads}SMAD4_{cyt}\,pSMAD3_{cyt} + k_{offSmads}pSMAD3SMAD4_{cyt}\right) \\ &- 2V_{cyt}\left(k_{onSmads}pSMAD3_{cyt}\,pSMAD3_{cyt} + k_{offSmads}pSMAD3\,pSMAD3_{cyt}\right) \end{aligned}$$

(4.36)

$$\begin{aligned} V_{nuc}\frac{dpSMAD3_{nuc}}{dt} &= V_{cyt}k_{impS3}pSMAD3_{cyt} - V_{nuc}k_{expS3}pSMAD3_{nuc} \\ &- V_{nuc}k_{dephoS3}pSMAD3_{nuc} \\ &- V_{nuc}\left(k_{onSmads}SMAD4_{nuc}\,pSMAD3_{nuc} + k_{offSmads}pSMAD3SMAD4_{nuc}\right) \\ &- 2V_{nuc}\left(k_{onSmads}pSMAD3_{nuc}\,pSMAD3_{nuc} + k_{offSmads}pSMAD3\,pSMAD3_{nuc}\right) \end{aligned}$$

(4.37)

4.5.2.6 Phospho-SMAD complexes

Phospho-SMAD and SMAD4 complexes are imported to the nucleus (but not exported unless decomplexed) and they can form complexes both in the cytoplasm and the nucleus. The initial conditions of all the complexes are zero.

$$\begin{aligned} V_{cyt}\frac{dpSMAD2\,pSMAD2_{cyt}}{dt} &= -V_{cyt}k_{impSmads}pSMAD2\,pSMAD2_{cyt} \\ &+ V_{cyt}\left(k_{onSmads}pSMAD2_{cyt}\,pSMAD2_{cyt} - k_{offSmads}pSMAD2\,pSMAD2_{cyt}\right) \end{aligned}$$

(4.38)

$$\begin{aligned} V_{nuc}\frac{dpSMAD2\,pSMAD2_{nuc}}{dt} &= V_{cyt}k_{impSmads}\,pSMAD2\,pSMAD2_{cyt} \\ &+V_{cyt}\left(k_{onSmads}\,pSMAD2_{nuc}\;pSMAD2_{nuc}-k_{offSmads}\,pSMAD2\,pSMAD2_{nuc}\right) \end{aligned} \tag{4.39}$$

$$\begin{aligned} V_{cyt}\frac{dpSMAD3\,pSMAD3_{cyt}}{dt} &= -V_{cyt}k_{impSmads}\,pSMAD3\,pSMAD3_{cyt} \\ &+V_{cyt}\left(k_{onSmads}\,pSMAD3_{cyt}\;pSMAD3_{cyt}-k_{offSmads}\,pSMAD3\,pSMAD3_{cyt}\right) \end{aligned} \tag{4.40}$$

$$\begin{aligned} V_{nuc}\frac{dpSMAD3\,pSMAD3_{nuc}}{dt} &= V_{cyt}k_{impSmads}\,pSMAD3\,pSMAD3_{cyt} \\ &+V_{cyt}\left(k_{onSmads}\,pSMAD3_{nuc}\;pSMAD3_{nuc}-k_{offSmads}\,pSMAD3\,pSMAD3_{nuc}\right) \end{aligned} \tag{4.41}$$

$$\begin{aligned} V_{cyt}\frac{dpSMAD2SMAD4_{cyt}}{dt} &= -V_{cyt}k_{impSmads}\,pSMAD2SMAD4_{cyt} \\ &+V_{cyt}\left(k_{onSmads}\,pSMAD2_{cyt}\,SMAD4_{cyt}-k_{offSmads}\,pSMAD2SMAD4_{cyt}\right) \end{aligned} \tag{4.42}$$

$$\begin{aligned} V_{cyt}\frac{dpSMAD3SMAD4_{cyt}}{dt} &= -V_{cyt}k_{impSmads}\,pSMAD3SMAD4_{cyt} \\ &+V_{cyt}\left(k_{onSmads}\,pSMAD3_{cyt}\,SMAD4_{cyt}-k_{offSmads}\,pSMAD3SMAD4_{cyt}\right) \end{aligned} \tag{4.43}$$

$$\begin{aligned} V_{nuc}\frac{dpSMAD2SMAD4_{nuc}}{dt} &= V_{cyt}k_{impSmads}\,pSMAD2SMAD4_{cyt} \\ &+V_{cyt}\left(k_{onSmads}\,pSMAD2_{nuc}\,SMAD4_{nuc}-k_{offSmads}\,pSMAD2SMAD4_{nuc}\right) \end{aligned} \tag{4.44}$$

$$V_{nuc}\frac{dpSMAD3SMAD4_{nuc}}{dt}=V_{cyt}k_{impSmads}\,pSMAD3SMAD4_{cyt}$$
$$+V_{cyt}\left(k_{onSmads}\,pSMAD3_{nuc}\,SMAD4_{nuc}-k_{offSmads}\,pSMAD3SMAD4_{nuc}\right)$$

(4.45)

4.5.2.7 Negative feedback molecule SMAD7, transcription and translation

New equations were added to capture SMAD7 transcription mediated by the binding of pSMAD2,3-SMAD4 complexes to the promoter region (Nicklas and Saiz, 2013). The rate of transcription is modeled as a function of the amount of SMAD7 mRNA released per molecule of pSMAD-SMAD4 complex and the efficiency of this process is captured by affinity constants multiplied to this rate. The degradation of SMAD7 mRNA transcripts is modeled as a first order reaction. SMAD7 protein is controlled by first order translation and degradation and loss due to participation in surface LRC degradation. Both SMAD7 species have initial concentration of zero.

$$V_{nuc}\frac{dSMAD7_{mRNA}}{dt}=-V_{nuv}k_{\deg S7mrna}SMAD7_{mRNA}$$
$$+V_{nuc}\frac{k_{S7prodpS2S4}k_{S7affpS2S4}\,pSMAD2SMAD4+k_{S7prodpS3S4}k_{S7affpS3S4}\,pSMAD3SMAD4}{1+k_{S7affpS2S4}\,pSMAD2SMAD4+k_{S7affpS3S4}\,pSMAD3SMAD4}$$

(4.46)

$$V_{cyt}\frac{dSMAD7_{prot}}{dt}=-V_{cyt}k_{\deg S7prot}SMAD7_{prot}+V_{cyt}k_{S7protprod}SMAD7_{mRNA}$$
$$-V_{cyt}k_{lid}LRC_{surf}SMAD7_{prot}$$

(4.47)

4.5.2.8 Activin non-specific binding and intracellular fate of Activin

Activin in the medium may be lost by non-specific binding to cell surface as described in Equation 4.10. Further, the intracellular activin released by LRC dissociation may be degraded inside the cell. The equations for these species are given below:

$$V_{med}\frac{dAct_{ns}}{dt}=V_{med}\left(k_{onns}Act_{med}-k_{offns}Act_{ns}\right) \tag{4.48}$$

$$V_{cyt}\frac{dAct_{endo}}{dt}=V_{cyt}k_{dissLRC}LRC_{endo}-V_{cyt}\left(k_{\deg Act}Act_{endo}\right) \tag{4.49}$$

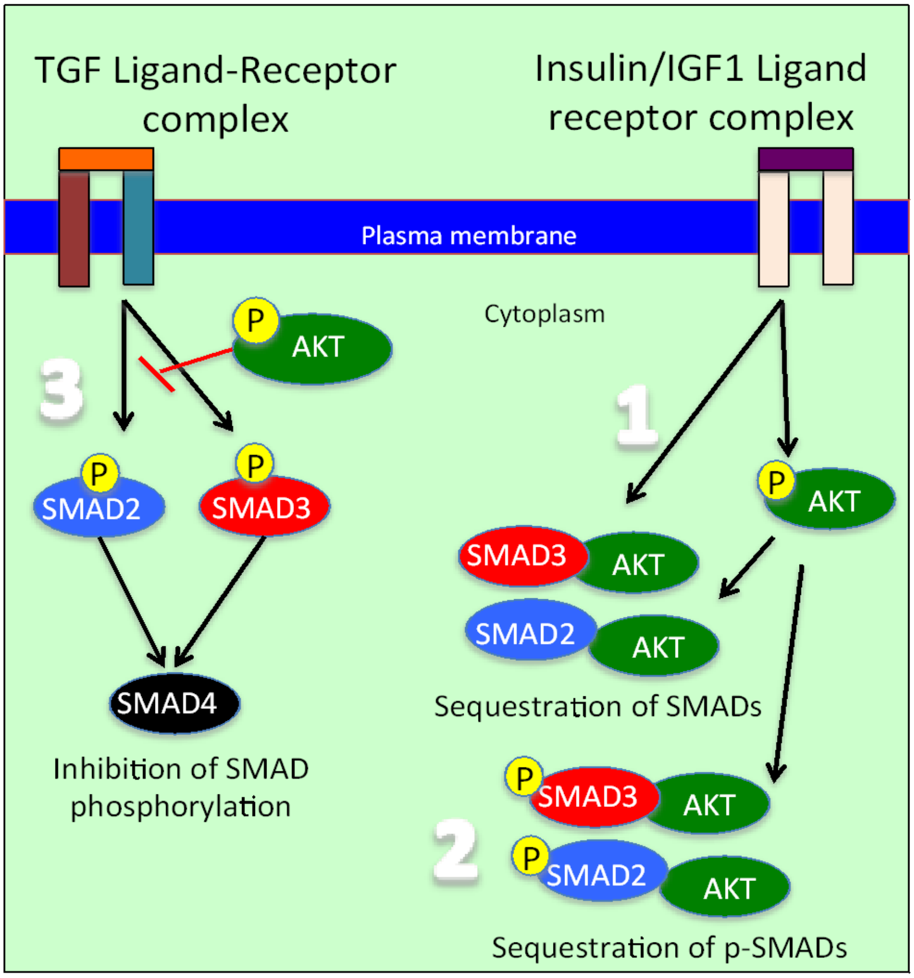


Figure 4.6 Schematic representing the three modes of AKT mediated crosstalk on SMADs

4.5.3 Modeling crosstalk between AKT and SMADs

Several experimental studies have shown that AKT may directly or indirectly influence the levels of SMAD signaling. These mechanisms are summarized in Figure 4.6. The crosstalk mechanisms involve sequestration of different SMADs in the cytoplasm or inhibition of SMAD activation.

4.5.3.1 Crosstalk 1: Sequestration of non-phospho SMADs by AKT

Earliest experiments by Remy *et al.* showed that AKT can physically interact with SMAD3 and form a complex with SMAD3 leading to its sequestration on the cell membrane in a human hepatoma cell line (Hep3B) (Remy *et al.*, 2004). As a result, some pool of SMAD3 is lost and not available to the LRC. In their analysis, this lead to indirect effects like decreases in SMAD3 phosphorylation, nuclear translocation and SMAD3 induced transcription. Further, they reported that kinase activity of AKT was not critical for existence of this crosstalk and AKT was not influenced by SMAD3. However, this crosstalk is minimal once SMAD3 gets phosphorylated, perhaps because SMAD3 forms non-AKT complexes and thus, SMAD nucleo-cytoplasmic shuttling becomes preferential to sequestration. Similar observations were made by Conery *et al.* (Conery *et al.*, 2004). They also reported that the effect was specific to SMAD3 and not observed for SMAD2. Our DBN inference also showed that AKT might influence SMAD3 levels but not SMAD2. However, report by Song *et al.*, found that AKT may bind to SMAD2 but its influence on SMAD2 phosphorylation is not clear (Song *et al.*, 2006). In a neutroblastoma cell line, PI3K/AKT pathway inhibition was found to increase SMAD2 phosphorylation and nuclear translocation (Qiao *et al.*, 2006). Therefore, in this crosstalk mechanism, both SMAD2 and SMAD3 were allowed to form a pool of non-available SMADs ($SMAD-c$). These reactions occur in the cytoplasm and are modeled as:

$$SMAD2 \leftrightarrow SMAD2-c \; k_{cS2}, k_{dcS2}$$

$$SMAD3 \leftrightarrow SMAD3-c \; k_{cS3}, k_{dcS3}$$

The rate coefficients represent the rate of complexation (k_c) and decomplexation (k_{dc}) of SMAD2 and SMAD3 with AKT and these have to be estimated from the experimental data. In this interaction, a constant level of AKT is assumed throughout the 24 h time period. But these reaction rate coefficients are allowed to vary between the high and low PI3K conditions to model different levels of AKT. The initial condition for each SMAD now has to account for the new pool of non-available SMADs. The initial conditions of free cytoplasmic SMADs are now calculated as:

$$[SMAD2]_{cyt}(t=0) = \frac{10^5}{N_{av} \times 10^{-9}} \frac{1}{V_{cyt}} \frac{1}{\left[\frac{k_{cS2}}{k_{dcS2}} + \frac{k_{impS2}}{k_{\exp S2}} + 1\right]} \tag{4.50}$$

$$[SMAD3]_{cyt}(t=0) = \frac{10^4}{N_{av} \times 10^{-9}} \frac{1}{V_{cyt}} \frac{1}{\left[\frac{k_{cS3}}{k_{dcS3}} + \frac{k_{impS3}}{k_{\exp S3}} + 1\right]} \tag{4.51}$$

The nuclear concentrations are obtained by using these values in Equation 4.31 and 4.33.

4.5.3.2 Crosstalk 2: Sequestration of phospho-SMADs by AKT

The study by Song *et al.* also showed that phospho SMADs may be sequestered by AKT in the cytoplasm (Song *et al.*, 2006). However, they did not analyze if reduction in SMAD nuclear accumulation is specific to SMAD or phospho SMAD sequestration. Since the possibility of p-SMAD sequestration exists, we modeled these reactions to check how well this model can capture the experimental data. The reactions are similar to Crosstalk 1, with non-phospho SMAD replaced by phospho SMAD.

$$pSMAD2 \leftrightarrow pSMAD2-c \;\; k_{cpS2}, k_{dcpS2}$$

$$pSMAD3 \leftrightarrow pSMAD3-c \;\; k_{cpS3}, k_{dcpS3}$$

The initial conditions of SMADs do not change with this crosstalk, however this pool of complexed p-SMADs is to be included in Equations (4.56-4.57) while calculating the model output that is to be compared to experimental data.

4.5.3.3 Crosstalk 3: AKT inhibition of SMAD phosphorylation

The most dominant model proposed by Song *et al.* was the direct inhibition of phosphorylation of SMAD by the PI3K/AKT pathway (Song *et al.*, 2003; Song *et al.*, 2006). In this model, a molecule downstream of AKT, called mTORC1, directly inhibits AKT phosphorylation. Since PI3K inhibition leads to reduction in both p-AKT and mTORC1 signaling, this effect could still exist in the PI3K perturbation explored in the experimental data. In order to model this effect, we chose a simplified scenario where the phosphorylation rates of SMAD2 and SMAD3 in the low PI3K condition are chosen equal to a factor multiplied by corresponding rates in the high PI3K condition, with the factor being greater than or equal to 1.

This ensures that low PI3K has a phosphorylation rate greater than or equal to the high PI3K case.

$$k_{phoS2\_lowPI3K} = f_{S2} \times k_{phoS2\_highPI3K} \quad \log_{10} f_{S2} \in [0,2] \tag{4.52}$$

$$k_{phoS3\_lowPI3K} = f_{S3} \times k_{phoS3\_highPI3K} \quad \log_{10} f_{S3} \in [0,2] \tag{4.53}$$

4.5.4 Why model crosstalk?

Each experimental study presented in the previous section focused on one model of crosstalk between AKT and SMAD and did not explicitly test the other models. The studies by Remy *et al.* and Conery *et al.* did not consider the influence of AKT inhibition of SMAD phosphorylation. The study by Song *et al.* could not verify if sequestration of SMADs could contribute in any way to their observation, and if their observations could be explained only by AKT mediated inhibition of SMAD phosphorylation. There is also a possibility that combinations of these mechanisms may exist. The mechanism may also be dependent on the cell type. In addition, completely removing SMAD sequestration effect is difficult since the actual biophysics of complex formation between AKT and SMAD is not well-known (Danielpour and Song, 2006; Song *et al.*, 2006). Furthermore, it is unclear if SMAD2 and SMAD3 show the same type of crosstalk behavior. Because of these reasons, it is challenging for a pure experimental study to tease out the differences. It is non-intuitive to decipher specific mechanism of crosstalk purely from the experimental dynamics. Hence using a mathematical modeling approach we wanted to verify if specific mechanism of crosstalk will result in unique dynamics of any signaling molecules. This can be tested using a detailed mathematical model of the entire SMAD pathway with each type of crosstalk represented explicitly. Once we have the complete model, we can test the performance of the model with crosstalk by fitting it to explain experimental data (high and low PI3K). This process may allow us to test individual mechanisms and eliminate those that do not adequately capture the experimental dynamics. The modeled mechanisms, which capture the experimental dynamics, can be further verified experimentally, for example by measurement of species whose dynamics/levels diverge between the competing mechanisms or performing additional perturbations or even checking the biological feasibility of the model predictions. We chose this particular route to test the three types of crosstalk hypotheses in the hESC system (which additionally shows distinct dynamics of p-SMAD2 and p-SMAD3). The parameter estimation process used to check the validity of the three hypotheses will be discussed next.

4.5.5 Parameter estimation

Values of the free rate parameters were inferred using a Bayesian parallel tempering (PT) approach (Brown and Sethna, 2003; Malkin *et al.*, 2015; Swigon, 2012), which utilizes Markov Chain Monte Carlo (MCMC) methods to sample the Bayesian posterior distribution, the probability of parameter set p given data y, given by the Bayes' formula,

$$P(p/y) = \frac{L(y/p)\theta(p)}{\int L(y/p)\theta(p)} \tag{4.54}$$

where, $L(y/p)$ is the likelihood of observing y for a model with parameters p, $\theta(p)$ is the prior distribution, and $\int L(y/p)\theta(p)$ is the normalizing constant. Additional sampling efficiency is gained by running multiple parallel chains evolving at different temperatures. Higher temperature increases the likelihood of acceptance of proposed steps. This allows the high temperature chains to move more freely through the parameter space, avoiding getting stuck in local minima. This results in more efficient exploration of parameter space, a method we have applied extensively in parameter estimation of complex non-linear models (Mochan *et al.*, 2014; Slezak *et al.*, 2010; Song *et al.*, 2012). This results in the creation of parameter ensembles, where each parameter is represented by a posterior distribution, rather than a single value.

4.5.5.1 Bayesian priors

For each parameter to be fitted, a uniform prior was used, with a suitably large range so as to encompass all reasonable parameter values. For the parameters with known literature values but which had to be fitted for the new system, the center of the interval was fixed at the nominal value in the literature.

4.5.5.2 Parameter set fitness

Fitness (log likelihood) of candidate parameter sets was determined by the percentage difference between the model simulations and experimental data, as determined by the sum of squared residuals cost function:

$$Fitness = \sum_{i,j,k} \frac{1}{2\sigma^2_{i,j,k}} \left(\frac{y_{i,j,k} - \hat{y}_{i,j,k}}{\hat{y}_{i,j,k}} \right)^2 \tag{4.55}$$

where, $y_{i,j,k}$ is the output for a simulation with single set of parameters, $\hat{y}_{i,j,k}$ is the experimental mean, $\sigma_{i,j,k}$ is the standard deviation allowed for the experimental mean at time point i, for observable j under treatment condition k. No additional penalties were added to the fitness function. Best-fit parameter set is the one with the least fitness value. But, we are interested in the parameter ensembles that can explain the experimental data.

4.5.5.3 Parallel tempering

To efficiently sample the posterior distribution, four separate Markov chains were run, initiated with parameter values randomly selected from the supplied prior distributions that met a maximum energy criterion. Each chain was initiated with a temperature and step size parameter, which controlled the chain's ability to fully explore the parameters space. Chains were allowed to swap from a higher temperature to a lower temperature every 25 steps to allow for local sampling of newly found local minima. Step size and temperature parameters dynamically changed every 6000 and 2000 steps respectively to attempt to reach an ideal step acceptance rate of 23% (Roberts *et al.*, 1997), and swap rates of 15%-30%. Once these targets were reached, the temperature schedule and step sizes were fixed. Parameter sets were saved every 25 steps. Full exploration of parameter space was confirmed by examining, for each parameter, the frequency histogram of its full marginal posterior distribution, confirming that it spanned the prior domain.

We measured convergence and chain stationarity using the Gelman-Rubin criteria (Brooks and Gelman, 1998; Gelman and Rubin, 1992). All parameters had converged with a potential scale reduction factor (PSRF) < 1.1 following 200,000 (x25) MCMC steps. Another 100,000 steps were taken to build a posterior distribution for each parameter that would be used for all model analysis and simulation. This ensured that all samples from the burn-in time for each chain were discarded, and only samples from the correct stationary distribution were used. The ensemble of all parameter sets from the lowest chain comprised the computed ensemble (posterior distribution). The ensemble process directly gives the uncertainty associated with each parameter and also contains the covariance information between the parameters in the high dimensional space.

4.5.5.4 Preprocessing of experimental data

For kinetic analysis, only p-SMAD2 and p-SMAD3 time series were used. In addition, we measured the levels of negative feedback molecule SMAD7 mRNA using qRT-PCR for time points of 0, 1, 6, 18 and 24 h. The data were converted to fold change over time 0. Therefore, we have three output levels measured at 22 non-zero time points (9 p-SMAD2 + 9 p-SMAD3 + 4 SMAD7) across three repeats of 2 conditions. Due to the high variability between repeats, we treated each repeat separately during the parameter fitting process. However, we treated each pair of condition (high and low PI3K) from the same passage together for the parameter fitting processes. The results from a representative repeat are presented in this chapter and it was seen that the same conclusions could be drawn when the process was repeated for the other two repeats. To summarize, for each fitting process we have 2 (conditions) x 22 (time points) = 44 number of experimental data points.

4.5.5.5 Selection of model outputs to be fitted to experimental data

The following equations describe the model outputs that were fitted to the experimental data for high and low PI3K.

$$\begin{aligned}[pSMAD2]_{\mathrm{mod}} &= [pSMAD2]_{cyt}V_{cyt} + 2[pSMAD2-pSMAD2]_{cyt}V_{cyt} \\ &\quad + [pSMAD2-SMAD4]_{cyt}V_{cyt} + [pSMAD2]_{nuc}V_{nuc} \\ &\quad + 2[pSMAD2-pSMAD2]_{nuc}V_{nuc} + [pSMAD2-SMAD4]_{nuc}V_{nuc}\end{aligned}$$

$$\begin{aligned}[pSMAD3]_{\mathrm{mod}} &= [pSMAD3]_{cyt}V_{cyt} + 2[pSMAD3-pSMAD3]_{cyt}V_{cyt} \\ &\quad + [pSMAD3-SMAD4]_{cyt}V_{cyt} + [pSMAD3]_{nuc}V_{nuc} \\ &\quad + 2[pSMAD3-pSMAD3]_{nuc}V_{nuc} + [pSMAD3-SMAD4]_{nuc}V_{nuc}\end{aligned}$$

$$[SMAD7]_{\mathrm{mod}} = [SMAD7mRNA] \qquad (4.56\text{-}4.58)$$

Each of these model outputs was multipled by a scaling factor. The scaling factor is a function of the specific detection molecules used in the experimental assay for each species of SMADs and should not change between the high and low PI3K condition. Therefore, each scaling factor is kept constant between the high and low PI3K condition but separately chosen for each species. The levels of p-SMADs are not zero in the experimental measurements at time zero (which is the no activin stimulation condition). This is a complex function of the background fluorescence, minimum detection limit and stimulation from growth factors present in the culture medium (that are not explicitly added). From our observations, the time zero fluorescence values were very low as compared to the values after stimulation. Hence, the experimental time zero values are taken as an offset for the model output at all time points and this also ensures that proper fold changes can be taken. Otherwise, fold change is meaningless since the model outputs for phospho SMADs and SMAD7 at time zero are zero.

4.5.5.6 Selection of key parameters

In the fitting process, a trial run was first conducted to identify the intervals of the model parameters that could capture important features of the experimental data (Section 4.3.1). This step was necessary since many features of the experimental dynamics was different from that reported in other cell systems (see description presented in Section 4.3.1). Once the correct ranges were identified, a GSA was conduced within this interval for total p-SMAD2, p-SMAD3 and SMAD7 mRNA to identify the most sensitive parameters that control these outputs. The goal of this approach is to identify the most sensitive parameters in the right parameter intervals and only include these sensitive parameters in the main parameter estimation process.

Figure 4.7 shows the sensitivity indices for first few sensitive parameters controlling p-SMAD2 and p-SMAD3 in the early and late time points. Same procedure is followed as described in Chapter 2. Here, 10^5 random parameter set samples were selected within the intervals identified by the above step, the ODE model with/without crosstalk was simulated and the dynamic profiles of total pSMAD2, total pSMAD3 and SMAD7 mRNA were calculated. Then a RS-HDMR based GSA was conducted at 9 time points within a 24 h period and for the area under the total pSMAD curve (AUC) over 24 h. It was seen that first order Sobol' indices amounted to ~70% of the variance on average and important second order indices were combinations of these first order indices. The entire 2nd order RS-HDMR explained close to 95% of variance in the outputs. As seen from Figure 4.7A, the most sensitive parameters for pSMAD2 and pSMAD3 were similar with phosphorylation and de-phosphorylation dominating the variance, but these parameters were more important in the early and late time points respectively. This was followed by LRC dissociation rate (important in the mid time point region) and the production of R2 receptor (important early on). Similar parameters were important when considering the AUC (Figure 4.7B). Additionally, some other secondary parameters included the receptor R2 degradation and SMAD7 mRNA production. Between pSMAD2 and pSMAD3, there were only minor differences. Similar results were obtained for GSA on SMAD7 mRNA (Figure C.6). Based on the results of this section, the parameters in Table 4.1 were kept fixed at the nominal values and the parameters in Table 4.2 were estimated during parameter estimation. Fixed parameters mainly included the import/export rates, the complex formation/dissociation rates etc. These rate constants are faster compared to the slow rates of phosphorylation and dephosphorylation (rate limiting). The fitted parameters additionally included the scaling parameters for comparison to the experimental data. The fitted parameters were either allowed the flexibility to be different between the high and low PI3K case or they were kept identical between high and low PI3K case based on the context. Additional parameters were added for each crosstalk scenario. These details are discussed in the results section.

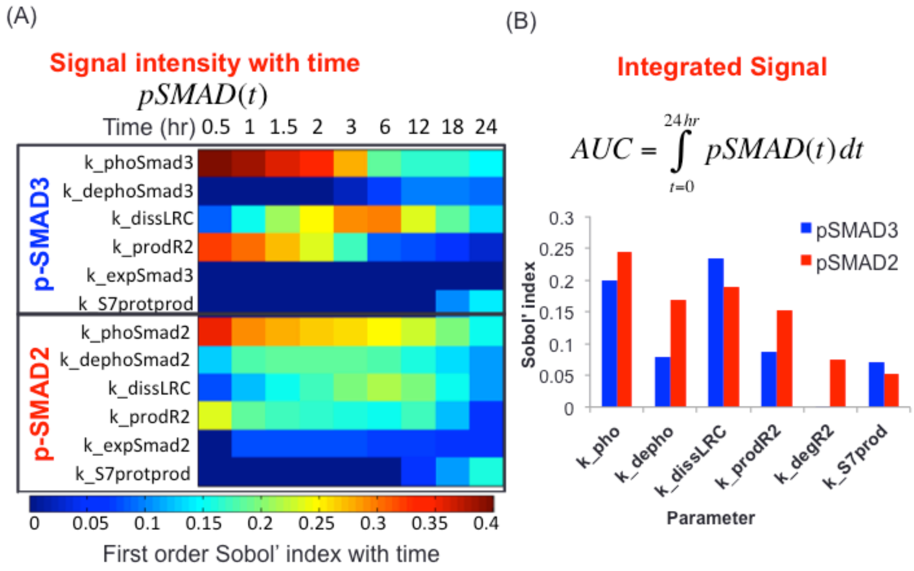


Figure 4.7 Results from GSA for pSMAD2 and pSMAD3 levels over time (A) and integrated levels (B).

