Understanding embryonic stem cell self-renewal and differentiation through systems biology

Raheem Peerani

***Abstract*—Understanding the molecular mechanisms behind embryonic stem (ES) cell fate is critical for controlled differentiation of ES cells for cell based therapies and tissue engineering. Three tools of systems biology, serial analysis of gene expression (SAGE), fluorescent high throughput screening (HTS), and Bayesian networks (BNs) are described with respect to how they can be used to understand embryonic stem cell fate. It is expected that these tools will elucidate the intrinsic and extrinsic control mechanisms that govern self-renewal and differentiation in molecular and quantitative terms. Such measurements will clarify some of the currently debated topics like stem cell plasticity and irreversible differentiation.**

***Index Terms*— Bayesian networks, differentiation, embryonic stem cell, high throughput screening (HTS), serial analysis of gene expression (SAGE), systems biology**

I. Introduction

Recent interest in embryonic stem (ES) cells has primarily flourished because of their remarkable ability to serve as models for development as well as potential sources for cells in pharmaceutical testing, cell based therapies, and tissue engineering. ES cells have been shown to be self-renewing and *pluripotent*, capable of generating all three germ layers (endoderm, mesoderm, ectoderm) and thus potentially capable of generating any somatic cell of the human body. In order to use ES cells in the above applications, precise control of their differentiation must be achieved. Any attempt to control ES cell differentiation must recognize the multitude and complexity of the signals involved in governing stem cell fate and thus systems biology, as an approach to study and model interacting networks, becomes considerably useful. Ultimately, the purpose of these network models would be to predict stem cell fate across a variety of culture conditions, cell lines, and model organisms. Such predictions would enable scientists to consolidate new and existing stem cell biology into specific paradigms as well as empower bioengineers with the ability to construct complex and functional tissues out of ES cells.

Systems biology is a discipline that attempts to predict the behaviour of a biological system by quantifying all the molecular elements present, determining their interactions and integrating these interactions into network models. Some of the key tools of systems biology that will assist in understanding ES cell regulatory mechanisms are transcriptome determination through serial analysis of gene expression (SAGE), high throughput screening (HTS) of DNA, protein, and cell microarrays, and mathematical modeling using Bayesian statistics. In this paper, these tools will be discussed in the context of how they can be used to determine the extrinsic and intrinsic mechanisms governing ES cell self-renewal and differentiation. The framework for using these tools in this context is presented in Fig. 1. SAGE can be used to determine the structural components of the network by identifying the active genes present in undifferentiated ES cells. Having identified the genes and proteins involved, HTS can produce the necessary datasets to determine the signal transduction pathways and rate constants involved in ES cell self-renewal and differentiation. These datasets will be inputted into a dynamic Bayesian network model to make hypotheses that predict stem cell fate under a variety of extrinsic conditions. In turn, HTS will provide a means to test several extrinsic conditions simultaneously either validating these hypotheses or suggesting modifications to them. Likewise, incorrect hypotheses may indicate missing structural components to the network which would suggest further SAGE analysis of ES cells.

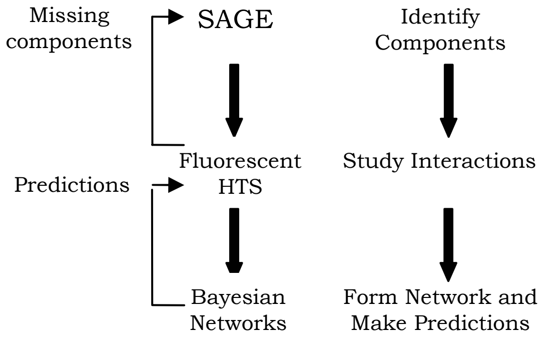

Fig.1. Framework for understanding ES self-renewal and differentiation through SAGE, fluorescent HTS, and Bayesian networks (BNs).

II. Mechanisms Directing Stem Cell Fate

In vivo, it is speculated that stem cells live in ‘niches’, spatially defined regions that contain a combination of soluble cytokines and insoluble extra-cellular matrix (ECM) proteins that govern self-renewal and differentiation. These factors are extrinsic mechanisms by which the organism can direct stem cell function in order to repair and maintain tissue. In contrast, intrinsic mechanisms pertain to the stem cell itself. This machinery includes the appropriate receptors, intracellular signaling species, and effector proteins that allow the cell to be competent to respond to the extra-cellular environment [1]. The complexity is further enhanced by the fact that both intrinsic and extrinsic factors can have dimensions in quantity, time, and space (Table 1). Systematic study of these factors through systems biology is key to deciphering the complexity of stem cell fate decisions.