Table 4.1 Fixed Parameters

Index	Parameter	Reaction step	Value	Unit	Reference
1	k_i	Receptor internalization	0.333	min^{-1}	(Vilar *et al.*, 2006)
2	k_r	Receptor recycling	0.0333	min^{-1}	(Vilar *et al.*, 2006)
3	k_{prodR1}	Receptor production	0.0137	$\text{nM}^{-1}\text{min}^{-1}$	(Zi and Klipp, 2007)
4	$k_{\deg R1}$	Receptor degradation	0.00256	min^{-1}	(Zi and Klipp, 2007)
5	$k_{\deg LRC}$	LRC degradation	0.00256	min^{-1}	(Zi and Klipp, 2007)
6	$k_{\deg Act}$	Ligand degradation	0.00256	min^{-1}	(Kaminska *et al.*, 2005)
7	k_{aLRC}	LRC complex formation	117.897	$\text{nM}^{-2}\text{min}^{-1}$	(Zi *et al.*, 2011)
8,9	k_{impS2}, k_{impS3}	Nuclear import (SMAD2,3)	0.156	min^{-1}	(Schmierer *et al.*, 2008) for Smad2

Table 4.1 (continued)

10,11	$k_{\exp S2}$, $k_{\exp S3}$	Nuclear export (SMAD2,3)	0.739	min^{-1}	(Schmierer *et al.*, 2008) for Smad2
12	k_{impS4}	Nuclear import (SMAD4)	0.156	min^{-1}	(Schmierer *et al.*, 2008)
13	$k_{\exp S4}$	Nuclear export (SMAD4)	0.355	min^{-1}	(Schmierer *et al.*, 2008)
14	$k_{impSmads}$	Nuclear import (complexes)	0.889	min^{-1}	(Schmierer *et al.*, 2008)
15	$k_{onSmads}$	Smad complex formation	0.1985	$\text{nM}^{-1}\text{min}^{-1}$	(Zi *et al.*, 2011)
16	$k_{offSmads}$	Smad complex dissociation	1	min^{-1}	(Schmierer *et al.*, 2008)
17	k_{onns}	Non-specific Ligand binding	0.0505	min^{-1}	(Zi *et al.*, 2011)
18	k_{D_ns}	Non-specific Ligand dissociation	40.2257	-	(Zi *et al.*, 2011)
19	$k_{S7affpS2S4}$	Chromosome affinity (complex pS2S4)	0.001	nM^{-1}	(Nicklas and Saiz, 2013)
20	$k_{S7affpS3S4}$	Chromosome affinity (complex pS2S4)	0.001	nM^{-1}	(Nicklas and Saiz, 2013)
21	k_{dcS2}	De-Sequestration (SMAD2)	1	min^{-1}	Fixed after first trial of parameter estimation
22	k_{dcS3}	De-Sequestration (SMAD3)	1	min^{-1}	Fixed after first trial of parameter estimation
23	k_{dcpS2}	De-Sequestration (pSMAD2)	1	min^{-1}	Fixed after first trial of parameter estimation
24	k_{dcpS3}	De-Sequestration (pSMAD3)	1	min^{-1}	Fixed after first trial of parameter estimation

Table 4.2 Fitted Parameters

Index	Parameter	Reaction step	Unit	Explored Range in $\log_{10}$ scale
1	k_{phoS2}	phosphorylation	$nM^{-1}min^{-1}$	[-4,2]
2	$k_{dephoS2}$	de-phosphorylation	min^{-1}	[-4,2]
3	k_{phoS3}	phosphorylation	$nM^{-1}min^{-1}$	[-4,2]
4	$k_{dephoS3}$	de-phosphorylation	min^{-1}	[-4,2]
5	k_{prodR2}	Receptor production	$nM^{-1}min^{-1}$	[-4,1]
6	$k_{\deg R2}$	Receptor degradation	min^{-1}	[-4,1]
7	k_{lid}	Negative feedback strength	$nM^{-1}min^{-1}$	[-4,1]
8	$k_{S7prodpS2S4}$	SMAD7 Transcription by pS2S4	$nM\,min^{-1}$	[0,4]
9	$k_{S7prodpS3S4}$	SMAD7 Transcription by pS3S4	$nM\,min^{-1}$	[0,4]
10	$k_{\deg S7mrna}$	SMAD7 mRNA degradation	min^{-1}	[-4,1]
11	$k_{\deg S7prot}$	SMAD7 protein degradation	min^{-1}	[-4,1]
12	k_{S7prod_prot}	SMAD7 protein production	$nM^{-1}min^{-1}$	[-4,1]
13	$k_{dissLRC}$	LRC dissociation	min^{-1}	[-4,4]
14	k_{iLRC}	LRC internalization	min^{-1}	[-4,4]
15	$Scale_{pS2}$	Scaling parameter	-	[-6,6]
16	$Scale_{pS3}$	Scaling parameter	-	[-6,6]

Table 4.2 (continued)

17	$Scale_{S7mRNA}$	Scaling parameter	-	[-6,6]
18	k_{cS2}	Sequestration (SMAD2)	min^{-1}	[-4,4]
19	k_{cS3}	Sequestration (SMAD3)	min^{-1}	[-4,4]
20	k_{cpS2}	Sequestration (pSMAD2)	min^{-1}	[-4,4]
21	k_{cpS3}	Sequestration (pSMAD3)	min^{-1}	[-4,4]

4.5.5.7 Refined model fitting

The selected parameters are estimated using the PT approach for each experimental repeat separately. During each fitting process, the model is fitted to the high and low PI3K data simultaneously and it is ensured that each pair of data is from the same cell culture passage. This eliminates passage-to-passage variability. Each fitting process results in a parametric ensemble that can be used to distinguish between the high and low PI3K conditions as well as distinguish between the three crosstalk mechanisms. Best fit parameter sets from each crosstalk mechanism are compared using the Akaike information criterion (AIC) (Akaike, 1998):

$$AIC_i = 2\ln(L) + 2N_{free} \tag{4.59}$$

AIC weighs model fit (log-likelihood from Equation 4.55) against model complexity, for example the number of free parameters (N_{free}). AIC values are mainly used for comparison of competing models, which can be formally done with the AIC weights:

$$AIC_{weight(i)} = \frac{e^{-0.5(AIC_i - AIC_{\min})}}{\sum_j e^{-0.5(AIC_j - AIC_{\min})}} \tag{4.60}$$

Here, $AIC_{\min}$ represents the minimum AIC value. Among a given set of competing models, the models with higher AIC weights are the best among the set.

4.6 RESULTS (KINETIC MODELING)

The following sections present the results for one experimental repeat. The goal of the fitting process is to evaluate the kinetics of the TGF-β/SMAD pathway in hESCs and the kinetics of three crosstalk mechanisms between AKT and SMAD. The parametric ensembles after convergence of MCMC chains are plotted against the experimental data for the two conditions. Then, the marginal distributions of the important fitted parameters are compared across the two conditions to evaluate the predictions made by each crosstalk model. Fitness of the best-fit parameter set and the number of parameters are given in Table 4.3.

4.6.1 Evaluating crosstalk 1

Figure 4.8A shows the ensemble outputs for each of the three species for high and low PI3K simulated using the model with non-phospho SMAD sequestration process. Here, the phosphorylation rate is kept the same between the high and low PI3K case to remove crosstalk 3 and simulate only crosstalk 1. In these plots, the lightly shaded region represents 5th-95th percentile trajectories of the simulated output, dark shaded region represents 25th-75th percentile, intermediate dark line shows the median output and the small circles represent the experimental data from one repeat. The fitted model faithfully recreates most of the important features in high PI3K condition. For example, the overshoot behavior in p-SMAD2, the first order increase of p-SMAD3 and the overshoot behavior of SMAD7 mRNA. For low PI3K condition, we obtain mixed results. Good fits are obtained for p-SMAD2 with increased phospho protein levels in the early time points. For p-SMAD3, the fits in the early time points (until 6 h) are good, although the variability in the data is high early on. The model predicts the rapid rise in p-SMAD3 in the low PI3K condition. After 6 h, the model predicts a saturation behavior with the levels equal to the high PI3K case, but the experimental data shows high levels at 12 and 24 h. The predicted 6 h SMAD7 mRNA levels in low PI3K are lower than the experimental value.

Figure 4.8B presents the posterior distributions of the parameters (box and whisker plot on log scale) with high and low PI3K conditions grouped together. Overall, the distributions of many parameters for the two conditions are close together, even though this was not explicitly imposed during the fitting process (except phosphorylation rate). Parameter distributions were compared using a two-sample Kolmogorov-Smirnov test. The de-phosphorylation rates for both molecules are unchanged between the two conditions. Comparing SMAD2 and SMAD3, the phosphorylation rates for both molecules are similar between the two conditions. However, the de-phosphorylation rate is very low for p-SMAD3 compared to p-SMAD2, possibly giving rise to the first order behavior. Receptor production and degradation are also similar between the conditions. Among the distributions that are different, SMAD7 shows increased degradation of protein but decreased degradation of mRNA in low PI3K. Therefore, the protein levels are lower in low PI3K condition (possibly due to reduced protein translation under low p-AKT). However this is compensated by increased negative feedback parameter leading to the same overall strength of feedback on p-SMAD2 and p-SMAD3. Interestingly, the comparison of crosstalk parameters showed that the complexation rate for SMAD2 is higher in low PI3K condition, which is opposite to the expectation. SMAD3 complexation rate is lower in the low PI3K condition as expected.

Under close inspection, it was found that the increase in p-SMAD2 in low PI3K is mainly due to increased levels of LRC in the endosomes (the signaling complex) (see Figure C.6A). In all the simulations of both conditions, the LRC peaks at the same time as p-SMAD2 and reduced dissociation of LRC in low PI3K increases its levels. This in turn increases p-SMAD2 and p-SMAD3. But this alone is not enough for pSMAD3 and reduced complexation also contributes to the increase. Overall, the correlations between the parameters are low in both conditions, with the strongest being the LRC dissociation rate and the phosphorylation rates of both SMADs (see Figure C.7). Thus, this mechanism does capture major features of the experimental data, but does not conform to the hypothesized mechanism for SMAD2.

Table 4.3 Comparison of best parameter outputs for each crosstalk mechanism*

Crosstalk #	# of fitted parameters (N_{free})	Best energy value (L)	AIC = 2 N_{free} +2ln$L$	AIC$_{\text{weights}}$	Is the mechanism biologically feasible?
1	36	42.64	79.51	0.1085	No, predicts increased SMAD2 complexation under low PI3K
2	36	99.13	81.19	0.0468	No, poor fits to p-SMAD2 and p-SMAD3 dynamics under low PI3K
3	34	41.23	75.44	0.8300	Yes
Combined 1+3	38	42.80	83.51	0.0147	Yes

* Note: Total number of experimental data points = 44 (from three outputs for two conditions)

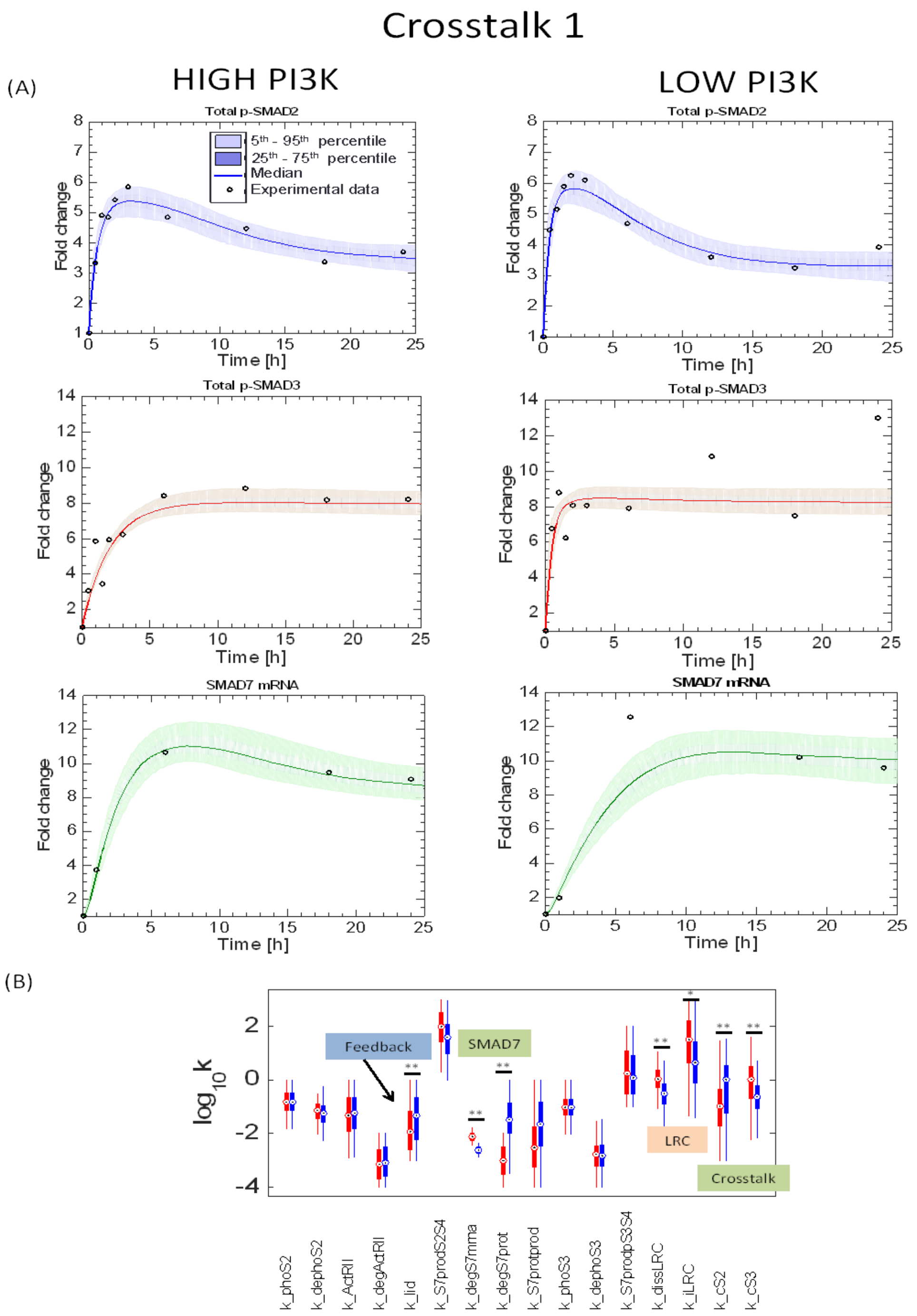


Figure 4.8 Ensemble modeling outputs for mechanism 1.

(A) Simulated model fits with the experimental training data. 10^5 MCMC samples are used for plotting. (B) Posterior distributions of the fitted parameters. Values for the mean (circle), 25th-75th percentile (end points of the boxes) and 2.5th to 97.5th percentiles as endpoints of the whiskers are shown. Parameter distributions were compared using a two sample Kolmogorov-Smirnov test, *p<0.05, **p<0.01.

4.6.2 Evaluating crosstalk 2

Figure 4.9A presents the simulated model outputs for mechanism 2. Here, it is observed that the model fits are satisfactory for the high PI3K case while very poor fits are obtained for the low PI3K case for p-SMAD2 and p-SMAD3. The low PI3K case does not capture the early rise in pSMAD3 and a delay in pSMAD3 is seen similar to the high PI3K case. Therefore, this mechanism, where AKT sequesters phospho protein in the cytoplasm, is unable to explain the increased pSMAD2 levels in low PI3K and faster increase in pSMAD3 in the low PI3K condition. The main reason for this is accumulation of pSMAD if it complexes with AKT (Figure 4.10). This reduces the amount of SMAD available for de-phosphorylation. Hence, it becomes difficult for the model to capture the increase in pSMAD when AKT is low if this mechanism is the only one leading to the increase in pSMAD.

Posterior parameter distributions shown in Figure 4.9B indicate that the two models differ mainly in the SMAD7 related parameters. SMAD7 mRNA degradation is increased and protein degradation is decreased in the low PI3K case which enables the model to capture the 1 h level of SMAD7 mRNA in low PI3K condition, but the model follows the high PI3K condition at the other time points. The LRC dissociation rate is increased in the low PI3K condition leading to lower levels of active LRC complex (Figure C.6B). Overall, this mechanism captures the SMAD7 dynamics better than previous mechanism but not the other phosphor protein dynamics.

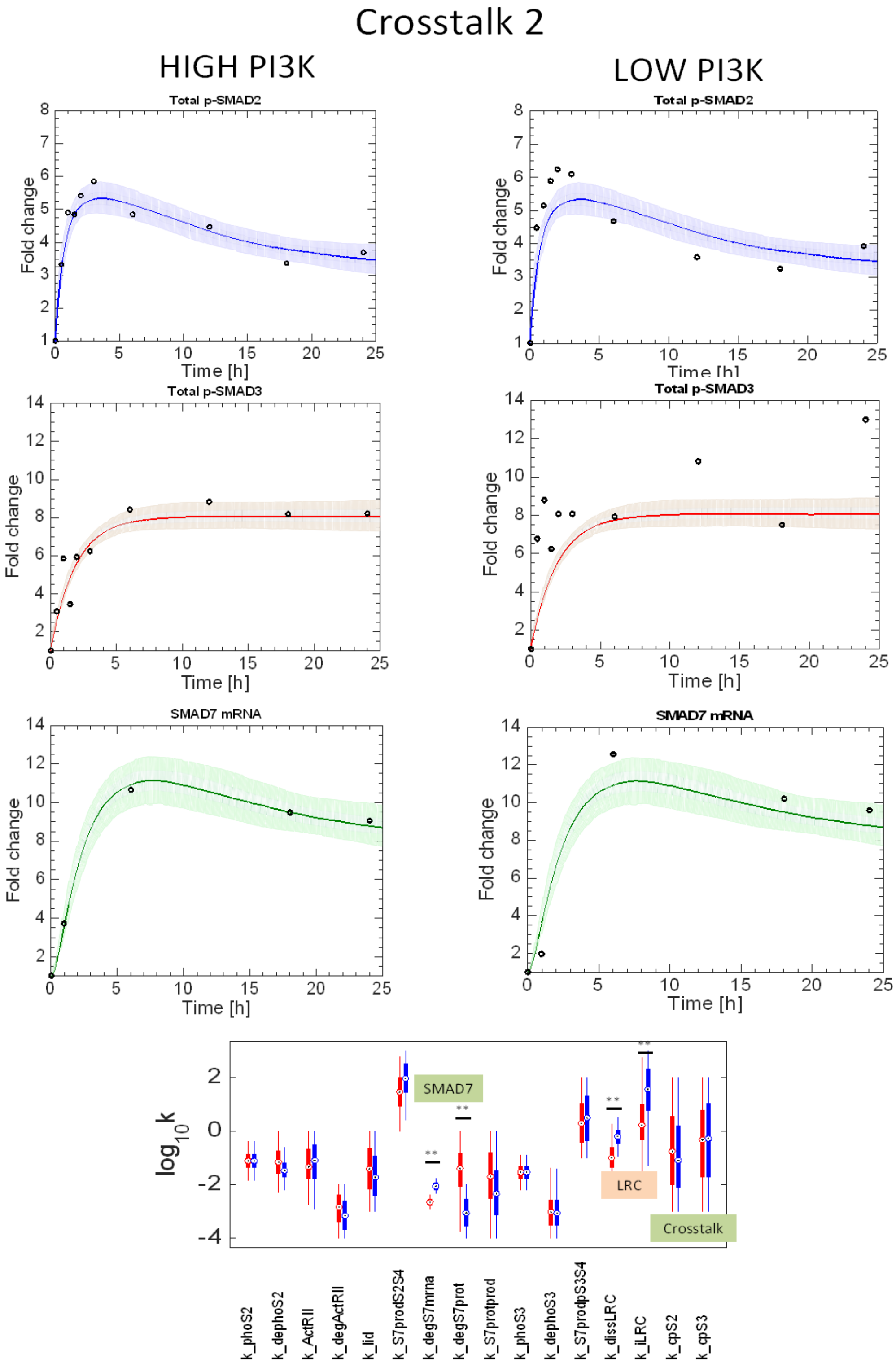


Figure 4.9 Ensemble modeling outputs for mechanism 2.

(A) Simulated model fits with the experimental training data. (B) Posterior distributions of the fitted parameters. Details identical to Figure 4.8.

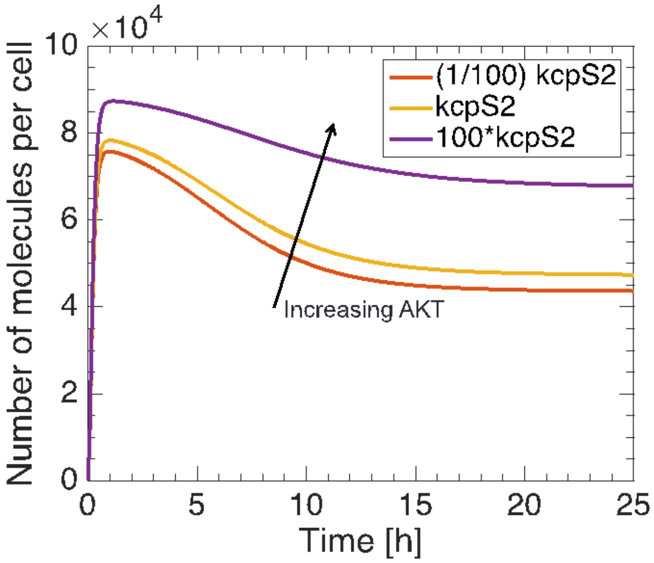


Figure 4.10 Changes in total pSMAD2 levels with increasing complexation rate with AKT.

Three levels of k_{cpS2} are considered here, best parameter from high PI3K case followed by 100-fold reduction and increase over this level. Increasing complexation rate (modeling increasing AKT) is leading to increase in total pSMAD2 levels.

4.6.3 Evaluating crosstalk 3

Figure 4.11A presents the simulated model outputs for mechanism 3. Here, it is observed that the model fits are satisfactory for the high PI3K case. The low PI3K fits are very similar to mechanism 1, with the early time points showing good fits for p-SMAD2 and pre 6h time points for p-SMAD3. Posterior parameter distributions shown in Figure 4.11B indicate that the two models differ in the SMAD phosphorylation rates. Both p-SMAD2 and p-SMAD3 show increased phosphorylation in the low PI3K condition, with the mean values of SMAD3 further apart than SMAD2. Therefore, the parameter ensembles capture right type of interaction between SMAD and AKT. In addition, the differences in LRC dissociation rate are minimal, indicating that the endosomal LRC levels are similar between the two conditions (Figure C.6C). Therefore the increase in p-SMADs is solely due to increased phosphorylation rate of SMADs in low PI3K. SMAD7 related parameters show similar distributions as mechanism 1.

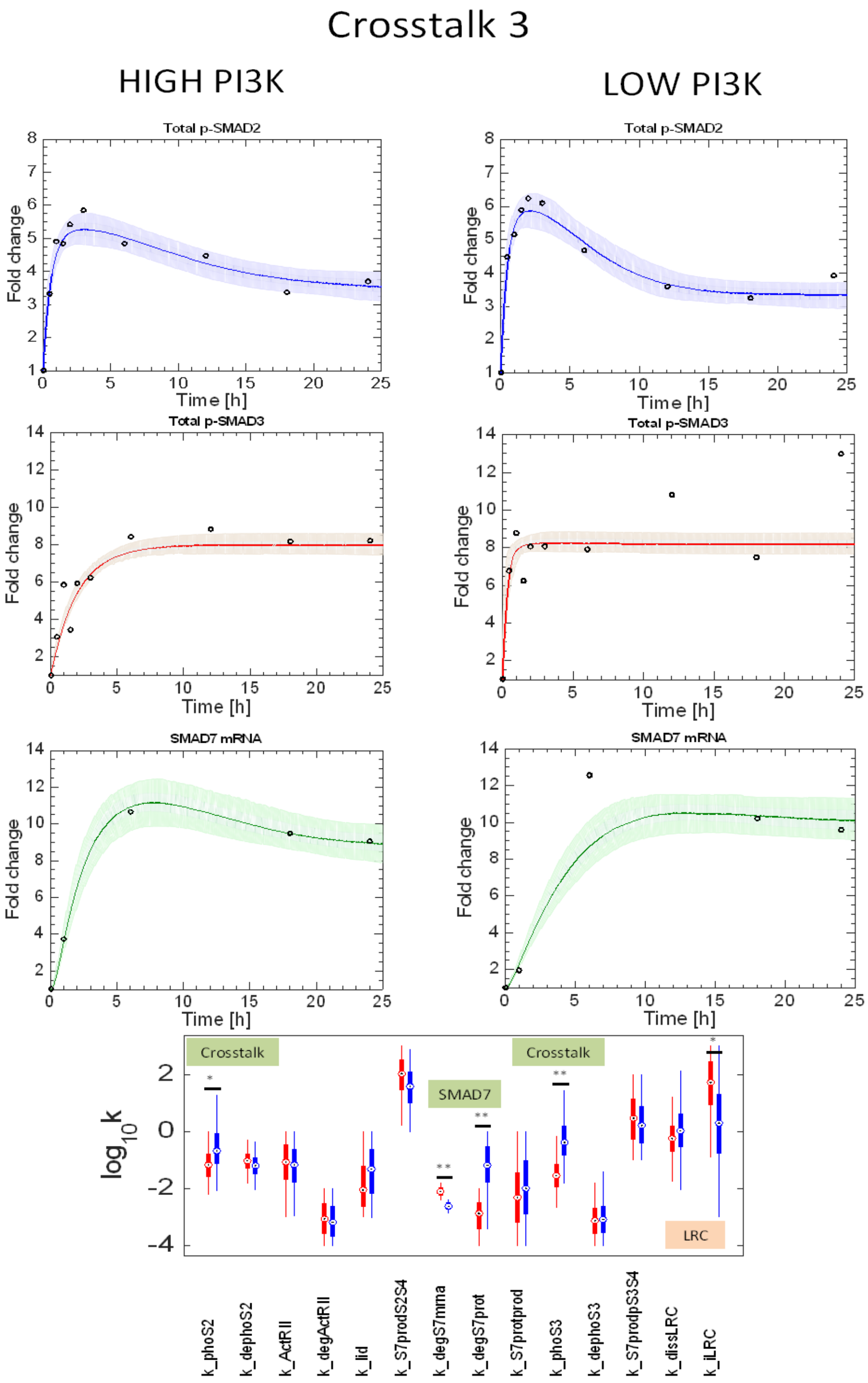


Figure 4.11 Ensemble modeling outputs for mechanism 3.

(A) Simulated model fits with the experimental training data. (B) Posterior distributions of the fitted parameters.

4.6.4 Evaluation of crosstalk combinations: Crosstalk 1 + 3

So far, mechanism 3 is able to explain most of the differences in p-SMAD2 and p-SMAD3 using increased phosphorylation rates and similar peak levels of LRC signaling complex in the endosomes. Mechanism 1 made predictions related to SMAD2 complexation that were opposite to the hypothesized mechanism and also predicted increased LRC complex in the endosomes for low PI3K. Mechanism 2 gave poor fits to the low PI3K condition. Next we wanted to check if combining mechanisms 1 and 3 gave better predictions. Figure 4.12A shows simulated model outputs for mechanism 1+3. The model fits were very similar to those obtained individually with 1 and 3. It was seen that LRC in the endosomes was identical between the two conditions, and therefore the differences in phospho-SMAD are only due to the modeled crosstalk interactions. However the parameter distributions (Figure 4.12B) showed that the phosphorylation rates are still higher in the low PI3K condition. Interestingly, the complexation rate of SMAD2 is lowered in the low PI3K condition unlike mechanism 1. Furthermore, the complexation rate of SMAD3 is similar between the two conditions. Therefore, according to this combination, SMAD3 probably is under the action of mechanism 3. For SMAD2, there is a combined action of mechanism 1 and 3. This improvement in biological feasibility comes with addition of two extra parameters as compared to mechanism 1 and four extra parameters as compared to mechanism 2 in the fitting process. The fits to SMAD7 mRNA are similar to each of the individual crosstalk mechanisms 1 and 3.

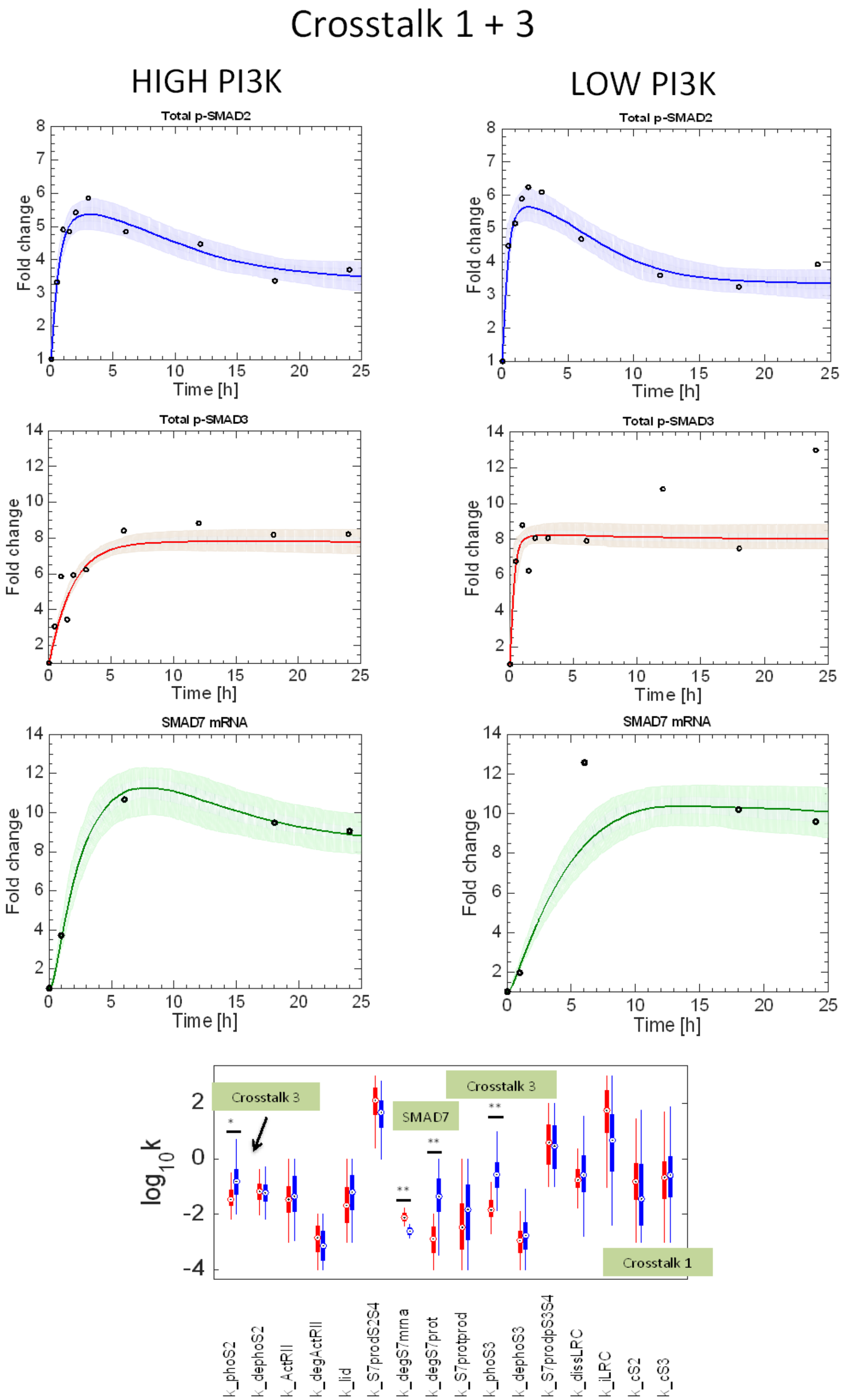


Figure 4.12 Ensemble modeling outputs for mechanism 1+3.

(A) Simulated model fits with the experimental training data. (B) Posterior distributions of the fitted parameters.

4.6.5 Model predictions of difference between crosstalk 1 and 3

In Section 4.6.4, the differences between crosstalk 1 and 3 from the simulation outputs were presented. The model also predicted other differences that can be tested experimentally. One such scenario is presented in Figure 4.13. Inhibition of activin signaling or non-stimulation condition followed by modulation of AKT levels will perturb the basal distribution of SMADs (non-phospho SMAD) between the cytoplasm and nucleus if crosstalk 1 is acting. In this situation, the non-phospho SMAD levels will be negligible. This effect will not be seen for crosstalk 3. Experimental measurement of total SMADs in the cytoplasm and nucleus (violet line for cytoplasm and orange line for nucleus in Figure 4.13) under high and low PI3K condition will enable verification of these differences. It is seen that the effect is more pronounced for the nucleus than the cytoplasm.

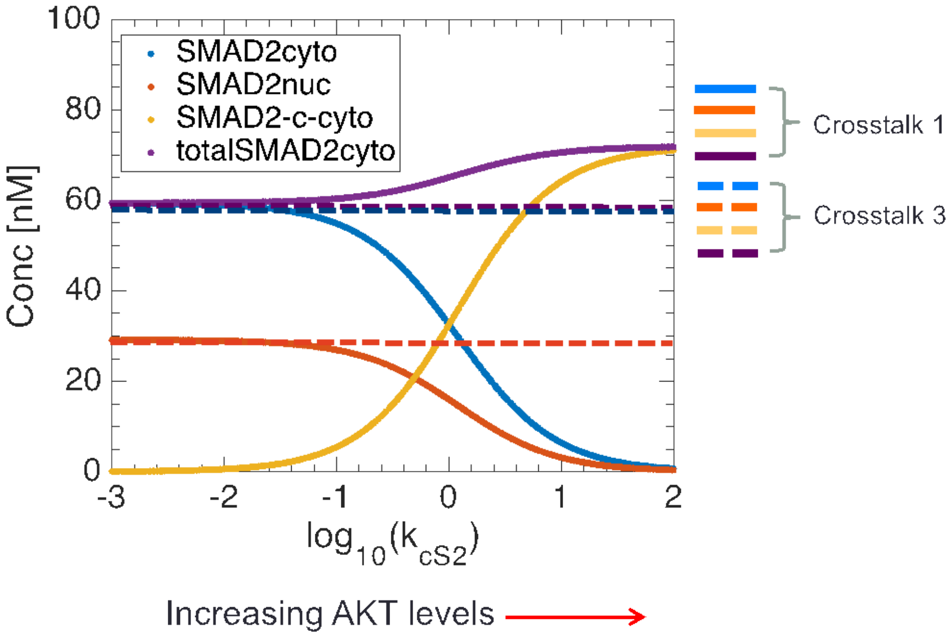


Figure 4.13 Model prediction distinguishing crosstalk 1 and 3.

Inhibition of Activin signaling (for example by inhibiting receptor activity) followed by modulation of AKT levels would perturb the nuclear-cytoplasmic equilibrium distribution of SMADs when crosstalk 1 is acting (thick lines). But no effect will be seen by modulation of AKT levels if crosstalk 3 is acting (dashed lines).

4.7 DISCUSSION (KINETIC MODELING)

In this section, we modeled the mechanism of AKT SMAD interactions and applied parametric ensemble analysis to check the plausibility of each mechanism to explain experimental observations. An important observation in hESCs was the difference in the dynamics of pSMAD2 and pSMAD3. pSMAD2 showed an overshoot behavior while pSMAD3 showed a first order increase in the initial phases of the signaling dynamics. Model predictions showed that this difference is likely due to the differences in the dephosphorylation rates. Some studies have indicated that low receptor levels could lead to delay in the activation of pSMADs (Schmierer and Hill, 2007), however this would affect SMAD2 and SMAD3 equally. Parameter estimation and sensitivity analysis on our model indicated that the receptor production and degradation rates showed a broad distribution. Therefore, the receptor levels have less control over SMAD dynamics as compared to phosphorylation and dephosphorylation rates in our system. Thus, our results indicate that low receptor levels alone are not sufficient to explain the difference in SMAD kinetics and a more direct influence via dephosphorylation could be necessary. But it is still important to note that the earliest peak in pSMAD2 and pSMAD3 occur around 2-3 hours and 1-3 hours respectively which is still later than 45-60 minutes seen in other cell types. Therefore, the receptor levels could be lower in hESCs that other mature cell types.

The dephosphorylation of SMADs is controlled by different phosphatases and several candidates have been identified in recent years (Bruce and Sapkota, 2012). However, there is no consensus on the most dominant phosphatase from the pool of identified candidates. Furthermore, the model of identical phosphatases regulating different SMAD types is now in question. The SMAD signaling community is increasingly recognizing that the phosphatase activity is likely to be different between SMAD2 and SMAD3, although a thorough experimental analysis is still lacking (Bruce and Sapkota, 2012). Our analysis provides the first evidence that hESCs are a good cell model to study the difference in the kinetics of SMAD2 and SMAD3 and future experimental work can identify the nature of the phosphatases catalyzing dephosphorylation of pSMAD2 and pSMAD3. Our modeling analysis has predicted the kinetics of the action of these phosphatases but their actual identity needs further experimental study. Negative feedback via SMAD7 was found to be less important in controlling the long term dynamics of SMAD2,3. This is possibly because SMAD7 is acting on the surface LRC complex which is a small fraction of the entire pool of LRC. Nevertheless, negative feedback in combination with dephosphorylation is the prime mechanisms to limit SMAD signaling in this model.

DBN inference indicated that AKT mediated crosstalk might be dominant in the early phases of signal transduction in the TGF-β/SMAD pathway. Kinetic analysis showed that addition of crosstalk interactions via AKT is helpful in capturing the early phase of the signal transduction differences between high and low PI3K conditions. Among the crosstalk mechanisms, the most promising model was where AKT inhibited the phosphorylation kinetics of SMAD2 and SMAD3. The effect was more dominant for SMAD3 than SMAD2. This is in line with the literature, where more studies have focused on AKT-SMAD3 crosstalk than AKT-

SMAD2 crosstalk in most cell lines. This was also seen in the DBN inference where AKT-SMAD3 connections were predicted but AKT-SMAD2 connections were not identified. One reason could be that correlation between AKT and SMAD2 time series is not strong enough relative to other connections of AKT and therefore, this connection is lost due to the 3-parent rule per child node imposed in the DBN algorithm.

The combination crosstalk model also explained the early signaling data well, although at the expense of additional parameters. From the modeling standpoint, this leads to higher AIC values than crosstalk 3 and thus not favorable. Comparison of the combination model with individual crosstalks is however interesting. SMAD3 phosphorylation is predicted to be inhibited by AKT in both scenarios. But, for SMAD2, there is a combination of both inhibition of SMAD phosphorylation and sequestration of SMADs. Though not experimentally proven, studies by Song *et al.* have indicated that crosstalk effects may be different for SMAD2 and SMAD3 (Danielpour and Song, 2006). Currently, our analysis cannot distinguish between these best cases. Future experiments probing the sequestration effect of AKT on SMAD2 will be necessary. This can be done by Fluorescence recovery of Photobleaching (FRAP) and Fluorescence loss in Photobleaching (FLIP) (Bancaud *et al.*, 2010; Phair and Misteli, 2001). Using FRAP/FLIP techniques, we can check if fluorescently tagged SMAD2 and SMAD3 accumulate in the cytoplasmic regions, possibly at the membranes where AKT is abundant (Meyer *et al.*, 2011). Such sequestration effects would be captured in the changes in the diffusivity coefficients of SMAD. Further, we can estimate the differences in the nuclear import and export kinetics of different SMADs, which were fixed to be the same in the current analysis (Schmierer and Hill, 2005). We have begun initial experiments to probe these connections (see Appendix D for preliminary experimental results and PDE modeling analysis for estimation of diffusivity of fusion protein, GFP-Profilin1 in a breast cancer cell line).

The kinetic analysis of crosstalk interactions with AKT presented in this section is unable to explain the long-term levels of pSMAD3 in low PI3K condition. This is also the region where AKT levels are increasing back to basal levels. Therefore, it is unlikely that AKT mediated interactions are responsible for the increase in pSMAD3 level in this time region of low PI3K condition. This was also indicated by the DBN inference in the late time zones of signaling dynamics. There could be influence of other crosstalk interactions that we have not measured and analyzed in this work. A prime candidate mechanism is the phosphorylation in the linker region of SMADs that can positively influence phospho SMAD3 levels, for example via proline directed cyclin dependent kinases that change during the cell cycle (Kamato *et al.*, 2013).

In our analysis, the levels of AKT were assumed to be constant in a particular condition since the rate parameters for crosstalk were assumed constant during the entire phase of the signaling dynamics. This is a valid assumption in the early phases of the signaling dynamics where we saw that the levels of AKT were either high or low (within experimental variability). Further, this is also the region where crosstalk effects are dominant. Future extensions of the model will require explicit consideration of AKT dynamics. This will enable integration of the mechanistic PI3K/AKT pathway model used in the self-renewal phase (Chapter 2) with the entire SMAD signaling machinery along-with the possible points of crosstalk.

4.8 CONCLUSIONS AND EXTENSIONS (KINETIC MODELING)

4.8.1 Major conclusions

In this work, we identified the unique dynamics of effector SMAD2,3 molecules in hESCs during early endoderm signaling. Through a thorough parametric ensemble analysis, we estimated the differences in the kinetics of reaction steps that lead to divergent dynamics of SMAD2,3. Further, detailed modeling of competing SMAD-AKT crosstalk interactions identified the differences in AKT mediated effects on SMAD2 and SMAD3. These results demonstrate the use of detailed mathematical models for evaluation of signaling mechanisms and regulatory interactions guiding endoderm differentiation.

4.8.2 Assumptions, potential pitfalls and proposed extensions

In the current work, the focus was on C-terminal phosphorylation of SMAD2,3 which is catalyzed by LRC and is the major phosphorylation event controlling processes in the early signaling phase. However, additional sites on the SMAD molecules undergo phosphorylation and this may lead to modulation of reactions in which SMAD participates. For example, phosphorylation in the linker region of SMAD proteins may affect the nuclear export rates (Pauklin and Vallier, 2013) and degradation of SMADs (Jason *et al.*, 2015). In the current modeling scenario, the total number of SMAD molecules in the cell is held constant. But protein degradation events, which have slower kinetics than the processes included in the current model, will become important in long term signaling and perturb the constant SMAD levels. SMAD7 dynamics under low PI3K was not captured effectively. Capturing the low PI3K SMAD7 behavior led to poor fits to pSMAD2 and pSMAD3 indicating some interactions not accounted for. Smad7 protein levels were not measured here and it is possible that the actual pool of SMAD7 proteins taking part in negative feedback may not be directly proportional to the mRNA levels. Measuring the total SMAD7 protein levels will provide additional data to confirm this.

In this work, only one source of crosstalk is modeled, namely AKT (ERK was found to be unimportant in general). However, ERK may affect linker phosphorylation, which was not measured in the current experiments. In other words, ERK (and other molecules like CDKs) may indirectly influence SMAD activity via linker phosphorylation and these effects may become more prominent in the later phases of signaling dynamics. It is important to note that linker phosphorylation has a slower kinetics than C-terminal phosphorylation (Kamato *et al.*, 2013). On the other hand, crosstalk with other developmental pathways like WNT/β-catenin might also affect SMAD activity. These open questions can be explored in future studies and our work will provide a framework to design the stimulation conditions for these experiments.

5.0 OVERALL CONCLUSIONS AND FUTURE DIRECTIONS

In this work, we have performed a systems level analysis of signal transduction in the self-renewal state and during endoderm differentiation process of hESCs. Through the use of data-driven and equation based mechanistic models, we have analyzed the process by which signals integrate between disparate yet interacting signaling pathways in hESCs. Through the three aims, we have identified critical interactions and regulatory processes that robustly control features of hESC signaling and therefore, are best candidates for further exploration and optimization of hESC cultures. In the following sections, the major contributions and possible applications of the techniques developed in each aim are presented.

5.1 AIM 1: SENSITIVE NODES OF COMPLEX PATHWAYS

Our first aim focused on application of computationally efficient algorithms to identify critical nodes from the PI3K/AKT pathway in the self-renewal state of hESCs. Our motivation was based on the hypothesis that these critical (or sensitive) nodes are the best places to target and modulate the pathway to achieve the desired aim of improving signals promoting self-renewal. Identification of such positions from complex signaling pathways is non-trivial, time consuming and computationally expensive. Because of the non-linearity in the kinetics of the interactions, the pathway of signal transduction becomes non-intuitive. To identify the sensitive nodes, we employed a novel meta-model based global sensitivity analysis based on high dimensional model representation technique developed for large-scale chemical reacting systems. We performed the sensitivity analysis in biologically realistic range for the rate parameters identified from experimental time series data. Through this work, we have demonstrated the application of meta-modeling techniques to reduce computational costs associated with traditional MC approaches for GSA. According to the current analysis, we observed a 10-fold reduction in the computational cost for evaluation of Sobol' indices and the accuracy of the indices of most sensitive parameters was good at low MC sample sizes (after around 1000 MC samples).

The meta-model approximation used for GSA is general and can be applied to different scenarios. For example, analysis of simulation outputs and analysis of experimental data (Li *et al.*, 2001a). For the latter, if fairly large input-output experimental datasets (~10^3 for first order analysis) are available, then we can model the input-output space for a high dimensional system. The outputs of second-order analysis indicate the influence of pairs of perturbations and often, these may provide better combinations of small molecule perturbations than single molecule perturbations. This is well known for pharmacological therapies in illnesses like cancer, but only empirically studied in stem cell biology. The meta-model technique can be expanded to evaluate third order indices, but this will require additional regularization strategies to reduce overfitting and inaccuracy (Miller *et al.*, 2012). Thus, overall we can identify best combinations of perturbations to improve self-renewal state using a systematic analysis that draws upon the power of mechanistic models. This will provide a parallel method to high throughput screening assays. In future, we can evaluate the entire network with crosstalks and identify co-modulation strategies to control the balance of self-renewal and differentiation. This will require model extension through addition of crosstalk interactions, some of which have been studied in AIM3.

5.2 AIM 2: CO-REGULATION OF TRANSCRIPTION FACTORS UNDER COMBINATORIAL SIGNAL INPUTS

In the second aim, we focused on the end stage of endoderm differentiation of hESCs using combinations of important signaling pathways. Starting from five commonly employed signaling pathways, we explored all possible combinations of these pathways to identify similarities and differences in the TF expression at the end-point of differentiation. We applied biclustering formulation to identify subsets of TFs that are co-regulated under sub-sets of conditions. To ensure that the identified biclusters are robust against experimental noise, we employed a bootstrapping with resampling approach. Our analysis primarily indicates that efficiency of endoderm induction depends on the context of BMP4 signaling. Therefore, protocols employing BMP4 signaling in combination with other pathways would require careful consideration when evaluating the later stage maturation potential.

Biclustering with bootstrapping framework has not found applications for analyzing signal transduction data. In the hESC literature, most signal transduction data is very sparse and composed of population averages and few different signaling molecules. Many hESC research labs use large-scale arrays for gene expression measurements and mass-spectroscopic measurements for phosphoproteins. Biclustering techniques can be used here to mine patterns of signal transduction dynamics that are relevant for fate choice, a question that is still unanswered (Schneider *et al.*, 2012). This could be an important step to effectively connect signaling pathways with gene regulatory networks to simulate the entire dynamics of fate choice. With newer platforms like Reverse Phase Protein Array for analyzing signal transduction molecules on a large scale (Iadevaia *et al.*, 2010; Tibes *et al.*, 2006), we envision that the techniques proposed in this aim will become commonplace. These experimental platforms, which provide large number of data points, will give better approximations for the bootstrapping approach that is sensitive to the empirical distribution. Our analysis was done on the end-point of differentiation. Performing the analysis with dynamic profiles will be useful to identify co-regulation patterns across intermediate stages like mesendoderm. This would require extension to the third dimension and this is possible using triclustering techniques (Mahanta *et al.*, 2011).

5.3 AIM 3: CROSSTALKS AND DYNAMICS OF SIGNAL PROPAGATION DURING ENDODERM DIFFERENTIATION

In the third aim, we focused on the main pathway of endoderm induction and associated crosstalks with parallel pathways. We first measured the dynamics of key signaling molecules from the TGF-β/SMAD pathway with additional molecules from PI3K/AKT and MAPK/ERK pathways. Using a DBN framework, we analyzed the within pathway and between pathway interactions, information about which was hidden in the experimental time series. DBN inference is applicable for wide variety of expression datasets and its application to protein expression datasets is fairly recent (Azhar *et al.*, 2013). Due to the large number of crosstalks in hESC signal transduction networks, DBN inference can provide a framework to generate hypotheses about possible interactions which can be tested further by perturbations of the associated nodes. Even for small number of signaling molecules, the associations provided by DBNs are informative.

As demonstrated by our analysis, DBN interactions can be probed further using detailed ODE models that enable identification of kinetics of these interactions. Application of ensemble parametric methods enables effective characterization of identifiability of parameters of systems biology models. Using this analysis, we identified possible reasons for the complex signaling behavior in hESCs. Although we have not characterized the actual identity of molecules catalyzing the differences, our model has made several predictions about the nature of these interactions. Mechanistic modeling of crosstalk interactions in a systematic way can be conducted using the same mathematical framework. These methods enable testing of competing hypotheses by identifying model predictions that are different between them. This can eventually help in building bigger network level representation of signaling during differentiation. Once complex predictive models are available, we can identify best possible treatments that will keep the differentiation relevant signals active and improve overall differentiation process. Taken together, we can utilize the information gained from different types of modeling frameworks employed in this work. These improvements will eventually enable development of well-defined cell types for cellular transplantation applications.

5.4 COMMON ASSUMPTIONS, PITFALLS, FUTURE DIRECTIONS

Throughout the dissertation, the experimental protein signaling data used for model calibration was obtained using a MagPix assay. Before analysis, cells in a well are lysed and their internal contents are collected together, similar to a western analysis. Therefore, the final fluorescence intensities are proportional to the cumulative protein content of the entire population in a well. This is the common method of analyzing signal transduction dynamics of many models (Janes, 2015). However, one major drawback of this method is the loss of single cell resolution. Phenomena like oscillations and cell-to-cell variability are hidden when data is collected at the population level (de Vargas Roditi and Claassen, 2015). Therefore, future studies using single cell measurements of the dynamics of signaling molecules will enable exploration of these phenomena. This requires generation of hESC cell lines with fluorescently tagged signaling proteins. The method of generation of such cell lines are critical, because this leads to additional complications like increased expression levels of the proteins etc. which may lead to non intended changes in the signaling behavior. We envision that population and single cell measurements must go hand in hand and this is an emerging area of analysis in signal transduction research. The population measurements may be used to verify the single cell behavior or to provide constraints, since population behavior should be a geometric combination of single cell measurements (Hasenauer *et al.*, 2014). Mathematical modeling will provide a link to connect these two measurement methods.

Variability is an inherent feature of hESC systems and the experimental variability seen in the three aims may have multiple origins that are specific to hESCs. Some of these are discussed below and these factors may need special attention in future studies of signal transduction in hESCs. hESCs grow as colonies and location within the colony may modify the behavior of these cells to cell fate manipulation (Rosowski *et al.*, 2015). The effect of spatial location within the colony on the signaling behavior has not been studied here, but this may contribute to variability seen in the experiments. Reorganization or breakage of colony structure during differentiation must be considered when modeling long term signaling since this changes the local signaling cues experienced by the cells. Further, random X-chromosome inactivation, during differentiation in female cell lines may lead to epigenetic alterations and this may affect the signaling behavior of daughter cells (Shen *et al.*, 2008). Future studies that will model long term signaling will have to consider the effect of cell cycle and cell density on the signaling outcomes. Location within the cell cycle is an important determinant of the culture heterogeneity in hESCs (Singh *et al.*, 2013).

In the current work, all experiments were done on H1 hESCs. Comparison between different hESC cell lines is an area of active research and many studies have shown that differences in hESC cell lines may lead to different behaviors under the same conditions (Allegrucci and Young, 2007). However, not much has been done to compare differences at the signal transduction level. Predictions from the current dissertation need to be verified in multiple cell lines including induced pluripotent cell types. This will lead to strategies that are common across cell lines and at the same time, help us to identify sources of variation in signal transduction between cell lines.

APPENDIX A

CHAPTER 2 ADDITIONAL MATERIAL

The mechanistic model of insulin signaling is based on the nonlinear ODE model developed by Sedaghat *et al.* (Sedaghat et al., 2002). The model is divided into several steps: insulin binding kinetics, receptor internalization and recycling, post-receptor activation of PI3K/AKT pathway and positive and negative feedback pathways. For the purpose of our analysis, the model was divided into two major modules, namely; (*M1*) Receptor-insulin binding kinetics and intracellular receptor trafficking, (*M2*) Post-receptor PI3K/AKT pathway with feedbacks. The complete version of the model consists of 20 state variables (including various states of the same molecule). For the sensitivity analysis, we utilized 21 rate parameters, 1 initial condition (insulin) and 3 modulators of the signaling pathway (PTP, PTEN and SHIP). Each reaction is modeled using mass-action kinetics unless stated otherwise. Total concentrations of the post-receptor signaling molecules are assumed to be constant for the time scale of the experiment.