TABLE I
MECHANISMS GOVERNING STEM CELL FATE DECISIONS

	Examples	
	Intrinsic	Extrinsic
Quantitative	Intracellular concentration of transcription factor	Concentration of exogenous cytokine applied
Temporal	Changes in receptor expression	Signalling duration
Spatial	Autocrine signaling	Immobilized cytokines on surfaces

III. Serial Analysis of Gene expression (SAGE)

Serial analysis of gene expression (SAGE) provides gene expression profiles based upon quantifying the frequency of appearance of a particular tag that identifies a unique mRNA transcript. A summary of the SAGE technique is presented here [2]. First, the mRNA from a population of cells is extracted by incubating the mRNA with streptavidin beads containing multiple thymines (‘T’s) that bind to the poly(A) tails of mRNA transcripts. The mRNA is then translated into cDNA and short tags (10-14bp) are generated by specific restriction enzymes that cut the cDNA at the ends. These tags are sequenced and then concatenated first into dimers and subsequently into *concatemers* that contain all the tags joined together. These concatemers are then amplified in bacteria. After amplification, the concatemers are bound to the original cDNA transcripts and the frequency of binding is quantified through computer software. The sequence of each tag can be matched with gene sequences found in databases thus giving an indication of which genes are active in the population and their respective levels of expression.

SAGE is useful as a high throughput technique to measure gene expression profiles since quantified levels of expression can be determined even if unknown genes are present. It can be used as a ‘gene finder’ in cancer cells or diseased tissue since any tag not found in a database could correspond to a mutation or new gene. However, its capacity to aid in gene discovery is limited since it only provides a very short sequence of the new or mutated gene. In addition, SAGE has other drawbacks including: possible lack of tag specificity, the cDNA transcript may not possess a restriction enzyme recognition sequence, and tags may not be of the same length making it more difficult to quantify binding [2].

Nonetheless, SAGE has the particular ability to aid the understanding of ES cell self-renewal and differentiation in the following ways. First, it provides a comprehensive means to compare the intrinsic regulatory mechanisms in different cell lines. There are dozens of human (hES) and mouse embryonic stem cell lines available around the world that seem to differentiate and proliferate at different rates under the same culture conditions. SAGE may provide the molecular differences between these lines. Second, since ES cells are derived from the inner cell mass of the embryo, a very short and unique phase in the development of the organism, active genes specific to these cells are difficult to find. SAGE is a means to identify these new genes as indicated by one study that found that only 9% of SAGE tags generated from the R1 ES cell line matched those found in databases and that >35% of the unique tags did not match any known sequence [3]. Third, SAGE provides a means to compare ES cells from different species which allows researchers to leverage the nearly two decades of mouse ES cell work against hES cell research that is in its infancy. For instance, one study that compared SAGE profiles between two hES cell lines (HES3 and HES4) and one mouse ES cell line (R1) indicated several similarities and differences in gene expression between all three lines [4]. In particular, quantified mRNA levels corresponding to the inactive leukemia inhibitory factor (LIF) pathway in hES cell lines were found which is significant since this pathway is critical in *in vitro* maintenance of ES cells. The addition of LIF alone is capable of preventing mouse ES cell differentiation thus allowing virtually infinite scale-up of mouse ES cells. In the case of hES cells, this pathway is known to be inactive and thus maintenance of hES cells requires mouse embryonic feeder (MEF) layers that secrete unknown factors that promote hES self-renewal. SAGE studies may indicate why this pathway is inactive in hES cells.

Fourth and perhaps most intriguing is that SAGE may allow stem cell biologists to relate the phenomenon of ‘plasticity’ directly to gene expression profiles. Plasticity refers to the apparent ability of somatic (adult) stem cells to switch their lineage specification. For example, evidence was presented that hematopoietic stem cells could cross the blood-brain barrier to become neurons in the brain [5]. An analogue to ES cells is that early ES differentiation into the primordial germ layers, endoderm, mesoderm, and ectoderm, merely specifies rather than determines their fate and that significant switching between lineages can occur. Currently, such evidence is hotly debated and SAGE can clarify this issue by relating gene expression profiles of undifferentiated ES cells to those of the earliest committed progenitors. The two important questions that SAGE can answer are; one, is there a point of irreversibility when it comes to ES cell differentiation? and two, is there a genetic basis to any irreversibility?

As presented above, SAGE provides an amenable means to measure the intrinsic control mechanisms or genetic basis of ES cell self-renewal and differentiation. It provides the basis of HTS of proteins and cells that will provide the kinetics of ES cell self-renewal and differentiation as well as the effect of extrinsic factors on ES cell fate, that is the epigenetic control of ES cells.

IV. High throughput screening

High throughput screening (HTS) is an integral part of systems biology and is at the core of discovery based science. Current HTS detection methods include: fluorescence, mass spectrometry in both isotope coded affinity tag (ICAT) or tandem forms, Raman spectroscopy, in-situ hybridization, and oligopeptide, DNA, and antibody microarrays. These techniques provide the means to collect large enough datasets to develop a statistical model like a Bayesian network to represent a biological system. Only fluorescence based HTS techniques are reviewed here.