Module M1

In the absence of feedback, *M1* is completely independent of the post receptor signaling. Therefore, this module is solved independently of *M2* to estimate the maximal phosphorylation of surface receptors achieved for every parameter set explored. This maximal value of $y_4 + y_5$ at steady state is referred to as IR_p in the original Sedaghat model. Once IR_p is evaluated the complete model (*M1*+*M2*) is evaluated again. However, for sensitivity analysis, the rate parameter, k_7, is combined with IR_p. The selected free parameters for global sensitivity analysis are provided in Table A.1. The state variables and the corresponding ODEs for this module are described below:

y_1 = insulininput
y_2 = Conc.unboundsurfaceinsulinreceptors
y_3 = Conc.non – phospho,once – boundsurfaceinsulinreceptors
y_4 = Conc.phospho,twice – boundsurfaceinsulinreceptors
y_5 = Conc.phospho,once – boundsurfaceinsulinreceptors
y_6 = Conc.unboundnon - phospho intracellularinsulinreceptors
y_7 = Conc.phosphotwice - bound intracellularinsulinreceptors
y_8 = Conc.phosphoonce - bound intracellularinsulinreceptors
[PTP] = Conc.of protein tyrosine phosphatases like PTP1B

$$
\begin{aligned}
&\dot{y}_1 = 0\\
&\dot{y}_2 = k_{-1}y_3 + k_{-3}[PTP]y_5 - k_1y_1y_2 + k_{-4}y_6 - k_4y_2\\
&\dot{y}_3 = k_1y_1y_2 - k_{-1}y_3 - k_3y_3\\
&\dot{y}_4 = k_2y_1y_5 - k_{-2}y_4 + k_{-4'}y_7 - k_{4'}y_4\\
&\dot{y}_5 = k_3y_3 + k_{-2}y_4 - k_2y_1y_5 - k_{-3}[PTP]y_5 + k_{-4'}y_8 - k_{4'}y_5\\
&\dot{y}_6 = k_5 - k_{-5}y_6 + k_6[PTP](y_7 + y_8) + k_4y_2 - k_{-4}y_6\\
&\dot{y}_7 = k_{4'}y_4 - k_{-4'}y_7 - k_6[PTP]y_7\\
&\dot{y}_8 = k_{4'}y_5 - k_{-4'}y_8 - k_6[PTP]y_8
\end{aligned}
$$

$$
\begin{aligned}
&IR_p = (y_4 + y_5)_{ss} \text{in the absence of feedback}\\
&\text{ss - steady state}
\end{aligned}
$$

The dependent rate parameters for *M1* are evaluated as follows:

$$
\begin{aligned}
&k_2 = k_1\\
&k_{-2} = 100k_{-1}\\
&k_{-3} = k_{-1}\\
&k_4 = k_{-4}/9\\
&k_5 = \begin{cases} 10k_{-5} & if\ (y_6 + y_7 + y_8) > 10^{-13} \\ 60k_{-5} & if\ (y_6 + y_7 + y_8) \le 10^{-13} \end{cases}
\end{aligned}
$$

Module M2

This module contains the cascade of signaling events following the activation of insulin receptors on the cell surface as described in the text. The signal terminates at the AKT and PKCζ nodes. Activation of these terminal nodes results in two types of feedback: positive feedback by p-AKT resulting in modulation of PTP and negative feedback from p-PKCζ resulting in the serine phosphorylation of IRS-1 node. The state variables and the corresponding ODEs are described below:

y_9 = Conc. non - phospho IRS1

y_{10} = Conc. of tyrosine phospho - IRS1

y_{11} = Conc. unactivated PI3K

y_{12} = Conc. tyrosine phosphorylated IRS1/activated PI3K complex

y_{13} = Percentage of PI(3,4,5)P_3 out of the total lipid population

y_{14} = Percentage of PI(4,5)P_2 out of the total lipid population

y_{15} = Percentage of PI(3,4)P_2 out of the total lipid population

y_{16} = Percentage of non - phospho - AKT

y_{17} = Percentage of phospho - AKT

y_{18} = Percentage of non - phospho - PKCζ

y_{19} = Percentage of phospho - PKCζ

y_{20} = Conc. of serine phospho - IRS1

$$\dot{y}_9 = k_{-7}[PTP]y_{10} - k_7 y_9 \frac{(y_4 + y_5)}{IR_p} + k_{-7'} y_{20} - k_{7'}[PKC]y_9$$

$$\dot{y}_{10} = k_7 y_9 \frac{(y_4 + y_5)}{IR_p} + k_{-8} y_{12} - (k_{-7}[PTP] + k_8 y_{11}) y_{10}$$

$$\dot{y}_{11} = k_{-8} y_{12} - k_8 y_{10} y_{11}$$

$$\dot{y}_{12} = -k_{-8} y_{12} + k_8 y_{10} y_{11}$$

$$\dot{y}_{13} = k_9 y_{14} + k_{10} y_{15} - k_{-9} y_{13} - k_{-10} y_{13}$$

$$\dot{y}_{14} = k_{-9} y_{13} - k_9 y_{14}$$

$$\dot{y}_{15} = k_{-10} y_{13} - k_{10} y_{15}$$

$$\dot{y}_{16} = k_{-11} y_{17} - k_{11} y_{16}$$

$$\dot{y}_{17} = k_{11} y_{16} - k_{-11} y_{17}$$

$$\dot{y}_{18} = k_{-12} y_{19} - k_{12} y_{18}$$

$$\dot{y}_{19} = k_{12} y_{18} - k_{-12} y_{19}$$

$$\dot{y}_{20} = k_{7'}[PKC]y_9 - k_{-7'} y_{20}$$

In the original model, the dependent rate parameters for *M2* are evaluated as follows:

$$k_{-7} = 0.336 k_7 \frac{1}{[PTP]_{basal}}$$

$$k_{-7'} = k_{7'} \frac{1.24\times 10^{-13} \frac{1}{[PTP]_{basal}}}{\left(6.24\times 10^{-13} - 1.24\times 10^{-13} \frac{1}{[PTP]_{basal}}\right)}$$

[PTP]$_{\text{basal}}$ is the basal level of PTP selected by the parameter sampling. The levels of PTP change because of the feedback by phospho-AKT. Hence, PTP is modeled as:

$$[PTP] = \begin{cases} PTP_{basal}\left(1 - 0.25 y_{17} \frac{11}{100}\right) & for\ \ y_{17} \le \frac{400}{11} \\ 0 \ \ otherwise & \end{cases}$$

p-AKT results in the 25% inhibition of PTP under nominal conditions. At steady state, the level of p-AKT is 100/11 under nominal conditions without feedback. When the p-AKT levels rise four times above the nominal levels, the PTP levels are set to zero.

$$k_8 = 7.065\times 10^{10} k_{-8}$$

$$k_9 = \left(k_{9stim} - k_{9basal}\right)\frac{y_{12}}{PI3K} + k_{9basal}$$

$$k_{9basal} = 3.12\times 10^{-3} k_{-9}$$

$$k_{-9} = 30.32 k_{9stim}$$

$$k_{10} = 1.069 k_{-10}$$

The coefficients for the above equations are set from the nominal case.

$$k_{11} = 0.1k_{-11}\frac{y_{13} - 0.31}{3.1 - 0.31}$$

$$k_{12} = 0.1k_{-12}\frac{y_{13} - 0.31}{3.1 - 0.31}$$

$$PI3K = \frac{k_8\left(3.7 \times 10^{-13} \times 1 \times 10^{-13}\right)}{k_8 3.7 \times 10^{-13} + k_{-8}}$$

The constants in the equation for PI3K are again selected from the nominal case.

The action of negative feedback by p-PKCζ is modeled by a Hill equation for enzyme kinetics, $[PKC] = \frac{V_{\max} Y_{19}{}^{n}}{K^{n}{}_{d} + Y^{n}{}_{19}}$. We do not incorporate delay into the equation as originally included in the Sedaghat model since we are performing a steady state analysis while doing GSA. The total number of free rate parameters for the complete model is as follows:

Module *M1*: 8

Module *M2*: 8

Additional modifications for GSA

As seen in the previous section, many of the rate parameters of the signaling cascade (*M2*) are related by equilibrium ratios. The equilibrium ratios enable one to calculate one of the forward or backward reaction rates when the other is known. However, for GSA, these parameters were chosen independently so that the effect of each of these reactions on the final steady state can be effectively studies. Otherwise, the final steady state will be the same since both the forward and backward reactions are perturbed to the same extent. Also, for application to completely new systems, it is not known whether the same equilibrium ratios hold. It is important to note that the rate constants, k_{11}, k_{-11} and k_{12}, k_{-12} were kept constant as they were phenomenological values in the original model and were not calibrated to any experiments. In short, the final list of free input parameters was expanded to include k_{-7}, $k_{-7'}$, k_8, k_{-9}, k_{10}. This new list of parameters now includes 21 rate parameters (16 original + 5 additional), 1 initial concentration (insulin) and 3 modulators of the pathway. The nominal values and the explored ranges of these parameters are presented in Table A.1 of the main manuscript. The equations/parameters that were modified are given below:

(1) k_7 was changed to k_7 / IR_p with a nominal value of 4.64×10^{12}.

(2) The parameter, k_{9stim}, was combined with $PI3K$. The parameter, k_9, was modified to avoid negative values under parameter perturbations. The new equation is $k_9 = k_{9stim} y_{12} + k_{9basal}$

(3) The parameters, k_{11} and k_{12}, were modified to avoid negative values. The new equations are:

$$k_{11} = k_{11basal} y_{13}$$

$$k_{12} = k_{12basal} y_{13}$$

$$k_{11basal} = \frac{0.1 k_{-11}}{(3.1 - 0.31)}$$

and similar equation for $k_{12basal}$. Here, the basal values of the parameters are the same as the original model.

(4) In the original model, PTP was described by a piecewise function that was linear in the low p-AKT regime and abruptly became zero when p-AKT was 4 times the nominal value. However, this was not based on any precise experimental measurements. In our hESC system, the levels of p-AKT were still in the lower range and negative feedback was still dominant. Therefore, we chose to model PTP using a continuous linear function that went to small values

(but not zero) when p-AKT was high (> 36.36). The levels reach zero when all the p-AKT molecules are phosphorylated (i.e. pAKT = 100%).

$$PTP = PTP_{basal}\left(1-0.01y_{17}\right)$$

The basal PTP levels are perturbed during the sensitivity analysis in the range 0.5 to 1.5. Therefore, under nominal conditions without feedback (i.e. p-AKT equal to 100/11), PTP levels can change to 0.45 and to 1.36 for the basal levels of 0.5 and 1.5 respectively.

Table A.1 Free input parameters used for the MC simulation and their explored ranges**

No.	Symbol	Nominal value	Range
M1			
1	k_1	6×10^{7} $M^{-1} \bullet min^{-1}$	$[10^{6}, 10^{8}]$
2	k_{-1}	0.2 min^{-1}	$[10^{-2}, 5]$
3	k_3	2500 min^{-1}	$[10^{2}, 10^{4}]$
4	k_{-4}	0.003 min^{-1}	$[10^{-4}, 10^{-2}]$
5	$k_{4'}$	2.1×10^{-3} min^{-1}	$[10^{-4}, 10^{-2}]$
6	$k_{-4'}$	2.1×10^{-4} min^{-1}	$[10^{-5}, 10^{-3}]$
7	k_{-5}	1.67×10^{-18} min^{-1}	$[10^{-19}, 10^{-17}]$
8	k_6	0.461 min^{-1}	$[10^{-2}, 10^{0}]$
M2			
9	$\frac{k_7}{IR_p}$	4.64×10^{12} $M^{-1} \bullet min^{-1}$	$[4\times10^{12}, 10^{13}]$
10	k_{-7}	1.396 min^{-1}	[0.1, 6]

Table A.1 (continued)

11	k_8	7.06×10^{11} $M^{-1} \cdot min^{-1}$	$[5\times10^{11}, 10^{12}]$
12	k_{-8}	10 min^{-1}	[1, 50]
13	k_{-9}	42.15 min^{-1}	[1, 50]
14	k_{9stim}	4.96×10^{14} $M^{-1} \cdot min^{-1}$	$[4\times10^{14}, 10^{15}]$
15	k_{10}	2.96 min^{-1}	[0.1, 50]
16	k_{-10}	2.77 min^{-1}	[0.1, 10]
17	$k_{7'}$	0.347 min^{-1}	[0.01, 0.5]
18	$k_{-7'}$	0.0858 min^{-1}	$[10^{-3}, 10^{-1}]$
19	$V_{\max}$	20	[1, 50]
20	K_d	12	[1, 20]
21	n	4	[1, 5]
22	PTP_{basal}	1.00	[0.5, 1.5]
23	$[SHIP]$	1.00	[0.5, 1.5]
24	$[PTEN$	1.00	[0.5, 1.5]
25	$y_1(0)$	10^{-7} M	$[10^{-9}, 10^{-6}]$

**The values of the parameters, (k_{11}, k_{12}), and (k_{-11}, k_{-12}) were kept constant at 0.693 and 6.93 respectively. The initial conditions for all the species except insulin was kept constant at values same as (Sedaghat *et al.*, 2002).

Table A.2 Selection of number of clusters

Number of clusters	Mean Silhouette value (S_{mean})
1	-
2	0.55
3	0.61 (Optimal)
4	0.56
5	0.53

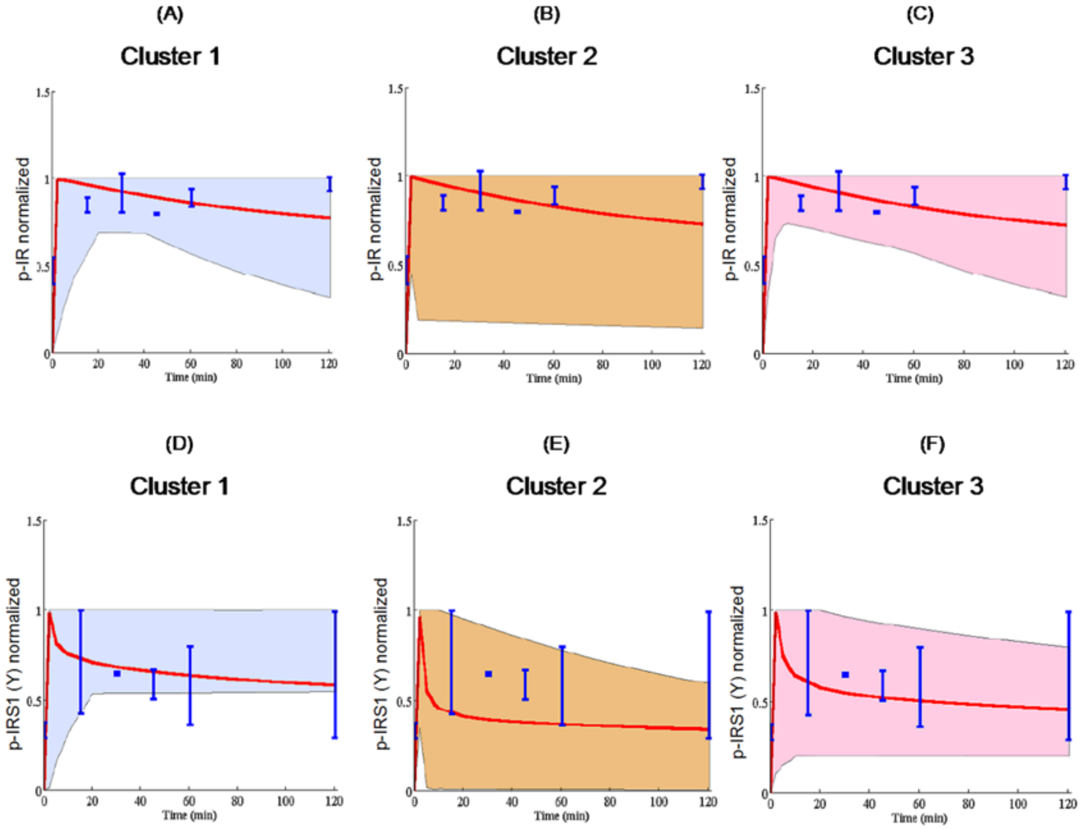


Figure A.1 K-means clusters of p-IR and p-IRS1 (Y) dynamics observed in 10^5 samples.

These profiles are presented for the k-means clustering done on the p-AKT profiles from Figure 2.3. The shaded areas represent the range of the profiles present in the cluster and the red curve represents the cluster centroid. The blue error bars represent the experimental data normalized to maximum. (A-C) p-IR clusters do not show significant differences and all the clusters are close to the experimental data. This is consistent with the GSA results that show that p-AKT does not significantly affect the molecules upstream of IRS1 (Y) under the current input conditions. (D-F) p-IRS1 (Y) clusters 1 and 3 fall closer to the experimental data as judged by the centroids.

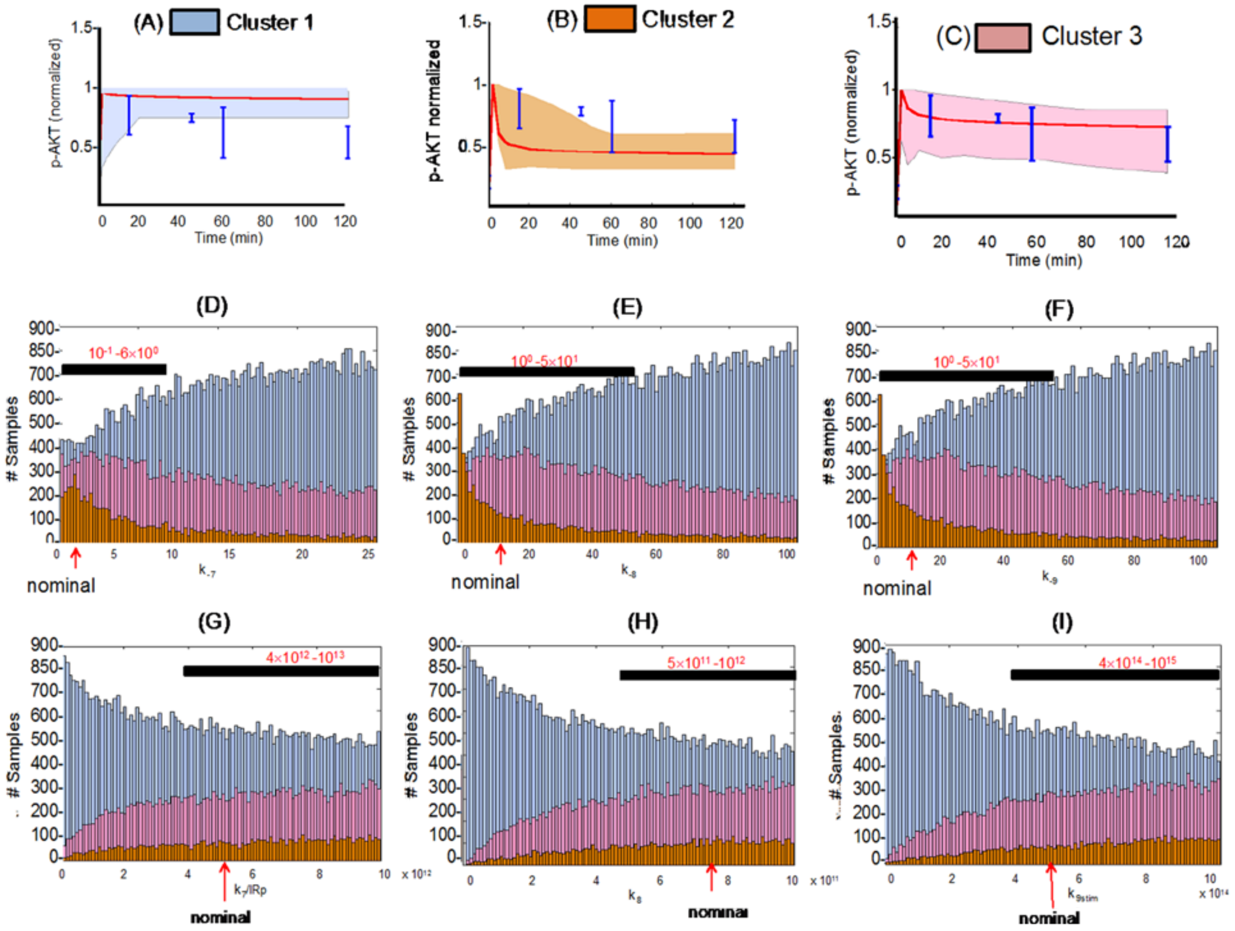


Figure A.2 Selection of parameter ranges for hESC dynamics.

(A-C) K-means clustering of p-AKT dynamics observed in 10^5 MC samples. (D-I) Parameter collection in the three k-clusters of p-AKT. The histograms present the number of samples in each cluster in a given parameter interval. The bottom red arrow shows the location of the nominal value. For each plot, black bars show final range chosen for sensitivity analysis.

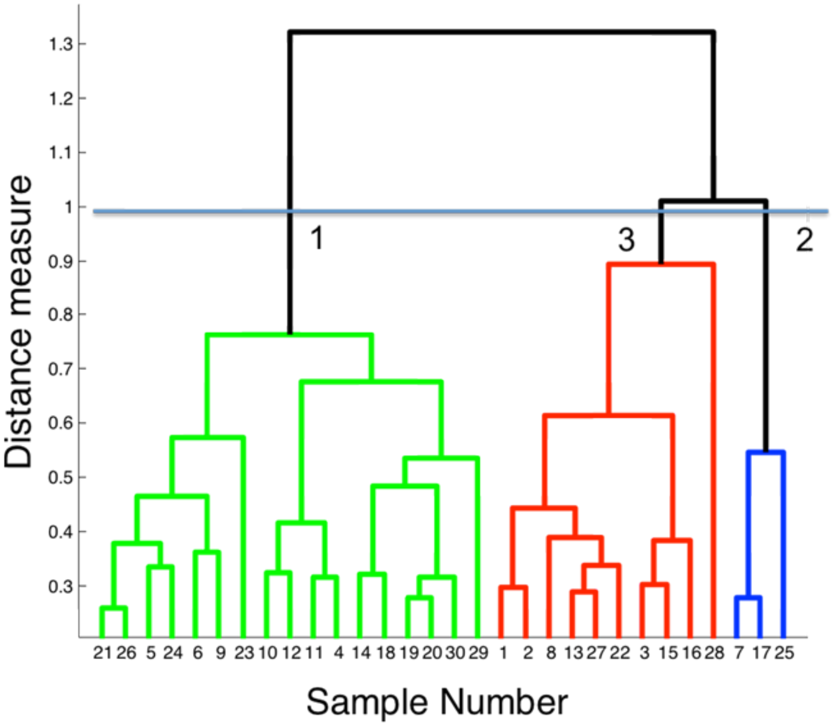


Figure A.3 Dendrogram showing the clustering of p-AKT profiles from the same samples as used for Figure A.2.

A cluster node limit of 30 was used for plotting purposes, but the clustering was performed on the entire 10^5 samples. Increasing the cluster node limit will expand the current terminal nodes further. It is seen that a total of three clusters were found to represent most dissimilar dynamics that were the same as k-means clustering analysis (and same parameter ranges were seen). The numbers at the nodes of these clusters denote the k-means clusters from Figure A.2 to which they belong. Cophenetic correlation coefficient was estimated to be 0.79 and the Spearman correlation coefficient was estimated to be 0.85 showing an acceptable clustering.

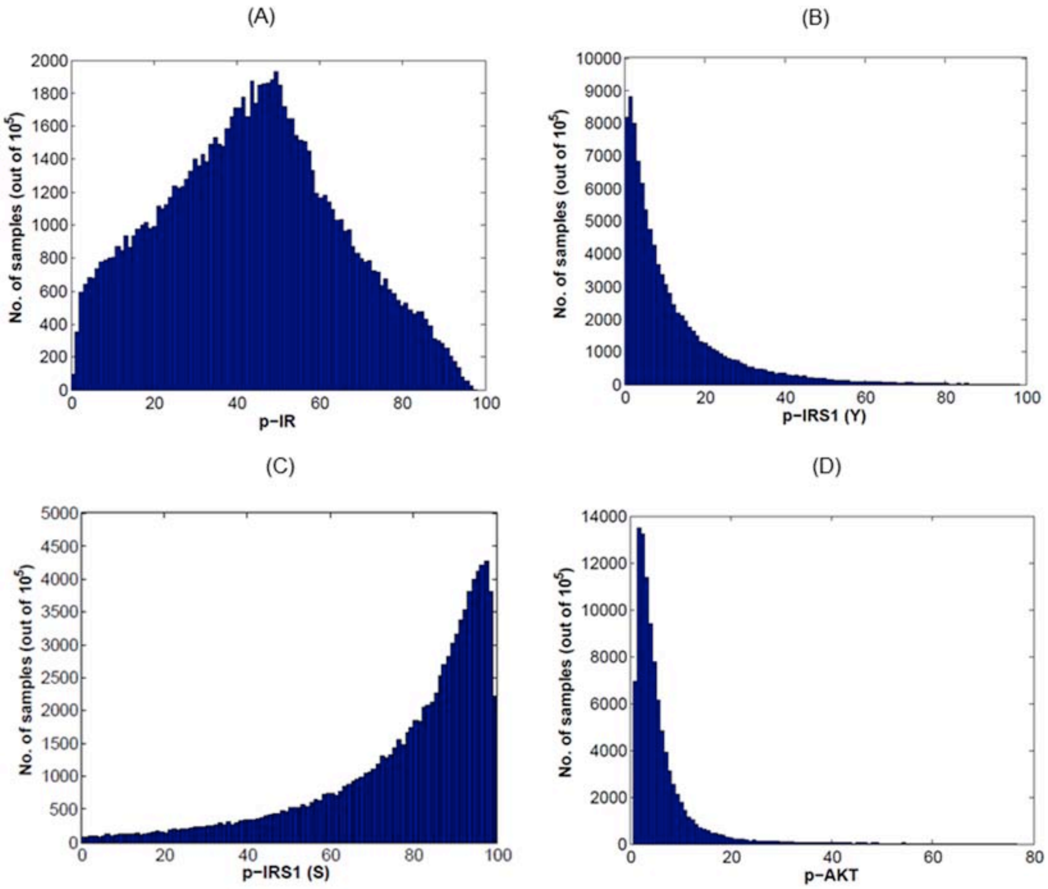


Figure A.4 Histogram of output distributions in the 10^5 MC samples used for RS-HDMR.

(A) % p-IR shows abundant intermediate levels (B) % p-IRS1 (Y) shows a skewed distribution with abundant low phosphorylation states (C) % p-IRS1 (S) shows a skewed distribution with abundant high phosphorylation states (D) % p-AKT distribution shows a skewed distribution with abundantly low levels at steady state.

Table A.3 Performance of second order RS-HDMR meta-model**

10^5 samples	Mean and Range of Output (%)	Total Variance** ($\%^2$)	$\sum_{i=1}^{25} S_i$	$\sum_{i=1}^{25} \sum_{j=i+1}^{25} S_{ij}$	R^2
p-IR	44 [0-100]	453	0.89	0.08	0.97
p-IRS1 (Y)	12 [0-100]	195	0.69	0.21	0.90
p-IRS1 (S)	78 [0-100]	429	0.72	0.19	0.91
p-AKT	6.5 [0-80]	62	0.59	0.28	0.87

****From our observation of total variance, we see that it is relatively easier to perturb the levels of p-IR and p-IRS1 (S) and (Y) that show large variances while it is difficult to perturb p-AKT levels that shows a relatively small variance. For outputs, p-IR, p-IRS1 (Y) and p-IRS1 (S), the first order contributions are sufficient while for p-AKT there are additional contributions from second order processes.**

Table A.4 Important second order processes affecting p-AKT

Group	Location in the pathway and function	Parameters	Actual contribution to total variance (%)
A	Negative regulators upstream of PIP_3	k_{-9}, k_{-8}, k_{-7}	7.3
B	Negative regulators upstream of PIP_3 with negative feedback regulators downstream of PIP_3	k_{-9}, k_{-8} with $k_{7'}$, $k_{-7'}$, V_{max}	6.2

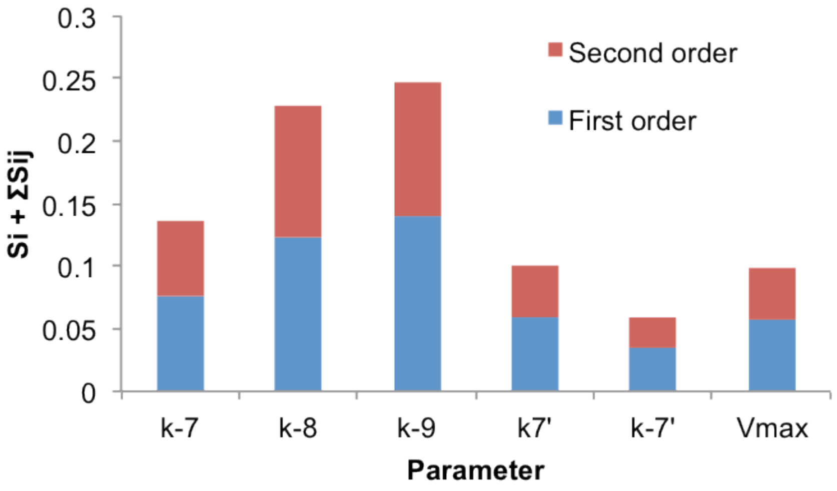


Figure A.5 Sum of first and second order indices.

Combined contribution of first and all relevant second order indices for the two important groups: negative regulators and negative feedback parameters. Second order interactions increase the sensitivity of the most sensitive parameters (for example, from 0.15 to 0.25 for k_{-9}).

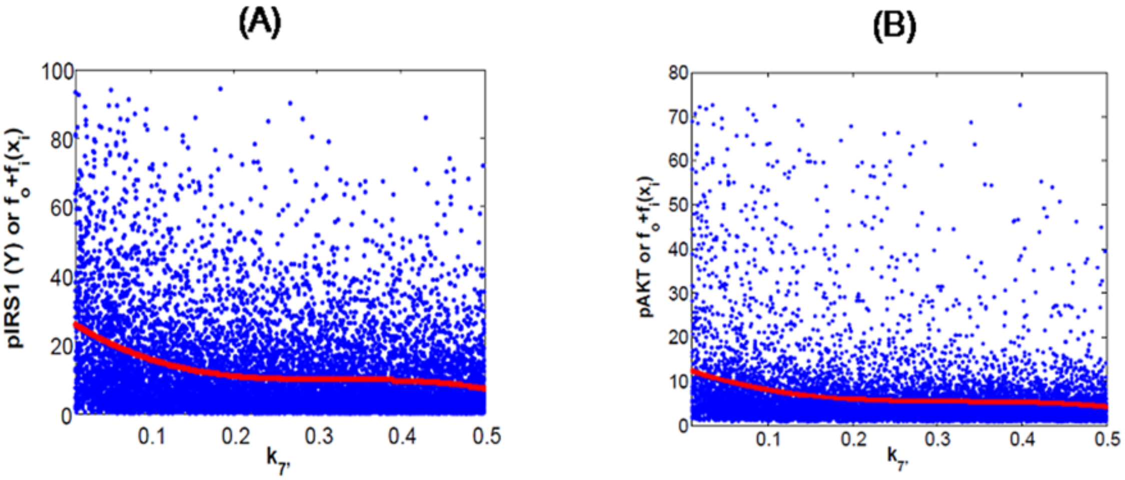


Figure A.6 Model predictions for p-IRS1 (Y) and p-AKT during perturbation of negative feedback strength.

(A) Influence on p-IRS1 (Y) output. (B) Influence on p-AKT output. The blue scatter points are the actual model output and the red curve is the first order RS-HDMR approximation. Inhibition of negative feedback (decrease in parameter $k_{7'}$) increases p-IRS1 (Y) and p-AKT levels and we also see large variability in the levels when negative feedback is weak.

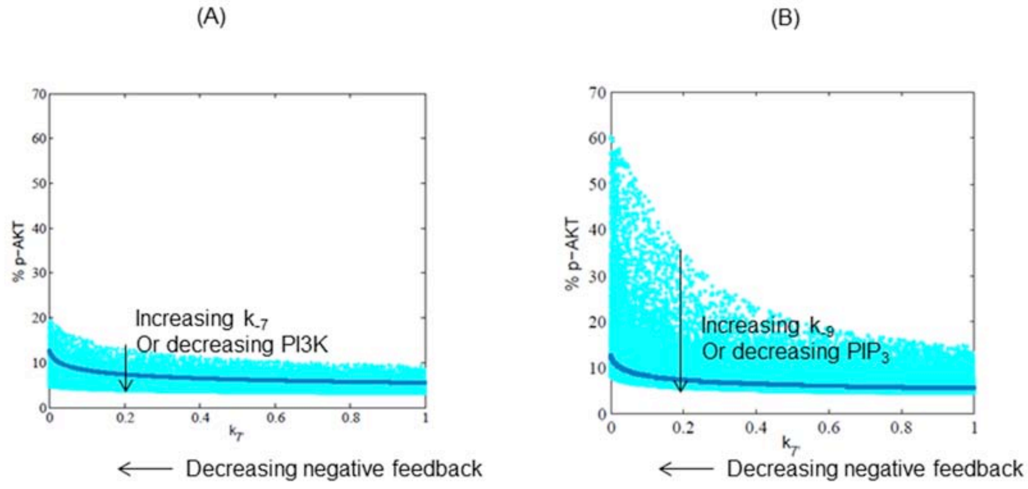


Figure A.7 Influence of perturbations in PI3K levels and PIP3 levels under varying strength of negative feedback

(A) Influence of PI3K levels. The parameter k_{-7} is perturbed to change PI3K levels. (B) Influence of PIP3 levels. The parameter k_{-9} is perturbed to change PIP3 levels. The X-axis shows the variation in the strength of negative feedback. The remaining parameters are kept at their nominal values. Changing PIP3 levels can considerably change p-AKT levels when negative feedback is weak while the variation is low when negative feedback is strong.

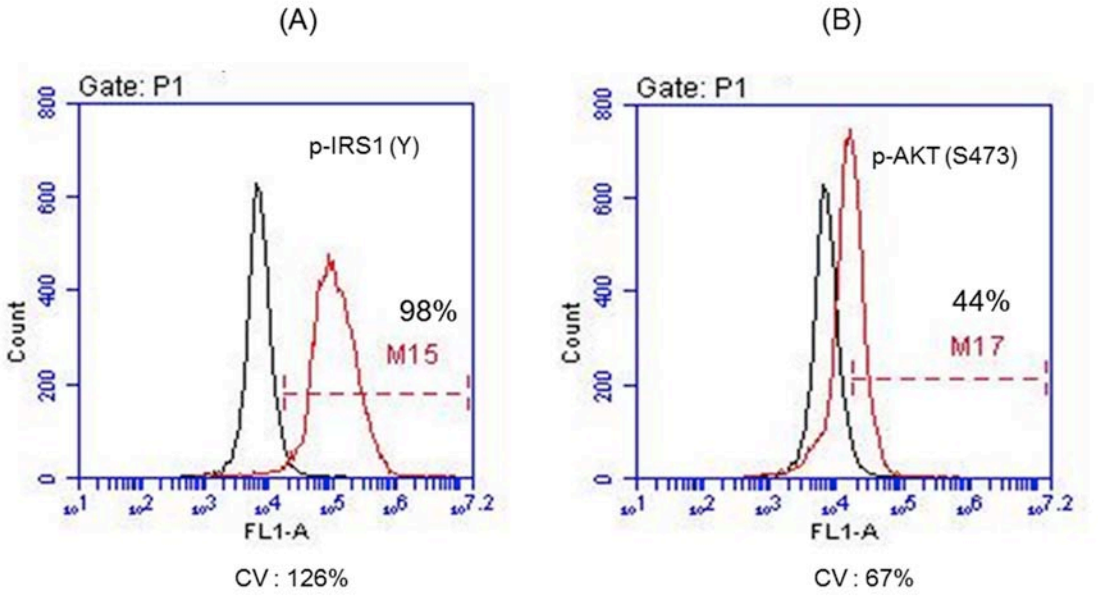


Figure A.8 Variability in phosphorylated IRS1 and AKT protein levels in hESC population.

(A) p-IRS1 (Y) levels. (B) p-AKT (S) levels. H1 cells were cultured in Activin + FGF (100 ng/ml) for 18 hr and were subjected to flow cytometry analysis as described in Task *et al.* (Task et al., 2012). Primary antibodies, rabbit anti-human p-IRS1 (pY612) and rabbit anti-human p-AKT (pS473) at 1:250 dilution were used to quantify the levels of signaling molecules. The red curve presents the positive readout within the gated region and the black curve represents the secondary antibody-only staining. p-IRS1 (Y) shows almost 98% positive staining while p-AKT shows 44% positive staining. The coefficient of variation (CV) of the curve falling in the positive region shows a large variability in p-IRS1 (Y) levels as compared to p-AKT.

APPENDIX B

CHAPTER 3 ADDITIONAL MATERIAL

Table B.1 Primers used for qRT-PCR of TF expression

MARKER (TFs)	PRIMERS FOR qRT-PCR	Reference
OCT4	CTGGGTTGATCCTCGGACCT	(D'Amour *et al.*, 2006b)
	CACAGAACTCATACGGCGGG	
CXCR4	CACCGCATCTGGAGAACCA	(D'Amour *et al.*, 2006b)
	GCCCATTTCCTCGGTGTAGTT	
SOX17	CTCTGCCTCCTCCACGAA	(Osafune *et al.*, 2008)
	CAGAATCCAGACCTGCACAA	
BRACHYURY	TGCTTCCCTGAGACCCAGTT	(D'Amour *et al.*, 2006b)
	GATCACTTCTTTCCTTTGCATCAAG	
PTF1α	GAAGGTCATCATCTGCCATCG	(D'Amour *et al.*, 2006b)
	GGCCATAATCAGGGTCGCT	
PDX1	AAGTCTACCAAAGCTCACGCG	(Kroon *et al.*, 2008)
	GTAGGCGCCGCCTGC	
CER	ACAGTGCCCTTCAGCCAGACT	(D'Amour *et al.*, 2006b)
	ACAACTACTTTTTCACAGCCTTCGT	
FOXA2 (HNF3β)	GGAGCGGTGAAGATGGAA	(Osafune *et al.*, 2008)
	TACGTGTTCATGCCGTTCAT	
GATA4	GGAAGCCCAAGAACCTGAAT	(Rust *et al.*, 2006)
	GGGAGGAAGGCTCTCACTG	
HNF1β	TCACAGATACCAGCAGCATCAGT	(Kroon *et al.*, 2008)
	GGGCATCACCAGGCTTGTA	
HNF4α	CATGGCCAAGATTGACAACCT	(Kroon *et al.*, 2008)
	TTCCCATATGTTCCTGCATCAG	
HNF6	TGTGGAAGTGGCTGCAGGA	(Zhang *et al.*, 2009b)
	TGTGAAGACCAACCTGGGCT	
GAPDH	ACGACCACTTTGTCAAGCTCATTTC	(D'Amour *et al.*, 2006b)
	GCAGTGAGGGTCTCTCTCTTCCTCT	

EFFECT OF MODEL PARAMETERS ON BICLUSTERING

In a recent work by Zhang *et al.*, the SEBI algorithm was applied to a transcriptional factor data-set of embryonic stem cells (Zhang *et al.*, 2012a). The SEBI algorithm was successful in identifying biologically relevant biclusters stable under the free parameters of the algorithm. This section elaborates the selection of free parameters of the SEBI algorithm.

GA Parameters

GA has been shown to be efficient in solving this class of NP hard problems, but the common criticism in using GA is its lack of convergence criteria and sensitivity to various search parameters. In the present simulation, a population size of 20 was used which was simulated for 700 generations, at which point no further improvement of the optimal objective was observed. A crossover probability of 0.5 and a mutation probability of 0.2 were used to maintain sufficient diversity in the population. Table B.2 summarizes all the GA parameters in detail..

Table B.2 Summary of the GA parameters

Parameter	Value
Population Size	20
Number of generations	700
Crossover probability	0.5
Mutation probability	0.2
Elitism probability	1
Weight for conditions (W_e)	1
Weight for genes (W_r)	1

Bicluster Parameters

While parameters associated with the GA formulation influences the optimal objective, there are additional parameters associated with the biclustering formulation which affects the quality of optimal bicluster. Equation 3.1 in the text details the objective function for optimizing the bicluster formulation, which consists of the following free parameters: δ, the user defined threshold on residue; W_c, W_r the relative weights associated with the columns and the rows of the bicluster respectively. The optimal bicluster obtained is significantly affected by the values of these parameters.

In order to analyze the effect of these parameters on the optimum bicluster, the optimization problem was solved at various values of δ, W_c and W_r, as summarized in Figures B.1-B.5. Figure B.1 shows the variation of the number of genes and conditions in the optimal bicluster when the threshold on the residue in varied. Very low threshold identifies smaller biclusters. For example, low values of $\delta = 0.5$ identifies optimal biclusters containing 2 genes and 2 conditions. Increasing the threshold relaxes the problem and therefore, the algorithm can search for biclusters with acceptable residue as well as larger volumes. The size of the biclusters increases with the relaxation of the threshold residue. Larger thresholds, however, compromise the quality of the bicluster, hence we select a value of $\delta = 1.5$ which gives optimal biclusters containing 3 genes and 5 conditions and acceptable residue.

Figure B.2 illustrates the effect of the relative weights on columns and rows on the volume of the identified optimal bicluster and on the number of identified genes and conditions in the bicluster. The weights W_c and W_r allow user the flexibility to bias the bicluster to include more genes or more conditions. Such flexibility is useful with prior knowledge of the structure of existing network. Comparing Figures B.2 (A), (C), it is found that the bicluster volume does not change appreciably with changes in row weight and the column weight, the volume increases from 10 to 30 when changing W_r from 0.5 to 2 while it changes from 12 to 24 over the same range for W_c. For W_c greater than 2, we see rapid increase in the number of conditions because the search is sensitive to W_c. Figure B.2 (B), (D) further breaks up the volume into genes and conditions and illustrates how it changes in the number of both genes and conditions with W_c and W_r respectively. We find that increasing the column (row) weight increases the number of conditions (genes) while the number of genes (conditions) remain almost constant until W_c (W_r) = 2.

Figure B.3 shows the effect of row and column weights on the residue of the bicluster for a fixed threshold value of δ = 1.5. Changing the row weights is found to increase the residue appreciably. However, the residue is found to be less sensitive to the column weights. It is interesting to note that the residue is never found to be higher than the threshold even though this check was not explicitly introduced in the formulation.

Following above analysis, we chose the value of δ = 1.5 in order to capture reasonable volume of the bicluster. Regarding the weights W_r and W_c, in the absence of prior knowledge regarding the structure of expected bicluster, all were chosen to be on the lower end of 1.

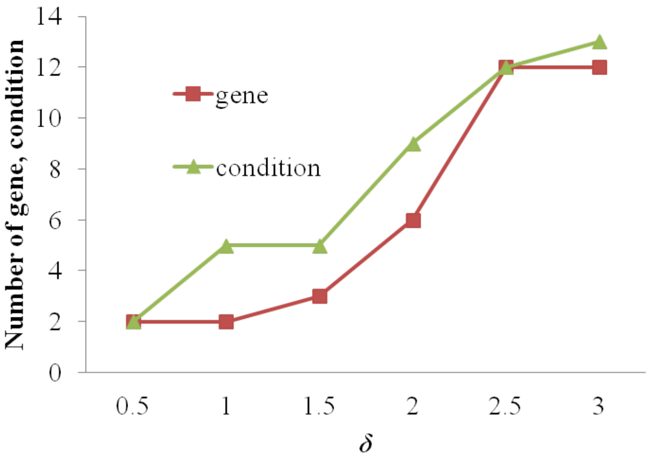


Figure B.1 Variation of number of genes and conditions in the optimal bicluster with different values of the threshold, δ

Increasing the threshold increases the number of genes and conditions contained by the optimal bicluster. A rapid increase in the number of TFs and conditions is observed after a δ of 1.5.

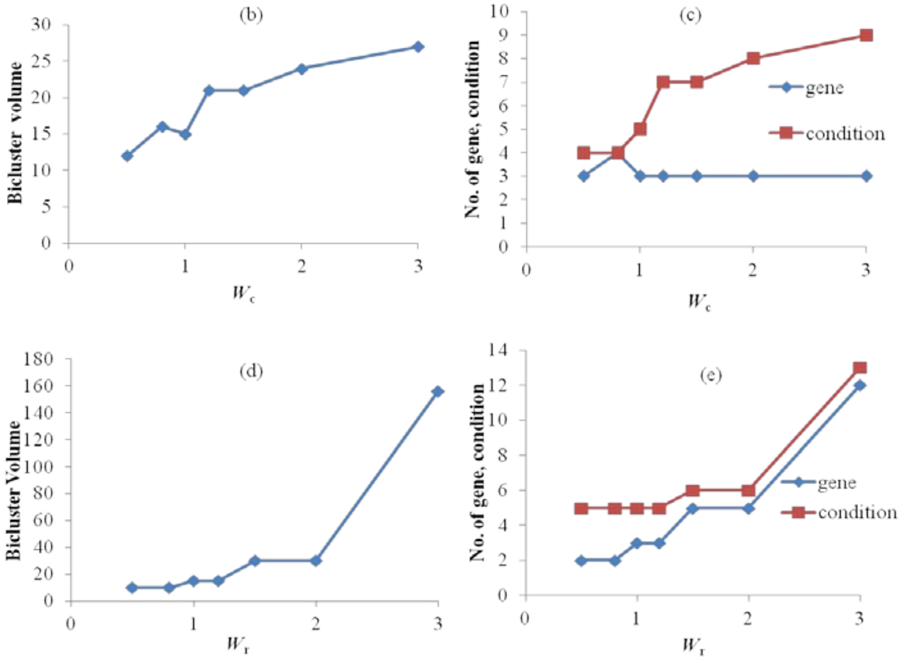


Figure B.2 Effect of model parameters on features of optimal bicluster.

Variation in the bicluster volume and the number of genes, conditions in the optimal bicluster with changes in the column weights (A-B) and row weights (C-D) respectively.

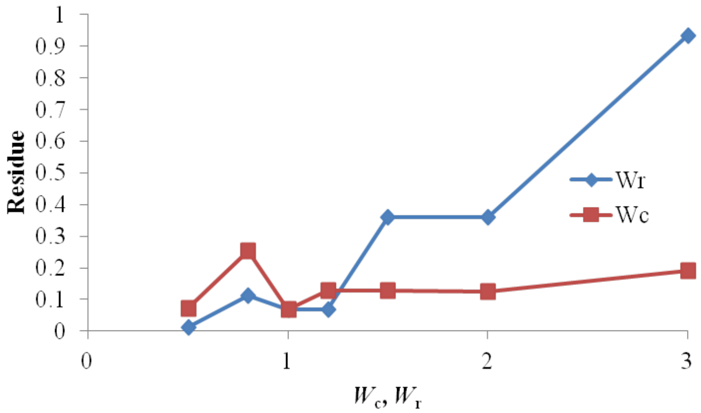


Figure B.3 Variation of the residue as a function of row and column weights.

The residue is found to be sensitive to the row weights. All the residues remain well within the threshold limit of δ = 1.5.

Effect of model parameters on the robust subsets

While the bootstrap + biclustering algorithm enables determination of biclusters which remain robust to experimental noise, these are still evaluated for certain specific values of model parameters. Hence to analyze its sensitivity to the model parameters, the entire procedure was repeated for different values of model parameters: the threshold on residue (δ); row weight (W_r) and column weights (W_c). The frequency of occurrence of the two groups was subsequently measured by changing the parameter values, as illustrated in Figure B.4 (A-C). Figure B.4 (A) shows the variation in the frequency of occurrence of the robust bicluster for varying values of δ. It was observed that for a broad range of the threshold the subsets are being repeated over 50% of time. Also, for low values of δ, the number of repeats of Group 1 remains almost constant indicating that it is indeed robust. At larger values of δ, the number of repeats for this group decrease and Group 2 takes over. Higher values of δ relax the constraint on the residue of the biclusters and therefore, increase the volume of the biclusters and the residue. Hence, this increases the occurrence of other genes and conditions in the biclusters and therefore, we see a decrease in the number of repeats for Group 1. It is interesting to note that the number of repeats for Group 2 increases with delta indicating that it is possibly the next robust subset present in the array but has higher residue as compared to Group 1. Figure B.4 (B) shows the variation in the number of repeats with the column weight. Again, we see that the number of repeats for Group 1 goes through a maximum at 1.0 and on average stays above 500. At lower W_c, the biclusters are very small and therefore, the subsets are repeated fewer number of times and the repeats increase with W_c. However, when the δ crosses 1.0, the larger biclusters tend to have relatively high residue and thus contain mostly genes-condition groups with less similar profiles. Thus, we see a decrease in the repeats at larger W_c. Again, we note that Group 2 subset occurs more frequently with increase in W_c. Figure B.4 (C) shows the variation in the number of repeats with the row weight.

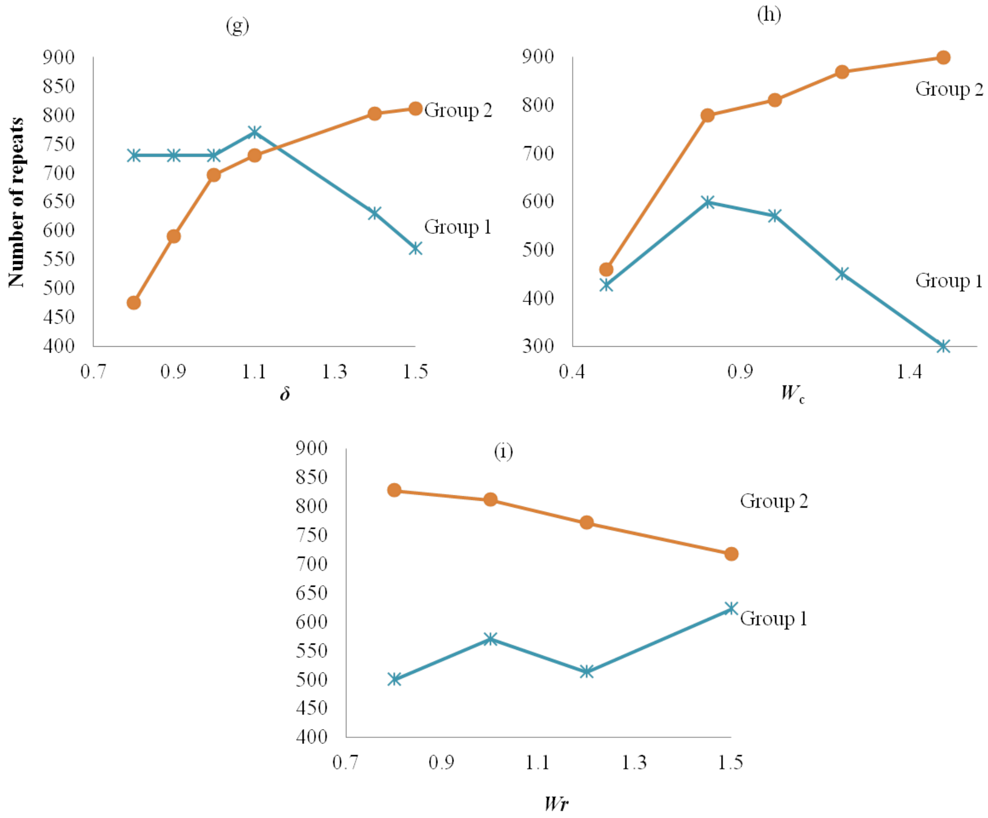


Figure B.4 Sensitivity of the identified robust bicluster on model parameters.

Biclustering of the bootstrap data identifies 2 groups of robust bi-clusters. The figure illustrates the number of repeats of these robust bi-clusters with changes in (A) threshold, δ (B) column weight, W_c. (C) row weight, $W_{r.}$

APPENDIX C

CHAPTER 4 ADDITIONAL MATERIAL

DBN INFERENCE

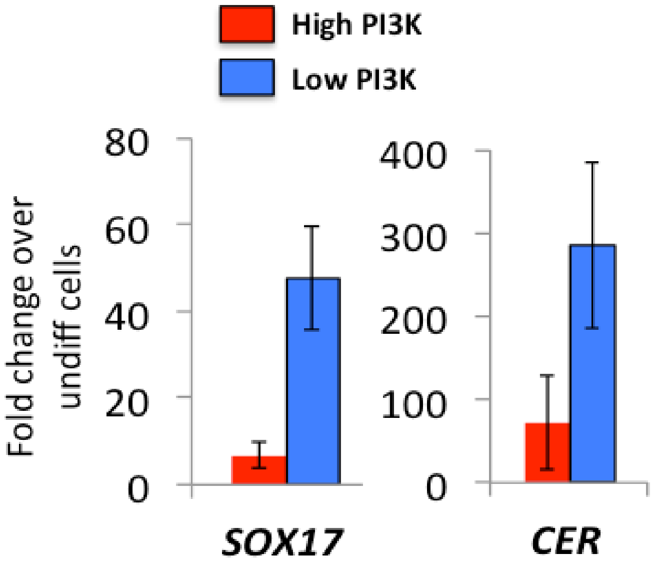


Figure C.1 Endoderm markers *SOX17* and *CER1* at day 4 of differentiation under high and low PI3K.

The fold change is calculated over undifferentiated hESCs (number of repeats = 3).

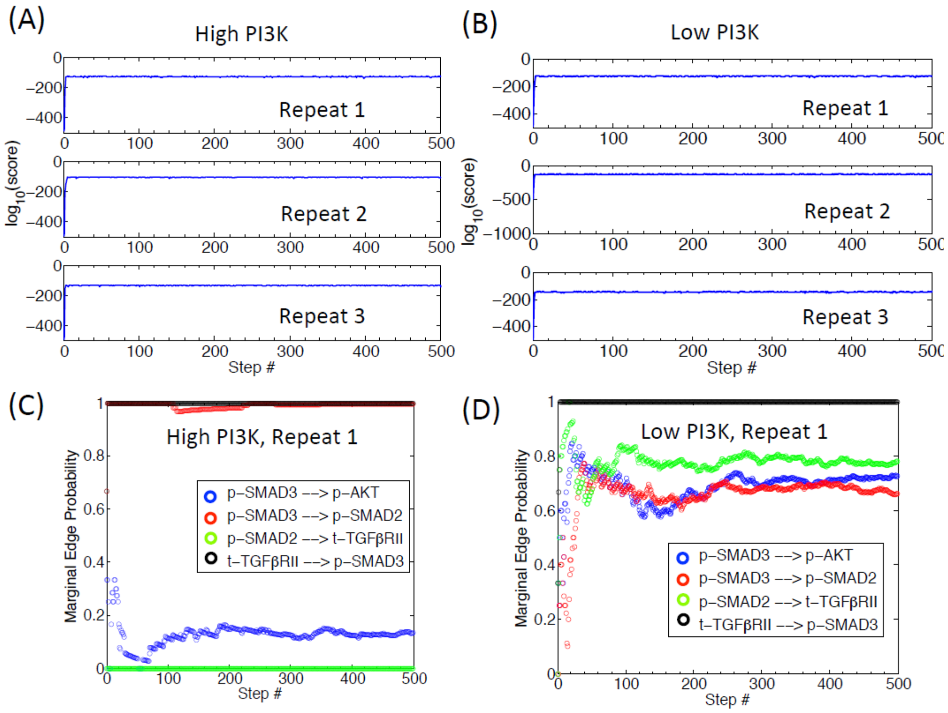


Figure C.2 Convergence characteristics for DBN inference of entire time series.

(A-B) Log likelihood score for each repeat of high PI3K and low PI3K. (C-D) Marginal edge probabilities for selected nodes in the two conditions for each successive Gibbs sampling step. The marginal edge probability at a given step was calculated by using later half of the samples until that step. It is evident from the edges presented in this plot (as well as those not shown here) that the probabilities converge to the mean value by 250 steps.

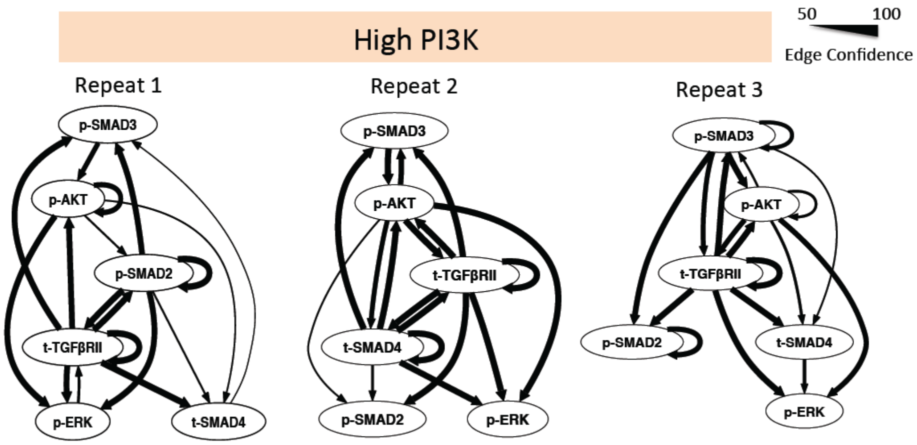


Figure C.3 DBNs for individual repeats in high PI3K condition over the entire time series data

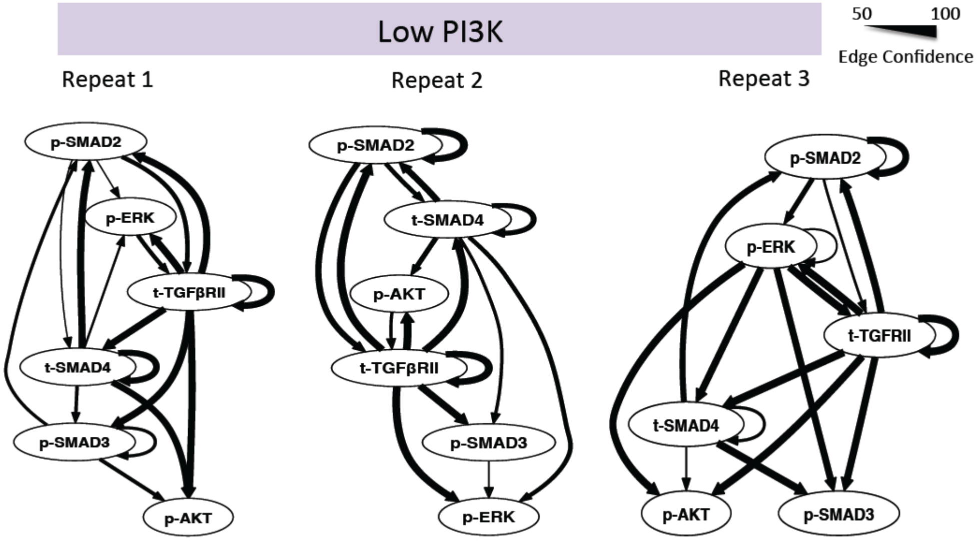
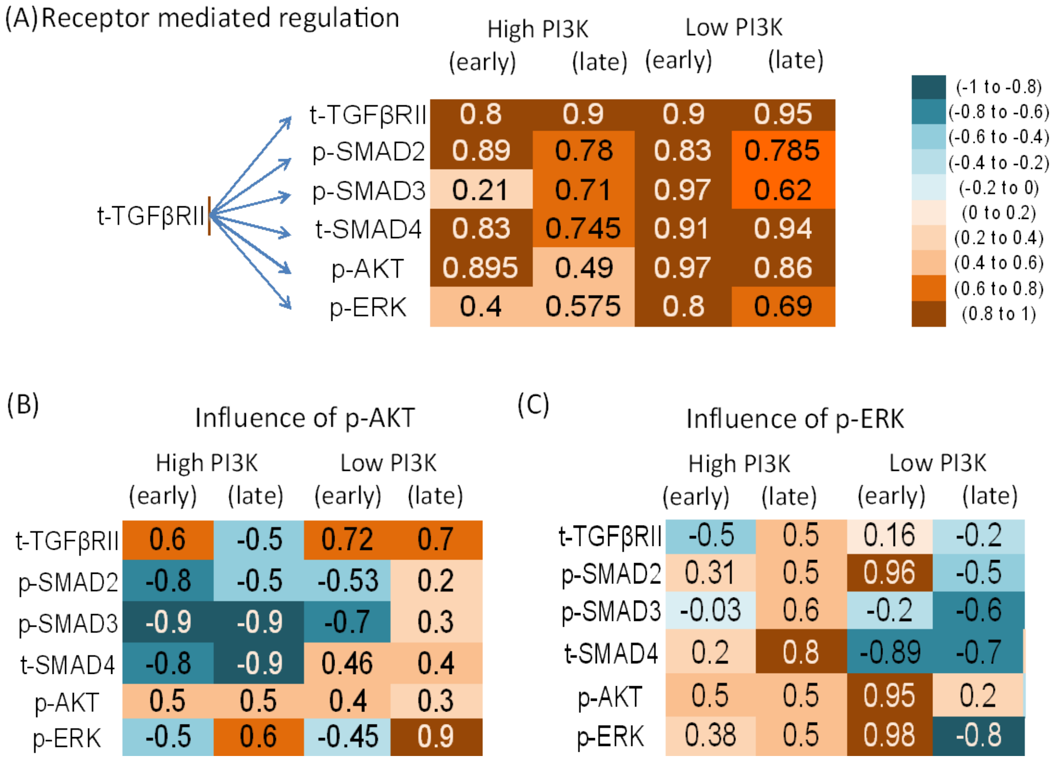


Figure C.4 DBNs for individual repeats in low PI3K condition over the entire time series data.



(A) Receptor mediated regulation

	High PI3K (early)	High PI3K (late)	Low PI3K (early)	Low PI3K (late)
t-TGFβRII	0.8	0.9	0.9	0.95
p-SMAD2	0.89	0.78	0.83	0.785
p-SMAD3	0.21	0.71	0.97	0.62
t-SMAD4	0.83	0.745	0.91	0.94
p-AKT	0.895	0.49	0.97	0.86
p-ERK	0.4	0.575	0.8	0.69

(-1 to -0.8)
(-0.8 to -0.6)
(-0.6 to -0.4)
(-0.4 to -0.2)
(-0.2 to 0)
(0 to 0.2)
(0.2 to 0.4)
(0.4 to 0.6)
(0.6 to 0.8)
(0.8 to 1)

(B) Influence of p-AKT

	High PI3K (early)	High PI3K (late)	Low PI3K (early)	Low PI3K (late)
t-TGFβRII	0.6	-0.5	0.72	0.7
p-SMAD2	-0.8	-0.5	-0.53	0.2
p-SMAD3	-0.9	-0.9	-0.7	0.3
t-SMAD4	-0.8	-0.9	0.46	0.4
p-AKT	0.5	0.5	0.4	0.3
p-ERK	-0.5	0.6	-0.45	0.9

(C) Influence of p-ERK

	High PI3K (early)	High PI3K (late)	Low PI3K (early)	Low PI3K (late)
t-TGFβRII	-0.5	0.5	0.16	-0.2
p-SMAD2	0.31	0.5	0.96	-0.5
p-SMAD3	-0.03	0.6	-0.2	-0.6
t-SMAD4	0.2	0.8	-0.89	-0.7
p-AKT	0.5	0.5	0.95	0.2
p-ERK	0.38	0.5	0.98	-0.8

Figure C.5 Correlation tables for high and low PI3K condition.

(A) Receptor mediated regulation; (B) p-AKT mediated regulation; (C) p-ERK mediated regulation. The Pearson correlation is calculated between the parent nodes at time step (t 1) and all other nodes at time step t. The early time points 0.5, 1, 1.5 h (both conditions) and the late time points correspond to 6, 12, and 18 for high PI3K and 12, 18, 24 for low PI3K.

KINETIC ANALYSIS ADDITIONAL FIGURES

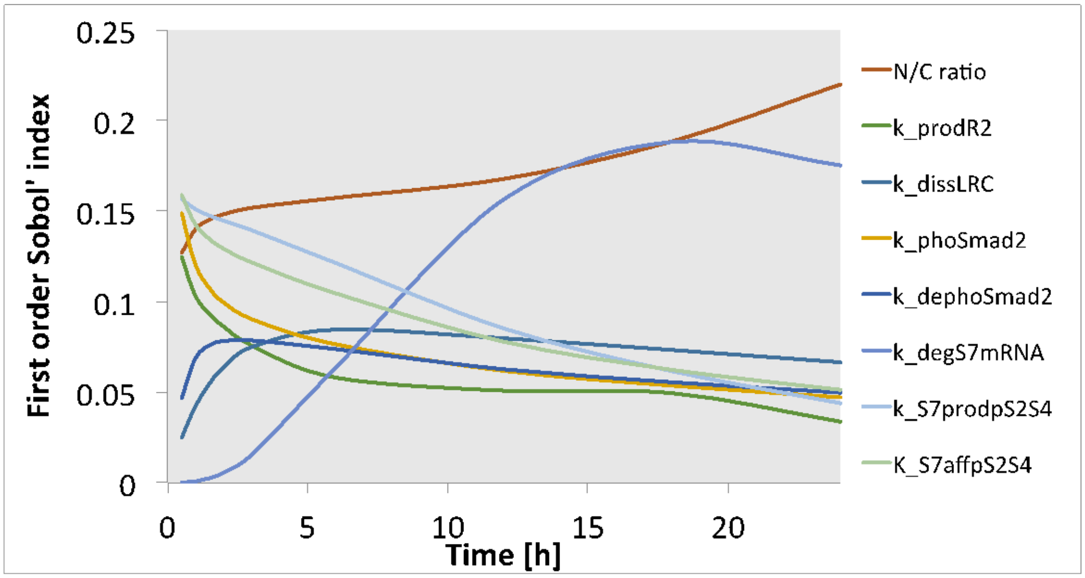


Figure C.6 Sobol' indices for parameters controlling SMAD7 mRNA levels with time

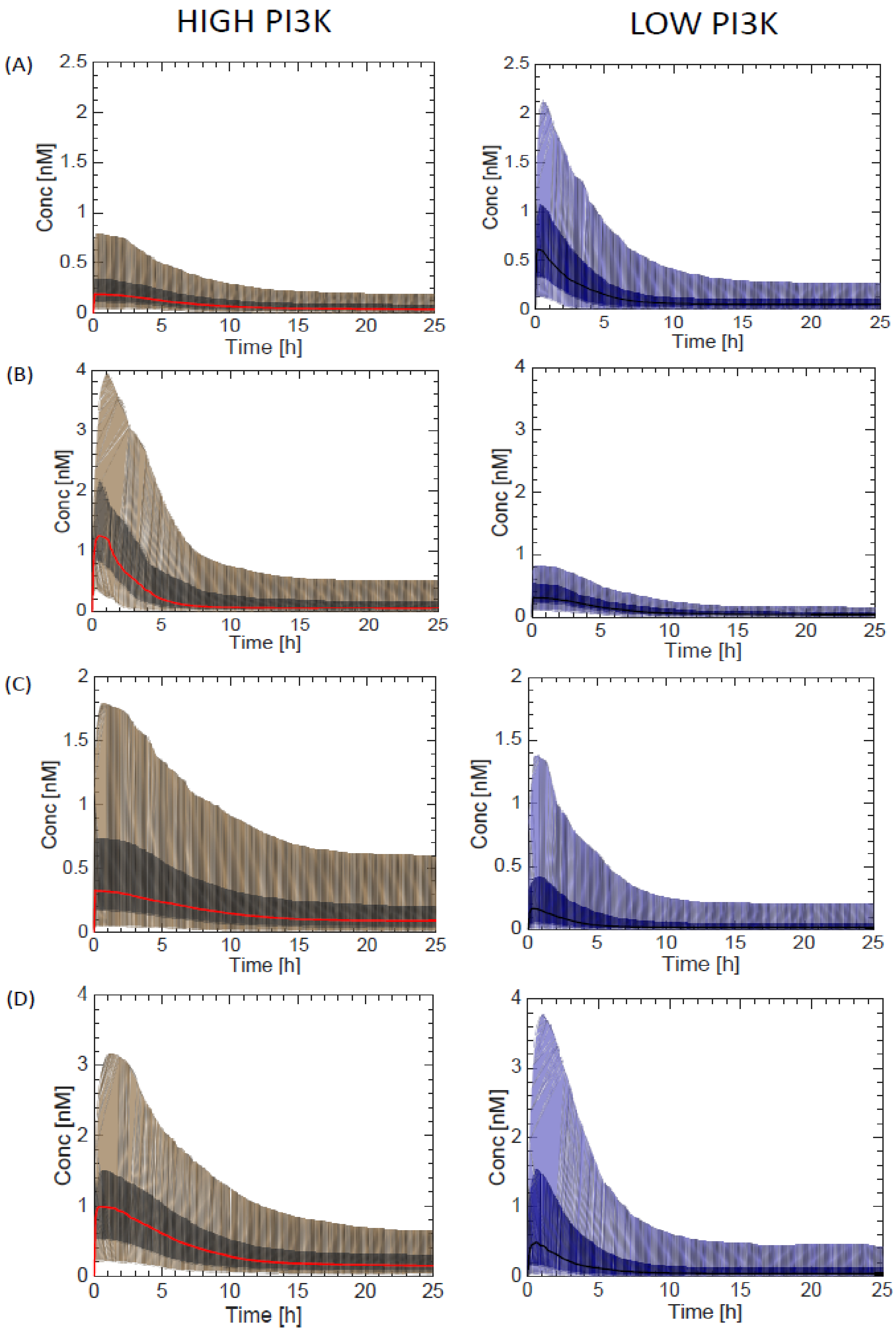


Figure C.7 Endosomal LRC levels in high and low PI3K condition.

Panels are for (A) Crosstalk 1, (B) Crosstalk 2, (C) Crosstalk 3, (D) Crosstalk 1+3.

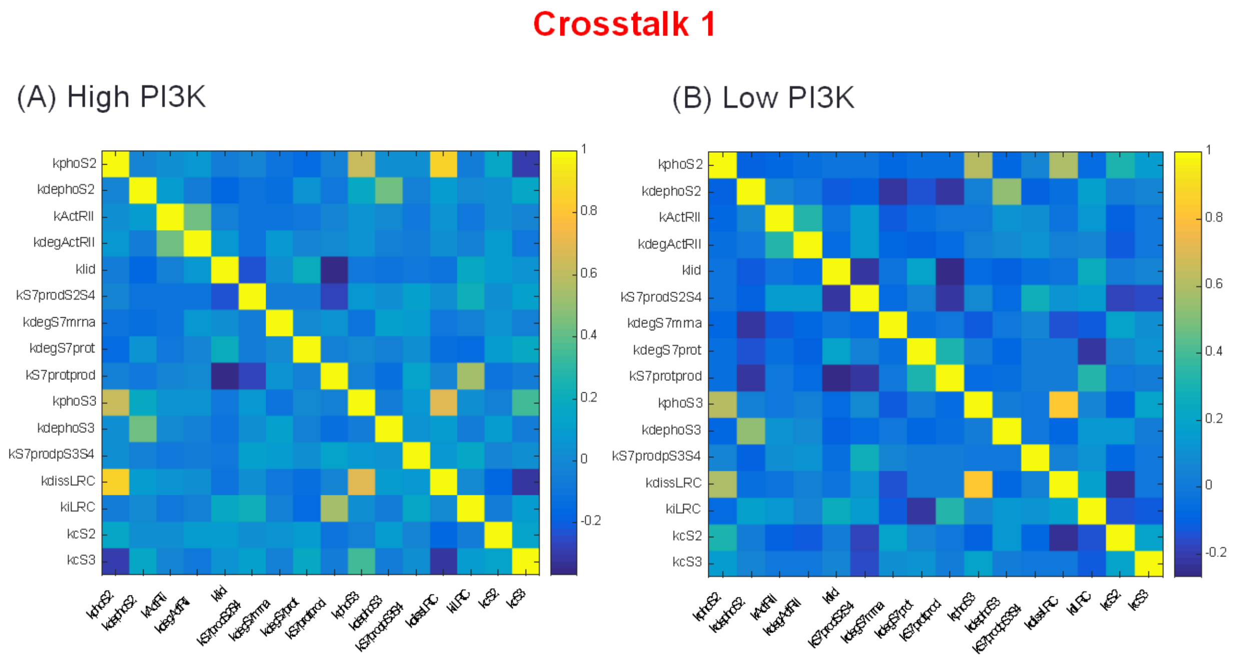


Figure C.8 Pearson correlation plots for Crosstalk 1.

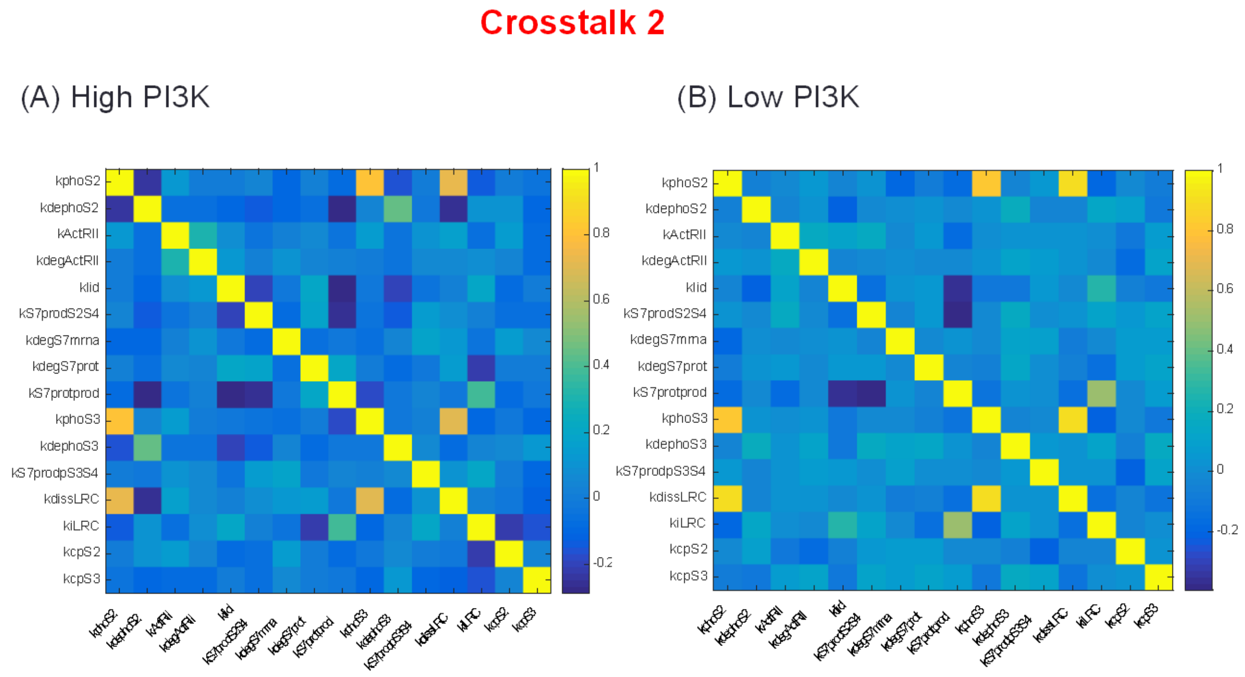


Figure C.9 Pearson correlation plots for Crosstalk 2.

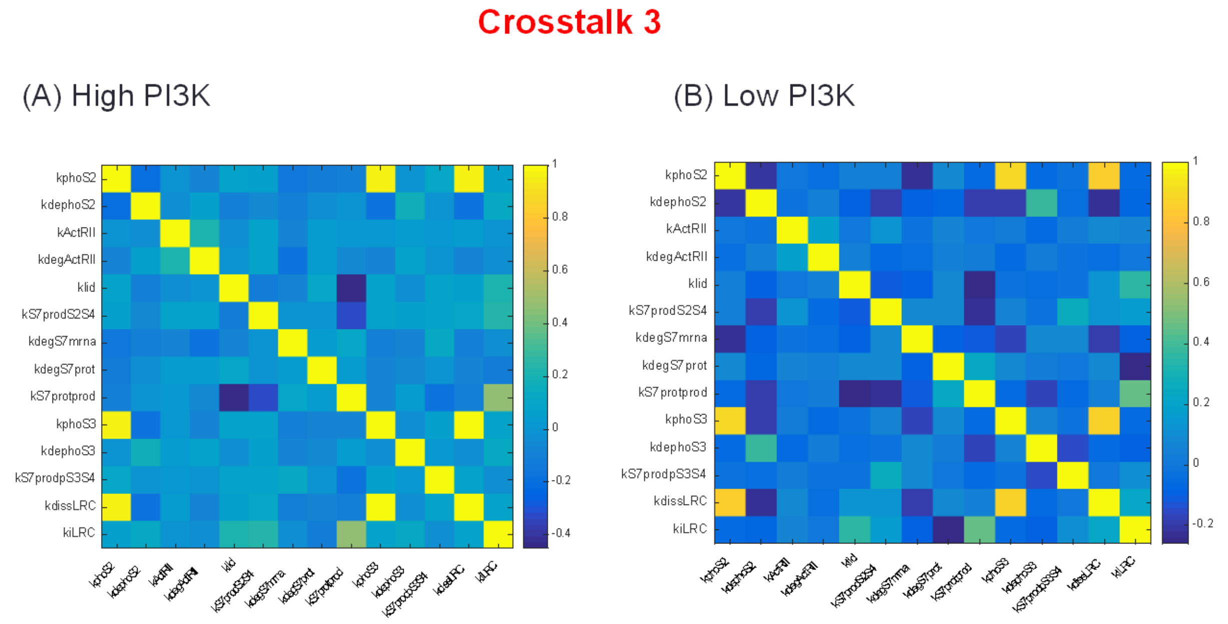


Figure C.10 Pearson correlation plots for Crosstalk 3.

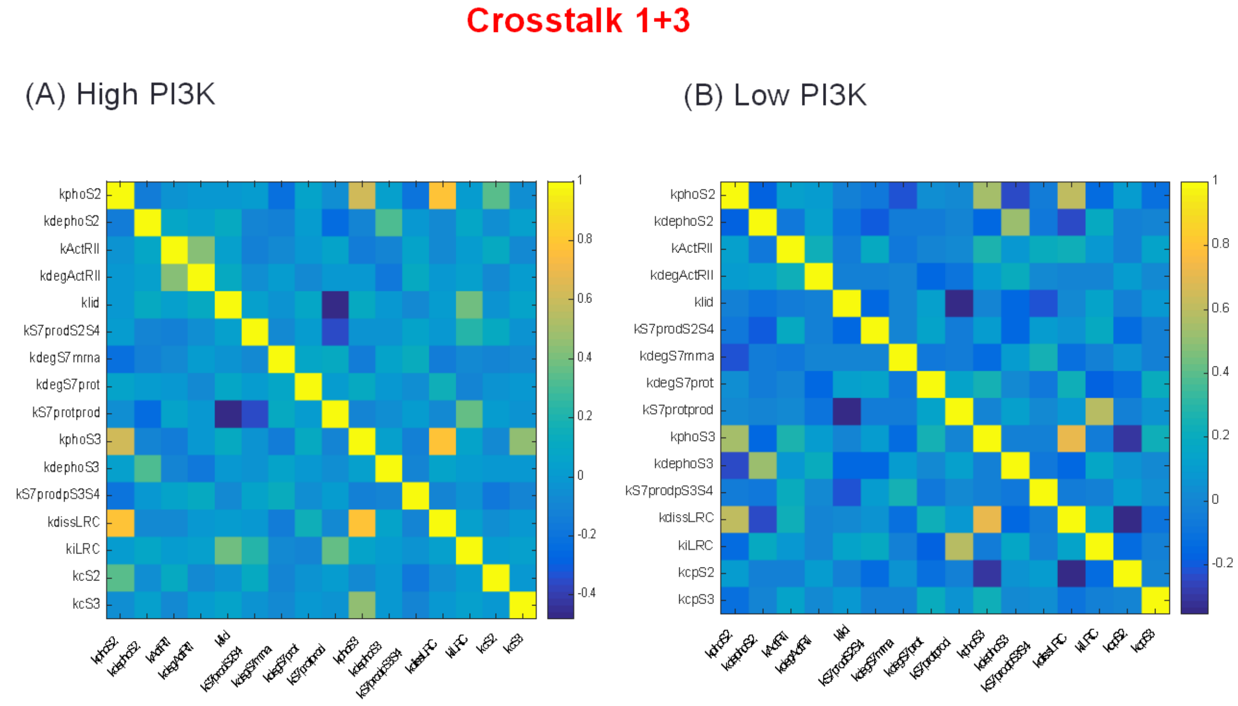


Figure C.11 Pearson correlation plots for Crosstalk 1+3.

APPENDIX D

ANALYZING PROTEIN MOBILITY USING FRAP TECHNIQUE

BACKGROUND

FRAP technique is commonly used to measure the kinetics of protein diffusion in cells, for example movement of proteins in the membrane or within and between cellular compartments (Phair and Misteli, 2001). Tagging proteins with small fluorescent molecules like GFP makes them visible by light microscopy. Combining laser induced photobleaching and confocal laser scanning microscopy enables qualitative and quantitative analysis of various features of protein mobility, for example diffusivity, compartmental import and export rates and association and dissociation constants.

METHODS: EXPERIMENTAL

In a typical FRAP experiment; fluorescence in a small area of the cell (called Region of Interest or ROI) where diffusion is to be measured is repeatedly bleached using a high intensity laser (Figure D.1A). After this bleaching episode, the movement of unbleached molecules from the neighboring regions into the ROI is recorded by time-lapse microscopy. This data gives the recovery curve of fluorescence in the ROI. Proteins may enter the ROI by pure diffusion or by active transport via transporter molecules. Further, during the transport process, the protein may encounter binding sites and the equilibrium kinetics of binding influence the overall recovery of fluorescence in the ROI. Standard confocal microscopes are suitable for the FRAP technique;

since they are equipped with an acousto-optical tunable filter, which allows for rapid switching of the laser power from low intensity for imaging the cells to high intensity for bleaching the desired area and back again (Davies *et al.*, 2010). Integration of mathematical methods (conservation equations) to the recovery curve and the spatial fluorescence distribution (proportional to the concentration distribution) enables estimation of the diffusivity coefficient of the protein in and around the ROI.

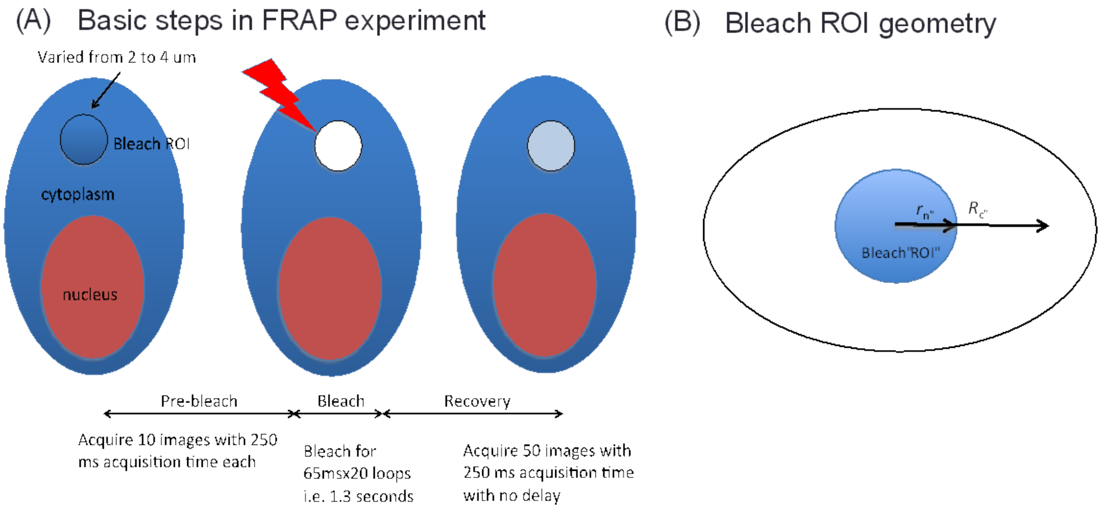

(A) Basic steps in FRAP experiment
Varied from 2 to 4 um
Bleach ROI
cytoplasm
nucleus
Pre-bleach
Bleach
Recovery
Acquire 10 images with 250 ms acquisition time each
Bleach for 65msx20 loops i.e. 1.3 seconds
Acquire 50 images with 250 ms acquisition time with no delay
(B) Bleach ROI geometry
r_n"
R_c"
Bleach"ROI"

Figure D.1 FRAP experiments and ROI geometry.

(A) Three major stages in a typical FRAP experiment. (B) Circular bleach geometry showing the bleach radius and effective radius of bleach (defined using Gaussian fit of Equation D.1 to the fluorescence data along a middle strip of the ROI).

A.2 EVALUATING DIFFUSION COEFFICIENTS USING MATHEMATICAL MODELING

Consider a circular ROI with radius r_n bleached at time 0 (Figure D.1B). Although bleaching is done only in the ROI, there is movement of bleached molecules into the surrounding region during the bleaching process. Therefore, the actual area of bleaching is greater than the ROI. The effective radius of bleaching (r_e) is empirically determined from the first FRAP image after bleaching. Note that during the bleaching process, no image is taken as the laser is in the bleaching mode. Microscope parameters decide how long it takes for switching from bleaching to acquisition mode. The radial distribution of fluorescence (F) just after bleach (due to instantaneous diffusion) usually takes an inverted bell shaped curve, given by:

$$F(r,t=0)=f(r)=F(r,i)\left[1-K\exp\left[\frac{-2r^2}{r_e^{\ 2}}\right]\right] \qquad \text{Equation (D.1)}$$

Here K represents the bleach depth parameter and $F(r,i)$ represents the pre-bleach fluorescence at a given radial location. The fluorescence values are normalized to remove background fluorescence (F_{bkgd}) and also net loss of fluorescence due to the bleaching process itself. The normalization is performed using the following relation:

$$u(ROI,t)=\frac{F(ROI,t)-F_{bkgd}}{F(cell,t)-F_{bkgd}}\times\frac{F(cell,i)-F_{bkgd}}{F(ROI,i)-F_{bkgd}} \qquad \text{Equation (D.2)}$$

Here, $F(cell,i)$ is the intensity of the entire cell and $F(ROI,i)$ is the intensity of the ROI. Similar normalization is performed on the fluorescence intensities at different radial locations and time points. For a pure diffusion process, the concentration distribution at different radial locations evolves in time according to the diffusion equation:

$$D\left[\frac{\partial^2 u}{\partial r^2}+\frac{1}{r}\frac{\partial u}{\partial r}\right]=\frac{\partial u}{\partial t},\ 0\le r\le R_c \qquad \text{Equation (D.3)}$$

The initial condition is given by the Gaussian curve:

$$u(r,0)=f(r)=\left[1-K\exp\left[\frac{-2r^2}{r_e^{\ 2}}\right]\right]$$

At sufficient distance from the center of the ROI, the change in concentration with distance is negligible leading to the following boundary condition at all times:

$$At r = R_c, \frac{\partial u}{\partial r} = 0, t > 0$$

Solution to this PDE is given by:

$$u(r,t) = A_0 + \sum_{n=1}^{\infty} A_n J_0(\alpha_n r) e^{-\alpha^2{}_n Dt} \qquad \text{Equation (D.4)}$$

Here,

$$A_0 = \frac{2}{R^2{}_c} \int_0^{R_c} r\, f(r) dr$$

$$A_n = \frac{2}{R^2{}_c J^2{}_0(\alpha_n R_c)} \int_0^{R_c} r J_0(\alpha_n r) f(r) dr$$

$$f(r) = \exp\left[-K \exp\left[\frac{-2r^2}{r_e{}^2}\right]\right]$$

The Bessel functions ($J_0(\alpha_n R_c)$) are fixed according to the boundary condition:

$$J_0'(\alpha_n R_c) = 0$$

$\alpha_n = \dfrac{x_n}{R_c}$, where x_n is a positive root of $J_0'(x_n) = 0$. The first three positive roots are

x_n : 3.8317, 7.0156 and 10.1735.

The average intensity in the ROI and its evolution with time can be obtained using the relation:

$$u_{ROI}(t) = \frac{\iint u(r,t) \times 2\pi r dr dz}{\iint 2\pi r dr dz} = \frac{\int_{r=0}^{r_n} u(r,t) \times 2\pi r dr}{\int_{r=0}^{r_n} 2\pi r dr} \qquad \text{Equation (D.5)}$$

In the above equation, the intensity profile from Equation (D.4) can be used to obtain an expression for $u_{ROI}(t)$ as a function of the diffusivity. This analytical expression can be fitted to the experimental data to obtain an estimate of the diffusivity coefficient using a least squares error function:

$$\min_{D} \sum_{i=1}^{\#time\,points} \frac{1}{\#time\,points}\left[u_{ROI}(t_i) - f(t_i)\right]^2 \qquad \text{Equation (D.6)}$$

A.3 EXAMPLES OF DIFFUSIVITY MEASUREMENTS

We estimated the diffusivity for two cases; freely diffusing GFP and GFP fused Profilin1 protein in MDA-MB-231 cancer cell line and obtained values close to the literature reports (Table D.1). The diffusivity values were in the same order of magnitude as the literature and the differences are mainly due to the increase in protein size from GFP to GFP-Profilin1. Figure D.2 shows the results and model fits and predictions for GFP-Profilin1.

Table D.1 Diffusivity coefficients

Protein	Diffusion coefficient (D, μm^2/s)	
	Our model	Literature value (Reference)
GFP	30	20-50 (Kang *et al.*, 2012)
GFP-Profilin1	3	5 (Novak *et al.*, 2008)

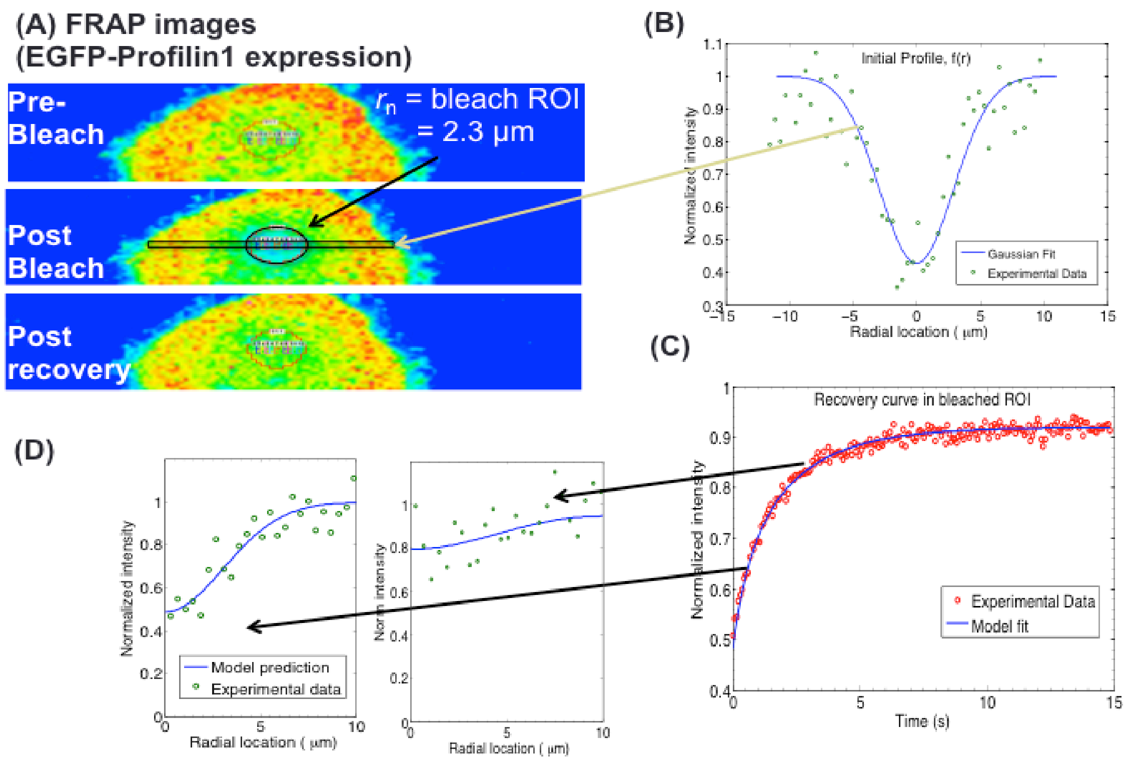


Figure D.2 FRAP analysis to estimate intracellular diffusion coefficient.

(A) Pseudo-color FRAP images acquired using NIS-Elements and confocal laser scanning microscope. The intensity increases from blue to red. FRAP was performed to measure EGFP-Profilin1 diffusion coefficient in MDA-MB-231 breast cancer cells. Bleach parameters: 488 nm laser at 100 % power, 1.3 s bleach time. Image acquisition is performed at 1% laser power and 200 ms acquisition time. (C) Initial radial bleach profile smoothed with a gaussian profile. (D) Recovery curve with time in the bleached ROI. Model output from Equation D.5 is fitted to the experimental data in this region with time. (E) Model predictions of the radial intensity profile with time (Equation D.4) compared to non-fitted experimental data. The intensity profile for 149 ms and 3 s after bleach are shown.

SIGNIFICANCE TO SMAD SIGNALING

FRAP can be used to quantify the influence of complexation of SMAD2,3 either with SMADs or crosstalk molecules like AKT on the transport properties of SMADs. Studies have looked at the diffusivity of GFP-SMAD2 under no stimulation (therefore monomeric SMAD2) and under stimulation (therefore mixed/effective diffusivity due to various proportions of complexed

SMAD2) (Gonzalez-Perez *et al.*, 2011). Similar analysis can be done for GFP-SMAD3, which has not been analyzed before. Further the variation of diffusivity near the membrane and in the cytoplasm under high and low PI3K signaling can be used to test the strength of the sequestration effect by AKT. On the other hand, simple compartmental models can be used to describe recovery in the nuclear fluorescence after complete bleaching of the nucleus. These FRAP configurations have been used to estimate the import and export rates of SMAD2 and SMAD4 (Nicolas *et al.*, 2004; Schmierer and Hill, 2005). Future experiments using such integrated modeling and experimental analysis will be useful to thoroughly characterize the differences in SMAD2 and SMAD3 and also study the cell-to-cell variability in processes involving these proteins.

BIBLIOGRAPHY

Aerts, J. M., Haddad, W. M., An, G., Vodovotz, Y., 2014. From data patterns to mechanistic models in acute critical illness. J Crit Care 29, 604-10, doi:10.1016/j.jcrc.2014.03.018.

Akaike, H., 1998. Information theory and an extension of the maximum likelihood principle. Selected Papers of Hirotugu Akaike. Springer, pp. 199-213

Aksamitiene, E., Kiyatkin, A., Kholodenko, B. N., 2012. Cross-talk between mitogenic Ras/MAPK and survival PI3K/Akt pathways: a fine balance. Biochemical Society Transactions 40, 139-146, doi:Doi 10.1042/Bst20110609.

Albeck, J. G., MacBeath, G., White, F. M., Sorger, P. K., Lauffenburger, D. A., Gaudet, S., 2006. Collecting and organizing systematic sets of protein data. Nat Rev Mol Cell Biol 7, 803-12, doi:10.1038/nrm2042.

Aldridge, B. B., Burke, J. M., Lauffenburger, D. A., Sorger, P. K., 2006. Physicochemical modelling of cell signalling pathways. Nat Cell Biol 8, 1195-203, doi:10.1038/ncb1497.

Allegrucci, C., Young, L. E., 2007. Differences between human embryonic stem cell lines. Human reproduction update 13, 103-120

Ao, A., Hao, J., Hong, C. C., 2011. Regenerative chemical biology: current challenges and future potential. Chem Biol 18, 413-24, doi:10.1016/j.chembiol.2011.03.011.

Atala, A., Lanza, R., Thomson, J. A., Nerem, R., 2010. Principles of regenerative medicine. Academic Press.

Atkinson, S. P., Lako, M., Armstrong, L., 2013. Potential for pharmacological manipulation of human embryonic stem cells. Br J Pharmacol 169, 269-89, doi:10.1111/j.1476-5381.2012.01978.x.

Attisano, L., Wrana, J. L., Montalvo, E., Massague, J., 1996. Activation of signalling by the activin receptor complex. Mol Cell Biol 16, 1066-73.

Azhar, N., Ziraldo, C., Barclay, D., Rudnick, D. A., Squires, R. H., Vodovotz, Y., Pediatric Acute Liver Failure Study, G., 2013. Analysis of serum inflammatory mediators identifies unique dynamic networks associated with death and spontaneous survival in pediatric acute liver failure. PLoS One 8, e78202, doi:10.1371/journal.pone.0078202.

Bancaud, A., Huet, S., Rabut, G., Ellenberg, J., 2010. Fluorescence perturbation techniques to study mobility and molecular dynamics of proteins in live cells: FRAP, photoactivation, photoconversion, and FLIP. Cold Spring Harb Protoc 2010, pdb top90, doi:10.1101/pdb.top90.

Basma, H., Soto-Gutierrez, A., Yannam, G. R., Liu, L., Ito, R., Yamamoto, T., Ellis, E., Carson, S. D., Sato, S., Chen, Y., Muirhead, D., Navarro-Alvarez, N., Wong, R. J., Roy-Chowdhury, J., Platt, J. L., Mercer, D. F., Miller, J. D., Strom, S. C., Kobayashi, N., Fox, I. J., 2009. Differentiation and transplantation of human embryonic stem cell-derived hepatocytes. Gastroenterology 136, 990-9, doi:10.1053/j.gastro.2008.10.047.

Bernardo, A. S., Faial, T., Gardner, L., Niakan, K. K., Ortmann, D., Senner, C. E., Callery, E. M., Trotter, M. W., Hemberger, M., Smith, J. C., 2011. BRACHYURY and CDX2 mediate BMP-induced differentiation of human and mouse pluripotent stem cells into embryonic and extraembryonic lineages. Cell stem cell 9, 144-155.

Bessonnard, S., De Mot, L., Gonze, D., Barriol, M., Dennis, C., Goldbeter, A., Dupont, G., Chazaud, C., 2014. Gata6, Nanog and Erk signaling control cell fate in the inner cell mass through a tristable regulatory network. Development 141, 3637-3648, doi:10.1242/dev.109678.

Birtwistle, M. R., Kolch, W., 2011. Biology using engineering tools The negative feedback amplifier. Cell Cycle 10, 2069-2076, doi:10.4161/cc.10.13.16245.

Bleuler, S., Prelic, A., Zitzler, E., 2004. An EA framework for biclustering of gene expression data. Vol. 1. IEEE, pp. 166-173

Brooks, S. P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. Journal of computational and graphical statistics 7, 434-455

Brown, K. S., Sethna, J. P., 2003. Statistical mechanical approaches to models with many poorly known parameters. Phys Rev E Stat Nonlin Soft Matter Phys 68, 021904, doi:10.1103/PhysRevE.68.021904.

Bruce, D. L., Sapkota, G. P., 2012. Phosphatases in SMAD regulation. FEBS Lett 586, 1897-905, doi:10.1016/j.febslet.2012.02.001.

Cahan, P., Li, H., Morris, S. A., Lummertz da Rocha, E., Daley, G. Q., Collins, J. J., 2014. CellNet: network biology applied to stem cell engineering. Cell 158, 903-15, doi:10.1016/j.cell.2014.07.020.

Cheng, Y., Church, G. M., 2000. Biclustering of expression data. Vol. 8, pp. 93-103.

Chickarmane, V., Peterson, C., 2008. A computational model for understanding stem cell, trophectoderm and endoderm lineage determination. PLoS One 3, e3478, doi:10.1371/journal.pone.0003478.

Chickarmane, V., Troein, C., Nuber, U. A., Sauro, H. M., Peterson, C., 2006. Transcriptional dynamics of the embryonic stem cell switch. PLoS Comput Biol 2, e123, doi:10.1371/journal.pcbi.0020123.

Chung, S.-W., Miles, F. L., Sikes, R. A., Cooper, C. R., Farach-Carson, M. C., Ogunnaike, B. A., 2009. Quantitative Modeling and Analysis of the Transforming Growth Factor-β Signaling Pathway. Biophysical journal 96, 1733-1750

Clarke, D. C., Liu, X., 2008. Decoding the quantitative nature of TGF-beta/Smad signaling. Trends Cell Biol 18, 430-42, doi:10.1016/j.tcb.2008.06.006.

Clarke, D. C., Betterton, M. D., Liu, X., 2006. Systems theory of Smad signalling. IEE Proceedings-Systems Biology 153, 412-424

Conery, A. R., Cao, Y., Thompson, E. A., Townsend, C. M., Jr., Ko, T. C., Luo, K., 2004. Akt interacts directly with Smad3 to regulate the sensitivity to TGF-beta induced apoptosis. Nat Cell Biol 6, 366-72.

D'Amour, K. A., Agulnick, A. D., Eliazer, S., Kelly, O. G., Kroon, E., Baetge, E. E., 2005. Efficient differentiation of human embryonic stem cells to definitive endoderm. Nat Biotechnol 23, 1534-41, doi:10.1038/nbt1163.

D'Amour, K. A., Bang, A. G., Eliazer, S., Kelly, O. G., Agulnick, A. D., Smart, N. G., Moorman, M. A., Kroon, E., Carpenter, M. K., Baetge, E. E., 2006a. Production of pancreatic hormone-expressing endocrine cells from human embryonic stem cells. Nat Biotechnol 24, 1392-401, doi:10.1038/nbt1259.

D'Amour, K. A., Bang, A. G., Eliazer, S., Kelly, O. G., Agulnick, A. D., Smart, N. G., Moorman, M. A., Kroon, E., Carpenter, M. K., Baetge, E. E., 2006b. Production of pancreatic hormone–expressing endocrine cells from human embryonic stem cells. Nature biotechnology 24, 1392-1401.

Dalton, S., 2013. Signaling networks in human pluripotent stem cells. Curr Opin Cell Biol 25, 241-6, doi:10.1016/j.ceb.2012.09.005.

Danielpour, D., Song, K., 2006. Cross-talk between IGF-I and TGF-β signaling pathways. Cytokine & growth factor reviews 17, 59-74

Daun, S., Rubin, J., Vodovotz, Y., Clermont, G., 2008. Equation-based models of dynamic biological systems. J Crit Care 23, 585-94, doi:10.1016/j.jcrc.2008.02.003.

Davies, R. G., Jans, D. A., Wagstaff, K. M., 2010. Use of fluorescence photobleaching techniques to measure the kinetics of intracellular transport. Microscopy: Science, technology, applications and education. Spain: Formatex Research Center, 756-763.

Davis, M. J., Skodje, R. T., Tomlin, A. S., 2011. Global Sensitivity Analysis of Chemical-Kinetic Reaction Mechanisms: Construction and Deconstruction of the Probability Density Function. Journal of Physical Chemistry A 115, 1556.

de Vargas Roditi, L., Claassen, M., 2015. Computational and experimental single cell biology techniques for the definition of cell type heterogeneity, interplay and intracellular dynamics. Curr Opin Biotechnol 34, 9-15, doi:10.1016/j.copbio.2014.10.010.

Discher, D. E., Mooney, D. J., Zandstra, P. W., 2009. Growth factors, matrices, and forces combine and control stem cells. Science 324, 1673-7, doi:10.1126/science.1171643.

Divina, F., Aguilar-Ruiz, J. S., 2006. Biclustering of expression data with evolutionary computation. Ieee Transactions on Knowledge and Data Engineering 18, 590-602, doi:Doi 10.1109/Tkde.2006.74.

Dojer, N., Gambin, A., Mizera, A., Wilczynski, B., Tiuryn, J., 2006. Applying dynamic Bayesian networks to perturbed gene expression data. BMC Bioinformatics 7, 249, doi:10.1186/1471-2105-7-249.

Efron, B., Tibshirani, R. J., 1994. An introduction to the bootstrap (chapman & hall/crc monographs on statistics & applied probability).

Emr, B., Sadowsky, D., Azhar, N., Gatto, L. A., An, G., Nieman, G. F., Vodovotz, Y., 2014. Removal of Inflammatory Ascites Is Associated with Dynamic Modification of Local and Systemic Inflammation Along with Prevention of Acute Lung Injury: In Vivo and in Silico Studies. Shock 41, 317-323, doi:Doi 10.1097/Shk.0000000000000121.

Feil, B., Kucherenko, S., Shah, N., 2009. Comparison of monte carlo and quasi monte carlo sampling methods in high dimensional model representation. Advances in System Simulation, 2009. SIMUL'09. First International Conference on. IEEE, pp. 12-17.

Feng, X., Hooshangi, S., Chen, D., Li, G., Weiss, R., Rabitz, H., 2004. Optimizing genetic circuits by global sensitivity analysis. Biophysical journal 87, 2195-2202.

Filippone, M., Masulli, F., Rovetta, S., Mitra, S., Banka, H., 2006. Possibilistic approach to biclustering: An application to oligonucleotide microarray data analysis. Springer, pp. 312-322

Fischer, Y., Ganic, E., Ameri, J., Xian, X., Johannesson, M., Semb, H., 2010. NANOG reporter cell lines generated by gene targeting in human embryonic stem cells. PLoS One 5, doi:10.1371/journal.pone.0012533.

Fong, C. Y., Peh, G. S. L., Gauthaman, K., Bongso, A., 2009. Separation of SSEA-4 and TRA-1-60 Labelled Undifferentiated Human Embryonic Stem Cells from A Heterogeneous Cell Population Using Magnetic-Activated Cell Sorting (MACS) and Fluorescence-Activated Cell Sorting (FACS). Stem Cell Reviews and Reports 5, 72-80, doi:10.1007/s12015-009-9054-4.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning. Springer series in statistics Springer, Berlin.

Gafni, O., Weinberger, L., Mansour, A. A., Manor, Y. S., Chomsky, E., Ben-Yosef, D., Kalma, Y., Viukov, S., Maza, I., Zviran, A., Rais, Y., Shipony, Z., Mukamel, Z., Krupalnik, V., Zerbib, M., Geula, S., Caspi, I., Schneir, D., Shwartz, T., Gilad, S., Amann-Zalcenstein, D., Benjamin, S., Amit, I., Tanay, A., Massarwa, R., Novershtern, N., Hanna, J. H., 2013. Derivation of novel human ground state naive pluripotent stem cells. Nature 504, 282-6, doi:10.1038/nature12745.

Gelman, A., Rubin, D. B., 1992. Inference from iterative simulation using multiple sequences. Statistical science, 457-472

Glauche, I., Herberg, M., Roeder, I., 2010. Nanog Variability and Pluripotency Regulation of Embryonic Stem Cells - Insights from a Mathematical Model Analysis. Plos One 5, e11238-e11238 doi:ARTN e11238 10.1371/journal.pone.0011238.

Golberg, D. E., 1989. Genetic algorithms in search, optimization, and machine learning. Addion wesley 1989.

Gonzalez-Perez, V., Schmierer, B., Hill, C. S., Sear, R. P., 2011. Studying Smad2 intranuclear diffusion dynamics by mathematical modelling of FRAP experiments. Integr Biol (Camb) 3, 197-207, doi:10.1039/c0ib00098a.

Grzegorczyk, M., Husmeier, D., 2011a. Non-homogeneous dynamic Bayesian networks for continuous data. Machine Learning 83, 355-419

Grzegorczyk, M., Husmeier, D., 2011b. Improvements in the reconstruction of time-varying gene regulatory networks: dynamic programming and regularization by information sharing among genes. Bioinformatics 27, 693-699

Guo, X., Wang, X.-F., 2008. Signaling cross-talk between TGF-β/BMP and other pathways. Cell research 19, 71-88.

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., Sethna, J. P., 2007. Universally sloppy parameter sensitivities in systems biology models. PLoS Comput Biol 3, 1871-78, doi:10.1371/journal.pcbi.0030189.

Gutiérrez-Avilés, D., Rubio-Escudero, C., Martínez-Álvarez, F., Riquelme, J. C., 2014. TriGen: A genetic algorithm to mine triclusters in temporal gene expression data. Neurocomputing 132, 42-53

Hagos, E. G., Dougan, S. T., 2007. Time-dependent patterning of the mesoderm and endoderm by Nodal signals in zebrafish. BMC Dev Biol 7, 22, doi:10.1186/1471-213X-7-22.

Hanczar, B., Nadif, M., 2011. Using the bagging approach for biclustering of gene expression data. Neurocomputing 74, 1595-1605, doi:10.1016/j.neucom.2011.01.013.

Hasenauer, J., Hasenauer, C., Hucho, T., Theis, F. J., 2014. ODE constrained mixture modelling: a method for unraveling subpopulation structures and dynamics. PLoS Comput Biol 10, e1003686, doi:10.1371/journal.pcbi.1003686.

Heinrich, R., Neel, B. G., Rapoport, T. A., 2002. Mathematical models of protein kinase signal transduction. Mol Cell 9, 957-70.

Herberg, M., Roeder, I., 2015. Computational modelling of embryonic stem-cell fate control. Development 142, 2250-60, doi:10.1242/dev.116343.

Hindmarsh, A. C., 1983. ODEPACK, A Systematized Collection of ODE Solvers, RS Stepleman et al.(eds.), North-Holland, Amsterdam,(vol. 1 of), pp. 55-64. IMACS transactions on scientific computation 1, 55-64.

Hua, F., Hautaniemi, S., Yokoo, R., Lauffenburger, D. A., 2006. Integrated mechanistic and data-driven modelling for multivariate analysis of signalling pathways. J R Soc Interface 3, 515-26, doi:10.1098/rsif.2005.0109.

Huang, S., 2011. Systems biology of stem cells: three useful perspectives to help overcome the paradigm of linear pathways. Philos Trans R Soc Lond B Biol Sci 366, 2247-59, doi:10.1098/rstb.2011.0008.

Iadevaia, S., Lu, Y., Morales, F. C., Mills, G. B., Ram, P. T., 2010. Identification of optimal drug combinations targeting cellular networks: integrating phospho-proteomics and computational network analysis. Cancer Res 70, 6704-14, doi:10.1158/0008-5472.CAN-10-0460.

Janes, K. A., 2015. An analysis of critical factors for quantitative immunoblotting. Sci Signal 8, rs2, doi:10.1126/scisignal.2005966.

Janes, K. A., Yaffe, M. B., 2006. Data-driven modelling of signal-transduction networks. Nat Rev Mol Cell Biol 7, 820-8, doi:10.1038/nrm2041.

Jaqaman, K., Danuser, G., 2006. Linking data to models: data regression. Nat Rev Mol Cell Biol 7, 813-9, doi:10.1038/nrm2030.

Jaramillo, M., Mathew, S., Task, K., Barner, S., Banerjee, I., 2014. Potential for Pancreatic Maturation of Differentiating Human Embryonic Stem Cells Is Sensitive to the Specific Pathway of Definitive Endoderm Commitment. PloS one 9, e94307

Jason, S. L., Ramasamy, T. S., Murphy, N., Holt, M. K., Czapiewski, R., Wei, S.-K., Cui, W., 2015. PI3K/mTORC2 regulates TGF-[beta]/Activin signalling by modulating Smad2/3 activity via linker phosphorylation. Nature communications 6.

Jones, J. M., Thomson, J. A., 1999. Human embryonic stem cell technology. Vol. 18, pp. 219-223

Kamato, D., Burch, M. L., Piva, T. J., Rezaei, H. B., Rostam, M. A., Xu, S., Zheng, W., Little, P. J., Osman, N., 2013. Transforming growth factor-beta signalling: role and consequences of Smad linker region phosphorylation. Cell Signal 25, 2017-24, doi:10.1016/j.cellsig.2013.06.001.

Kaminska, B., Wesolowska, A., Danilkiewicz, M., 2005. TGF beta signalling and its role in tumour pathogenesis. Acta Biochim Pol 52, 329-37.

Kang, J. S., Liu, C., Derynck, R., 2009. New regulatory mechanisms of TGF-beta receptor function. Trends Cell Biol 19, 385-94, doi:10.1016/j.tcb.2009.05.008.

Kang, M., Day, C. A., Kenworthy, A. K., DiBenedetto, E., 2012. Simplified equation to extract diffusion coefficients from confocal FRAP data. Traffic 13, 1589-600, doi:10.1111/tra.12008.

Katoh, M., Katoh, M., 2006. CER1 is a common target of WNT and NODAL signaling pathways in human embryonic stem cells. International journal of molecular medicine 17, 795-799

Kaufman, L., Rousseeuw, P. J., 2009. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.

Kent, E., Neumann, S., Kummer, U., Mendes, P., 2013. What can we learn from global sensitivity analysis of biochemical systems?

Kim, K. A., Spencer, S. L., Albeck, J. G., Burke, J. M., Sorger, P. K., Gaudet, S., Kim do, H., 2010. Systematic calibration of a cell signaling network model. BMC Bioinformatics 11, 202, doi:10.1186/1471-2105-11-202.

Kiparissides, A., Kucherenko, S., Mantalaris, A., Pistikopoulos, E., 2009. Global sensitivity analysis challenges in biological systems modeling. Industrial & Engineering Chemistry Research 48, 7168-7180.

Kiselyov, V. V., Versteyhe, S., Gauguin, L., De Meyts, P., 2009. Harmonic oscillator model of the insulin and IGF1 receptors' allosteric binding and activation. Mol Syst Biol 5, 243, doi:10.1038/msb.2008.78.

Koller, D., Friedman, N., 2009. Probabilistic graphical models: principles and techniques. MIT press.

Kroon, E., Martinson, L. A., Kadoya, K., Bang, A. G., Kelly, O. G., Eliazer, S., Young, H., Richardson, M., Smart, N. G., Cunningham, J., 2008. Pancreatic endoderm derived from human embryonic stem cells generates glucose-responsive insulin-secreting cells in vivo. Nature biotechnology 26, 443-452.

Li, G., Rabitz, H., 2012a. General formulation of HDMR component functions with independent and correlated variables. Journal of Mathematical Chemistry 50, 99-130.

Li, G., Rosenthal, C., Rabitz, H., 2001a. High dimensional model representations. The Journal of Physical Chemistry A 105, 7765-7777

Li, G., Rosenthal, C., Rabitz, H., 2001b. High dimensional model representations. The Journal of Physical Chemistry A 105, 7765-7777.

Li, G., Wang, S. W., Rabitz, H., 2002a. Practical approaches to construct RS-HDMR component functions. The Journal of Physical Chemistry A 106, 8721-8733.

Li, G., Wang, S. W., Rabitz, H., Wang, S., Jaffé, P., 2002b. Global uncertainty assessments by high dimensional model representations (HDMR). Chemical Engineering Science 57, 4445-4460.

Li, G., Rabitz, H., Yelvington, P. E., Oluwole, O. O., Bacon, F., Kolb, C. E., Schoendorf, J., 2010. Global sensitivity analysis for systems with independent and/or correlated inputs. The Journal of Physical Chemistry A 114, 6022-6032.

Li, G. Y., Rabitz, H., 2012b. General formulation of HDMR component functions with independent and correlated variables. Journal of Mathematical Chemistry 50, 99-130, doi:Doi 10.1007/S10910-011-9898-0.

Liu, J., Dai, W., Hahn, J., 2014. Mathematical Modeling and Analysis of Crosstalk between MAPK Pathway and Smad-Dependent TGF-β Signal Transduction. Processes 2, 570-595.

MacArthur, B. D., Please, C. P., Oreffo, R. O., 2008. Stochasticity and the molecular mechanisms of induced pluripotency. PLoS One 3, e3086, doi:10.1371/journal.pone.0003086.

Madeira, S. C., Oliveira, A. L., 2004. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinform 1, 24-45, doi:10.1109/TCBB.2004.2.

Mahanta, P., Ahmed, H. A., Bhattacharyya, D. K., Kalita, J. K., 2011. Triclustering in gene expression data analysis: a selected survey. IEEE, pp. 1-6

Mahdavi, A., Davey, R. E., Bhola, P., Yin, T., Zandstra, P. W., 2007. Sensitivity analysis of intracellular signaling pathway kinetics predicts targets for stem cell fate control. PLoS Comput Biol 3, e130, doi:10.1371/journal.pcbi.0030130.

Malkin, A. D., Sheehan, R. P., Mathew, S., Federspiel, W. J., Redl, H., Clermont, G., 2015. A Neutrophil Phenotype Model for Extracorporeal Treatment of Sepsis. PLOS Comput Biol 11, e1004314

Marinho, P. A., Chailangkarn, T., Muotri, A. R., 2015. Systematic optimization of human pluripotent stem cells media using Design of Experiments. Scientific Reports 5, doi:ARTN 09834 10.1038/srep09834.

Massagué, J., 2012. TGFbeta signalling in context. Nat Rev Mol Cell Biol 13, 616-30, doi:10.1038/nrm3434.

Mathew, S., Banerjee, I., 2014. Quantitative Analysis of Robustness of Dynamic Response and Signal Transfer in Insulin mediated PI3K/AKT Pathway. Comput Chem Eng 71, 715-727, doi:10.1016/j.compchemeng.2014.07.018.

Mathew, S., Sundararaj, S., Mamiya, H., Banerjee, I., 2014. Regulatory interactions maintaining self-renewal of human embryonic stem cells as revealed through a systems analysis of

PI3K/AKT pathway. Bioinformatics Accepted, In Press, doi: 10.1093/bioinformatics/btu209.

Mathew, S., Jaramillo, M., Zhang, X., Zhang, L. A., Soto-Gutierrez, A., Banerjee, I., 2012. Analysis of alternative signaling pathways of endoderm induction of human embryonic stem cells identifies context specific differences. BMC Syst Biol 6, 154, doi:10.1186/1752-0509-6-154.

McLean, A. B., D'Amour, K. A., Jones, K. L., Krishnamoorthy, M., Kulik, M. J., Reynolds, D. M., Sheppard, A. M., Liu, H., Xu, Y., Baetge, E. E., 2007a. Activin a efficiently specifies definitive endoderm from human embryonic stem cells only when phosphatidylinositol 3- kinase signaling is suppressed. Stem Cells 25, 29-38

McLean, A. B., D'Amour, K. A., Jones, K. L., Krishnamoorthy, M., Kulik, M. J., Reynolds, D. M., Sheppard, A. M., Liu, H., Xu, Y., Baetge, E. E., 2007b. Activin a efficiently specifies definitive endoderm from human embryonic stem cells only when phosphatidylinositol 3- kinase signaling is suppressed. Stem Cells 25, 29-38.

Meyer, R., D'Alessandro, L. A., Kar, S., Kramer, B., She, B., Kaschek, D., Hahn, B., Wrangborg, D., Karlsson, J., Kvarnström, M., 2011. Heterogeneous kinetics of AKT signaling in individual cells are accounted for by variable protein concentration. Frontiers in physiology 3, 451-451.

Mfopou, J. K., Chen, B., Sui, L., Sermon, K., Bouwens, L., 2010. Recent advances and prospects in the differentiation of pancreatic cells from human embryonic stem cells. Diabetes 59, 2094-101, doi:10.2337/db10-0439.

Miller, M. A., Feng, X. J., Li, G., Rabitz, H. A., 2012. Identifying biological network structure, predicting network behavior, and classifying network state with high dimensional model representation (hdmr). PloS one 7, e37664.

Mimeault, M., Hauke, R., Batra, S. K., 2007. Stem cells: a revolution in therapeutics—recent advances in stem cell biology and their therapeutic applications in regenerative medicine and cancer therapies. Clinical Pharmacology & Therapeutics 82, 252-264

Miskov-Zivanov, N., Turner, M. S., Kane, L. P., Morel, P. A., Faeder, J. R., 2013. Duration of T Cell Stimulation as a Critical Determinant of Cell Fate and Plasticity. Science signaling 6, ra97.

Mitra, S., Banka, H., Paik, J. H., 2007. Evolutionary fuzzy biclustering of gene expression data. Rough Sets and Knowledge Technology. Springer, pp. 284-291

Mochan, E., Swigon, D., Ermentrout, G. B., Lukens, S., Clermont, G., 2014. A mathematical model of intrahost pneumococcal pneumonia infection dynamics in murine strains. J Theor Biol 353, 44-54, doi:10.1016/j.jtbi.2014.02.021.

Molofsky, A. V., Pardal, R., Morrison, S. J., 2004. Diverse mechanisms regulate stem cell self-renewal. Curr Opin Cell Biol 16, 700-7, doi:10.1016/j.ceb.2004.09.004.

Mummery, C., Van de Stolpe, A., Roelen, B., Clevers, H., 2014. Stem cells: scientific facts and fiction. Academic Press.

Murphy, K. P., 2002. Dynamic bayesian networks: representation, inference and learning. University of California, Berkeley.

Murray, J. D., 2002. Mathematical Biology I: An Introduction, vol. 17 of Interdisciplinary Applied Mathematics. Springer, New York, NY, USA.

Murry, C. E., Keller, G., 2008. Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. Cell 132, 661-80, doi:10.1016/j.cell.2008.02.008.

Needham, C. J., Bradford, J. R., Bulpitt, A. J., Westhead, D. R., 2007. A primer on learning in Bayesian networks for computational biology. PLoS Comput Biol 3, e129, doi:10.1371/journal.pcbi.0030129.

Nicholas, C. R., Gaur, M., Wang, S., Pera, R. A., Leavitt, A. D., 2007. A method for single-cell sorting and expansion of genetically modified human embryonic stem cells. Stem Cells Dev 16, 109-17, doi:10.1089/scd.2006.0059.

Nichols, J., Smith, A., 2012. Pluripotency in the embryo and in culture. Cold Spring Harb Perspect Biol 4, a008128, doi:10.1101/cshperspect.a008128.

Nicklas, D., Saiz, L., 2013. Computational modelling of Smad-mediated negative feedback and crosstalk in the TGF-β superfamily network. Journal of The Royal Society Interface 10.

Nicolas, F. J., De Bosscher, K., Schmierer, B., Hill, C. S., 2004. Analysis of Smad nucleocytoplasmic shuttling in living cells. J Cell Sci 117, 4113-25, doi:10.1242/jcs.01289.

Nim, T. H., Luo, L., White, J. K., Clement, M. V., Tucker-Kellogg, L., 2015. Non-canonical Activation of Akt in Serum-Stimulated Fibroblasts, Revealed by Comparative Modeling of Pathway Dynamics. PLoS Comput Biol 11, e1004505, doi:10.1371/journal.pcbi.1004505.

Nosova, E., Tagliaferri, R., Masulli, F., Rovetta, S., 2011. Biclustering by resampling. Computational Intelligence Methods for Bioinformatics and Biostatistics. Springer, pp. 147-158

Nostro, M. C., Sarangi, F., Ogawa, S., Holtzinger, A., Corneo, B., Li, X., Micallef, S. J., Park, I.-H., Basford, C., Wheeler, M. B., 2011. Stage-specific signaling through TGFβ family members and WNT regulates patterning and pancreatic specification of human pluripotent stem cells. Development 138, 861-871

Novak, I. L., Slepchenko, B. M., Mogilner, A., 2008. Quantitative analysis of G-actin transport in motile cells. Biophys J 95, 1627-38, doi:10.1529/biophysj.108.130096.

Nyman, E., Cedersund, G., Strålfors, P., 2012. Insulin signaling - mathematical modeling comes of age. Trends Endocrinol Metab 23, 107-15, doi:10.1016/j.tem.2011.12.007.

Osafune, K., Caron, L., Borowiak, M., Martinez, R. J., Fitz-Gerald, C. S., Sato, Y., Cowan, C. A., Chien, K. R., Melton, D. A., 2008. Marked differences in differentiation propensity among human embryonic stem cell lines. Nature biotechnology 26, 313-316.

Parker, R. S., Clermont, G., 2010. Systems engineering medicine: engineering the inflammation response to infectious and traumatic challenges. J R Soc Interface 7, 989-1013, doi:10.1098/rsif.2009.0517.

Pauklin, S., Vallier, L., 2013. The cell-cycle state of stem cells determines cell fate propensity. Cell 155, 135-47, doi:10.1016/j.cell.2013.08.031.

Payne, C., King, J., Hay, D., 2011. The role of activin/nodal and Wnt signaling in endoderm formation. Vitam Horm 85, 207-16, doi:10.1016/B978-0-12-385961-7.00010-X.

Phair, R. D., Misteli, T., 2001. Kinetic modelling approaches to in vivo imaging. Nat Rev Mol Cell Biol 2, 898-907, doi:10.1038/35103000.

Phillips, B. W., Hentze, H., Rust, W. L., Chen, Q. P., Chipperfield, H., Tan, E. K., Abraham, S., Sadasivam, A., Soong, P. L., Wang, S. T., Lim, R., Sun, W., Colman, A., Dunn, N. R., 2007. Directed differentiation of human embryonic stem cells into the pancreatic endocrine lineage. Stem Cells and Development 16, 561-578, doi:10.1089/scd.2007.0029.

Politis, D. N., Romano, J. P., 1994. The stationary bootstrap. Journal of the American Statistical association 89, 1303-1313.

Pontes, B., Giraldez, R., Aguilar-Ruiz, J. S., 2015. Biclustering on expression data: A review. J Biomed Inform 57, 163-180 doi:10.1016/j.jbi.2015.06.028.

Poulain, M., Furthauer, M., Thisse, B., Thisse, C., Lepage, T., 2006. Zebrafish endoderm formation is regulated by combinatorial Nodal, FGF and BMP signalling. Development 133, 2189-2200, doi:10.1242/dev.02387.

Prudhomme, W., Daley, G. Q., Zandstra, P., Lauffenburger, D. A., 2004. Multivariate proteomic analysis of murine embryonic stem cell self-renewal versus differentiation signaling. Proc Natl Acad Sci U S A 101, 2900-5, doi:10.1073/pnas.0308768101.

Qiao, J., Kang, J., Ko, T. C., Evers, B. M., Chung, D. H., 2006. Inhibition of transforming growth factor-beta/Smad signaling by phosphatidylinositol 3-kinase pathway. Cancer Lett 242, 207-14, doi:10.1016/j.canlet.2005.11.007.

Remy, I., Montmarquette, A., Michnick, S. W., 2004. PKB/Akt modulates TGF-β signalling through a direct interaction with Smad3. Nature cell biology 6, 358-365

Rendón, E., Abundez, I., Arizmendi, A., Quiroz, E., 2011. Internal versus external cluster validation indexes. International Journal of computers and communications 5, 27-34.

Rice, R. G., Do, D. D., 2012. Applied mathematics and modeling for chemical engineers. John Wiley & Sons.

Richardson, T., Kumta, P. N., Banerjee, I., 2014. Alginate Encapsulation of Human Embryonic Stem Cells to Enhance Directed Differentiation to Pancreatic Islet-like cells. Tissue Engineering Part A 20 3198-211, doi:doi:10.1089/ten.tea.2013.0659.

Roberts, G. O., Gelman, A., Gilks, W. R., 1997. Weak convergence and optimal scaling of random walk Metropolis algorithms. The annals of applied probability 7, 110-120

Rosowski, K. A., Mertz, A. F., Norcross, S., Dufresne, E. R., Horsley, V., 2015. Edges of human embryonic stem cell colonies display distinct mechanical properties and differentiation potential. Sci Rep 5, 14218, doi:10.1038/srep14218.

Rust, W. L., Sadasivam, A., Dunn, N. R., 2006. Three-dimensional extracellular matrix stimulates gastrulation-like events in human embryoid bodies. Stem cells and development 15, 889-904.

Schatten, G., 2013. Cellular promiscuity: explaining cellular fidelity in vivo against unrestrained pluripotency in vitro. EMBO Rep 14, 4, doi:10.1038/embor.2012.198.

Schier, A. F., 2009. Nodal morphogens. Cold Spring Harb Perspect Biol 1, a003459, doi:10.1101/cshperspect.a003459.

Schmid, A. C., Byrne, R. D., Vilar, R., Woscholski, R., 2004. Bisperoxovanadium compounds are potent PTEN inhibitors. FEBS Lett 566, 35-8, doi:10.1016/j.febslet.2004.03.102.

Schmierer, B., Hill, C. S., 2005. Kinetic analysis of Smad nucleocytoplasmic shuttling reveals a mechanism for transforming growth factor beta-dependent nuclear accumulation of Smads. Mol Cell Biol 25, 9845-58, doi:10.1128/MCB.25.22.9845-9858.2005.

Schmierer, B., Hill, C. S., 2007. TGFbeta-SMAD signal transduction: molecular specificity and functional flexibility. Nat Rev Mol Cell Biol 8, 970-82, doi:10.1038/nrm2297.

Schmierer, B., Tournier, A. L., Bates, P. A., Hill, C. S., 2008. Mathematical modeling identifies Smad nucleocytoplasmic shuttling as a dynamic signal-interpreting system. Proceedings of the National Academy of Sciences of the United States of America 105, 6608-6613, doi:Doi 10.1073/Pnas.0710134105.

Schneider, A., Klingmuller, U., Schilling, M., 2012. Short-term information processing, long-term responses: Insights by mathematical modeling of signal transduction. Early activation dynamics of key signaling mediators can be predictive for cell fate decisions. Bioessays 34, 542-50, doi:10.1002/bies.201100172.

Sedaghat, A. R., Sherman, A., Quon, M. J., 2002. A mathematical model of metabolic insulin signaling pathways. Am J Physiol Endocrinol Metab 283, E1084-101, doi:10.1152/ajpendo.00571.2001.

Selekman, J. A., Das, A., Grundl, N. J., Palecek, S. P., 2013. Improving Efficiency of Human Pluripotent Stem Cell Differentiation Platforms Using an Integrated Experimental and Computational Approach. Biotechnology and Bioengineering 110, 3024-3037, doi:10.1002/bit.24968.

Semb, H., 2008. Definitive endoderm: a key step in coaxing human embryonic stem cells into transplantable beta-cells. Biochemical Society Transactions 36, 272-275, doi:10.1042/Bst0360272.

Shen, Y., Matsuno, Y., Fouse, S. D., Rao, N., Root, S., Xu, R., Pellegrini, M., Riggs, A. D., Fan, G., 2008. X-inactivation in female human embryonic stem cells is in a nonrandom pattern and prone to epigenetic alterations. Proceedings of the National Academy of Sciences 105, 4709-4714

Shiraki, N., Yoshida, T., Araki, K., Umezawa, A., Higuchi, Y., Goto, H., Kume, K., Kume, S., 2008. Guided Differentiation of Embryonic Stem Cells into Pdx1‐Expressing Regional‐Specific Definitive Endoderm. Stem cells 26, 874-885

Singh, A. M., Reynolds, D., Cliff, T., Ohtsuka, S., Mattheyses, A. L., Sun, Y., Menendez, L., Kulik, M., Dalton, S., 2012a. Signaling network crosstalk in human pluripotent cells: a Smad2/3-regulated switch that controls the balance between self-renewal and differentiation. Cell stem cell 10, 312-326

Singh, A. M., Reynolds, D., Cliff, T., Ohtsuka, S., Mattheyses, A. L., Sun, Y. H., Menendez, L., Kulik, M., Dalton, S., 2012b. Signaling Network Crosstalk in Human Pluripotent Cells: A Smad2/3-Regulated Switch that Controls the Balance between Self-Renewal and Differentiation. Cell Stem Cell 10, 312-326, doi:Doi 10.1016/J.Stem.2012.01.014.

Singh, A. M., Chappell, J., Trost, R., Lin, L., Wang, T., Tang, J., Matlock, B. K., Weller, K. P., Wu, H., Zhao, S., Jin, P., Dalton, S., 2013. Cell-cycle control of developmentally regulated transcription factors accounts for heterogeneity in human pluripotent cells. Stem Cell Reports 1, 532-44, doi:10.1016/j.stemcr.2013.10.009.

Slezak, D. F., Suárez, C., Cecchi, G. A., Marshall, G., Stolovitzky, G., 2010. When the optimal is not the best: parameter estimation in complex biological models. PloS one 5, e13283.

Slonim, D. K., 2002. From patterns to pathways: gene expression data analysis comes of age. Nat Genet 32 Suppl, 502-8, doi:10.1038/ng1033.

Smith, Q., Stukalin, E., Kusuma, S., Gerecht, S., Sun, S. X., 2015. Stochasticity and Spatial Interaction Govern Stem Cell Differentiation Dynamics. Sci Rep 5, 12617, doi:10.1038/srep12617.

Song, K., Cornelius, S. C., Reiss, M., Danielpour, D., 2003. Insulin-like growth factor-I inhibits transcriptional responses of transforming growth factor-β by phosphatidylinositol 3-kinase/Akt-dependent suppression of the activation of Smad3 but not Smad2. Journal of Biological Chemistry 278, 38342-38351

Song, K., Wang, H., Krebs, T. L., Danielpour, D., 2006. Novel roles of Akt and mTOR in suppressing TGF-beta/ALK5-mediated Smad3 activation. EMBO J 25, 58-69, doi:10.1038/sj.emboj.7600917.

Song, S. O., Hogg, J., Peng, Z.-Y., Parker, R., Kellum, J. A., Clermont, G., 2012. Ensemble models of neutrophil trafficking in severe sepsis. PLoS Comput Biol 8, 1-16.

Soria, B., Gauthier, B. R., Martin, F., Tejedo, J. R., Bedoya, F. J., Rojas, A., Hmadcha, A., 2015. Using stem cells to produce insulin. Expert Opinion on Biological Therapy 15, 1469-1489, doi:10.1517/14712598.2015.1066330.

Sulzbacher, S., Schroeder, I. S., Truong, T. T., Wobus, A. M., 2009. Activin A-induced differentiation of embryonic stem cells into endoderm and pancreatic progenitors—the influence of differentiation factors and culture conditions. Stem Cell Reviews and Reports 5, 159-173

Sumi, T., Tsuneyoshi, N., Nakatsuji, N., Suemori, H., 2008. Defining early lineage specification of human embryonic stem cells by the orchestrated balance of canonical Wnt/β-catenin, Activin/Nodal and BMP signaling. Development 135, 2969-2979

Sun, T., Ye, F., Ding, H., Chen, K., Jiang, H., Shen, X., 2006. Protein tyrosine phosphatase 1B regulates TGF beta 1-induced Smad2 activation through PI3 kinase-dependent pathway. Cytokine 35, 88-94, doi:10.1016/j.cyto.2006.07.013.

Swigon, D., 2012. Ensemble modeling of biological systems. Math. Life Sci., The Publishing House, Berlin, 316.

Tanay, A., Sharan, R., Shamir, R., 2005. Biclustering algorithms: A survey. Handbook of computational molecular biology 9, 122-124.

Taniguchi, C. M., Emanuelli, B., Kahn, C. R., 2006. Critical nodes in signalling pathways: insights into insulin action. Nature Reviews Molecular Cell Biology 7, 85-96

Task, K., Jaramillo, M., Banerjee, I., 2012. Population based model of human embryonic stem cell (hESC) differentiation during endoderm induction. PLoS One 7, e32975, doi:10.1371/journal.pone.0032975.

Tchagang, A. B., Phan, S., Famili, F., Shearer, H., Fobert, P., Huang, Y., Zou, J., Huang, D., Cutler, A., Liu, Z., 2012. Mining biological information from 3D short time-series gene expression data: the OPTricluster algorithm. BMC bioinformatics 13, 54

Teukolski, S. A., Flannery, B. P., Press, W. H., Vetterling, W. T., 1989. Numerical Recipes in FORTRAN-The Art of Scientific Computing. University Press.

Tibes, R., Qiu, Y., Lu, Y., Hennessy, B., Andreeff, M., Mills, G. B., Kornblau, S. M., 2006. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. Mol Cancer Ther 5, 2512-21, doi:10.1158/1535-7163.MCT-06-0334.

Till, J. E., McCulloch, E., 1961. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. Radiat Res 14, 213-22.

Till, J. E., McCulloch, E. A., Siminovitch, L., 1964. A Stochastic Model of Stem Cell Proliferation, Based on the Growth of Spleen Colony-Forming Cells. Proc Natl Acad Sci U S A 51, 29-36.

Torres-Padilla, M. E., Chambers, I., 2014. Transcription factor heterogeneity in pluripotent stem cells: a stochastic advantage. Development 141, 2173-81, doi:10.1242/dev.102624.

Veber, D. F., Johnson, S. R., Cheng, H. Y., Smith, B. R., Ward, K. W., Kopple, K. D., 2002. Molecular properties that influence the oral bioavailability of drug candidates. Journal of Medicinal Chemistry 45, 2615-2623, doi:10.1021/jm020017n.

Vilar, J. M., Jansen, R., Sander, C., 2006. Signal processing in the TGF-beta superfamily ligand-receptor network. PLoS Comput Biol 2, e3, doi:10.1371/journal.pcbi.0020003.

Vizán, P., Miller, D. S., Gori, I., Das, D., Schmierer, B., Hill, C. S., 2013. Controlling long-term signaling: receptor dynamics determine attenuation and refractory behavior of the TGF-beta pathway. Sci Signal 6, ra106, doi:10.1126/scisignal.2004416.

Vodovotz, Y., An, G., 2014. Translational Systems Biology: Concepts and Practice for the Future of Biomedical Research. Elsevier.

Voskas, D., Ling, L. S., Woodgett, J. R., 2010. Does GSK-3 provide a shortcut for PI3K activation of Wnt signalling? F1000 biology reports 2, 82.

Waddington, C. H., 1957. The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser. London: George Allen & Unwin, Ltd.

Wegner, K., Bachmann, A., Schad, J. U., Lucarelli, P., Sahle, S., Nickel, P., Meyer, C., Klingmuller, U., Dooley, S., Kummer, U., 2012. Dynamics and feedback loops in the transforming growth factor beta signaling pathway. Biophys Chem 162, 22-34, doi:10.1016/j.bpc.2011.12.003.

White, D. E., Kinney, M. A., McDevitt, T. C., Kemp, M. L., 2013. Spatial pattern dynamics of 3D stem cell loss of pluripotency via rules-based computational modeling. PLoS Comput Biol 9, e1002952, doi:10.1371/journal.pcbi.1002952.

Wilding, L., Gannon, M., 2004. The role of pdx1 and HNF6 in proliferation and differentiation of endocrine precursors. Diabetes Metab Res Rev 20, 114-23, doi:10.1002/dmrr.429.

Woolf, P. J., Prudhomme, W., Daheron, L., Daley, G. Q., Lauffenburger, D. A., 2005. Bayesian analysis of signaling networks governing embryonic stem cell fate decisions. Bioinformatics 21, 741-53, doi:10.1093/bioinformatics/bti056.

Wu, J. C., Tzanakakis, E. S., 2012. Contribution of Stochastic Partitioning at Human Embryonic Stem Cell Division to NANOG Heterogeneity. Plos One 7, doi:ARTN e50715 10.1371/journal.pone.0050715.

Xu, C., Inokuma, M. S., Denham, J., Golds, K., Kundu, P., Gold, J. D., Carpenter, M. K., 2001. Feeder-free growth of undifferentiated human embryonic stem cells. Nat Biotechnol 19, 971-4, doi:10.1038/nbt1001-971.

Xu, R.-H., Chen, X., Li, D. S., Li, R., Addicks, G. C., Glennon, C., Zwaka, T. P., Thomson, J. A., 2002. BMP4 initiates human embryonic stem cell differentiation to trophoblast. Nature biotechnology 20, 1261-1264

Xu, X., Browning, V. L., Odorico, J. S., 2011. Activin, BMP and FGF pathways cooperate to promote endoderm and pancreatic lineage cell differentiation from human embryonic stem cells. Mechanisms of development 128, 412-427

Xu, Y., Shi, Y., Ding, S., 2008. A chemical approach to stem-cell biology and regenerative medicine. Nature 453, 338-44, doi:10.1038/nature07042.

Yang, J., Wang, H., Wang, W., Yu, P., 2003. Enhanced biclustering on expression data. IEEE, pp. 321-327

Yeo, J. C., Ng, H. H., 2013. The transcriptional regulation of pluripotency. Cell Res 23, 20-32, doi:10.1038/cr.2012.172.

Yoon, B. S., Jun, E. K., Park, G., Jun Yoo, S., Moon, J. H., Soon Baik, C., Kim, A., Kim, H., Kim, J. H., Young Koh, G., 2010. Optimal Suppression of Protein Phosphatase 2A Activity Is Critical for Maintenance of Human Embryonic Stem Cell Self- Renewal. Stem Cells 28, 874-884

Yoon, J., Deisboeck, T. S., 2009. Investigating differential dynamics of the MAPK signaling cascade using a multi-parametric global sensitivity analysis. PloS one 4, e4560

Yu, P., Pan, G., Yu, J., Thomson, J. A., 2011. FGF2 sustains NANOG and switches the outcome of BMP4-induced human embryonic stem cell differentiation. Cell stem cell 8, 326-334

Zhang, D., Jiang, W., Shi, Y., Deng, H., 2009a. Generation of pancreatic islet cells from human embryonic stem cells. Sci China C Life Sci 52, 615-21, doi:10.1007/s11427-009-0095-3.

Zhang, D., Jiang, W., Liu, M., Sui, X., Yin, X., Chen, S., Shi, Y., Deng, H., 2009b. Highly efficient differentiation of human ES cells and iPS cells into mature pancreatic insulin-producing cells. Cell research 19, 429-438.

Zhang, D. H., Jiang, W., Liu, M., Sui, X., Yin, X. L., Chen, S., Shi, Y., Deng, H. K., 2009c. Highly efficient differentiation of human ES cells and iPS cells into mature pancreatic insulin-producing cells. Cell Research 19, 429-438, doi:Doi 10.1038/Cr.2009.28.

Zhang, L., Zhou, F., ten Dijke, P., 2013. Signaling interplay between transforming growth factor-β receptor and PI3K/AKT pathways in cancer. Trends in biochemical sciences 38, 612-620

Zhang, X., Jaramillo, M., Singh, S., Kumta, P., Banerjee, I., 2012a. Analysis of Regulatory Network Involved in Mechanical Induction of Embryonic Stem Cell Differentiation. PLoS One 7, e35700.

Zhang, Y., Li, W., Laurent, T., Ding, S., 2012b. Small molecules, big roles -- the chemical manipulation of stem cell fate and somatic cell reprogramming. J Cell Sci 125, 5609-20, doi:10.1242/jcs.096032.

Zhang, Y. E., 2008. Non-Smad pathways in TGF-β signaling. Cell research 19, 128-139

Zheng, W.-H., Kar, S., Quirion, R., 2000. Stimulation of protein kinase C modulates insulin-like growth factor-1-induced akt activation in PC12 cells. Journal of Biological Chemistry 275, 13377-13385

Zi, Z., Klipp, E., 2007. Constraint-based modeling and kinetic analysis of the Smad dependent TGF-beta signaling pathway. PLoS One 2, e936, doi:10.1371/journal.pone.0000936.

Zi, Z., Feng, Z., Chapnick, D. A., Dahl, M., Deng, D., Klipp, E., Moustakas, A., Liu, X., 2011. Quantitative analysis of transient and sustained transforming growth factor-β signaling dynamics. Molecular systems biology 7.

Zielinski, R., Przytycki, P. F., Zheng, J., Zhang, D., Przytycka, T. M., Capala, J., 2009. The crosstalk between EGF, IGF, and Insulin cell signaling pathways--computational and experimental analysis. BMC Syst Biol 3, 88, doi:10.1186/1752-0509-3-88.

Ziraldo, C., Solovyev, A., Allegretti, A., Krishnan, S., Henzel, M. K., Sowa, G. A., Brienza, D., An, G., Mi, Q., Vodovotz, Y., 2015. A Computational, Tissue-Realistic Model of Pressure Ulcer Formation in Individuals with Spinal Cord Injury. PLoS Comput Biol 11, e1004309, doi:10.1371/journal.pcbi.1004309.

Zorn, A. M., Wells, J. M., 2009. Vertebrate endoderm development and organ formation. Annu Rev Cell Dev Biol 25, 221-51, doi:10.1146/annurev.cellbio.042308.113344.