Fluorescence based HTS is dependent upon either genetic modification by transfecting a fluorescent reporter gene such as green fluorescent protein (GFP) under the control of a specific promoter in a cell or conjugating a protein with a fluorochrome such as fluorescein isothiocyanate (FITC), either by direct attachment or indirectly through fluorescent antibodies. After conjugation, high throughput instruments like a flow cytometer for single cell based fluorescence detection or an inverted fluorescent microscope for image acquisition can be used. In both cases, advanced computer software packages are available that can provide population and sub-population statistics based on measured fluorescence.

Several measurements can be made through fluorescence based HTS including: protein expressional levels at the single cell and population levels, receptor-ligand binding affinity, trafficking parameters like internalization, translocation, degradation, and recycling rates, cell morphology and migration, protein synthesis and protein-protein interactions. Fluorescence based HTS is also compatible with other fluorescence based techniques like fluorescence recovery after photobleaching (FRAP) and fluorescence resonance energy transfer (FRET). Depending on the apparatus, these measurements can be taken in real-time with live cells. Currently, the number of individual biological phenomena measured simultaneously is limited by the quality of the fluorochromes in terms of overlapping excitation and emission spectra as well as the detection levels by the apparatus. With the emergence of quantum dot based fluorochromes that have tightly defined excitation and emission spectra and increasingly sensitive fluorescence detectors, the sophistication and capabilities of fluorescence based HTS can only increase.

The role of fluorescent HTS in ES cell biology is manifold. First, it can be used as a quantitative means to measure intrinsic factors such as protein levels which correlate to the genes identified through SAGE. In particular for undifferentiated ES cells, these proteins include Oct4, Stat3, Sox2, Nanog, LIF, E-cadherin, state specific embryonic antigen 1 (SSEA-1), and many others. The protein measurements conducted though HTS should include all the transcription, translation, binding, and internalization rates and protein-protein interactions described earlier. Essentially, the goal of fluorescence based HTS is to characterize the phenotype of an ES cell as it chooses to self-renew or differentiate by measuring real-time protein levels and protein interactions quantitatively.

One example of the ability of fluorescence based HTS to screen phenotype is measuring phosphorylated STAT3 (STAT3P) levels in the presence of exogenous LIF [6]. STAT3 is a kinase which is phosphorylated downstream of LIF binding to its co-receptors, LIFR and gp130. In this particular experiment, ES cells were deprived of LIF for 24 to 96 hours and thus allowed to differentiate. Exogenous LIF was then reapplied and it was shown that single cells became less responsive to LIF, indicated by lower measured STAT3P levels, even though these cells remained OCT4+, a marker of undifferentiated ES cells [6]. This example shows how HTS can complement SAGE analysis by illustrating the effects of protein-protein interactions on ES self-renewal and differentiation. In addition, it shows that stem cell plasticity must be studied both through SAGE and HTS in order to determine if there is a genetic and phenotypic basis to irreversible differentiation.

Second, fluorescence based HTS can be used as a quantitative means to screen extrinsic factors on ES self-renewal and differentiation. One recent study attempted to optimize the differentiation of ES cells into primitive endoderm by varying the presence and concentration of several growth and differentiation factors (GDFs) [7]. Subsequent two-level factorial analysis on the data indicated the positive or negative effects of each GDF on primitive endoderm differentiation. Clearly as HTS technology progresses, more GDFs will be able to be screened simultaneously and thus HTS will aid bioengineers in their attempt to differentiate ES cells into a single cell type.

Third, sophisticated mathematical models developed for fluorescent based HTS devices, such as flow cytometry, have allowed the deconvolution of fluorescent intensity measurements into subpopulation rate parameters for ES cell differentiation, proliferation of differentiated cells, and proliferation of undifferentiated cells, i.e. self-renewal [8]. Thus using mathematical models one can distinguish the effects of these three sub-populations on a measured observation, making flow cytometry a more powerful tool in understanding the population dynamics involved in ES cell self-renewal and differentiation. This particular example illustrates the importance of mathematical modeling in understanding ES cell self-renewal and differentiation; a topic discussed further in the following section.

V. Bayseian networks

While there are several means to model biological systems including differentiation equations, neutral networks, and boolean networks, Bayesian networks (BNs) offer a probabilistic model that can depict causal relationships between variables. The networks are graphical in nature with each node representing a variable. Connections between nodes indicate conditional probability. These networks are capable of both linear and non-linear groups of variables. BNs also reduce the number of parameters to describe the system as compared to other modeling approaches which is a direct result of Markov's rule which states that in order to predict the value of a variable, X, only the values that directly influence X must be considered. Indirect and non-descendent variables from X can be ignored. In addition, since BNs are probabilistic in nature, they can deal with noise in an automated and systematic way. Likewise, they can model certain cell processes considered to be stochastic in nature [9].

Bayesian models work by fitting models to data. When the structure is known, the fitting operation optimizes parameters that will maximize the probability of attaining the training data. In the case where structure is only partially known, Bayesian networks can be used to determine system structure [10]. Structure is determined by generating competing models which can be used to fit the experimental data. Each fit is assigned a Bayesian score depending on the likelihood of the fit. Models that have a high score are then compared and common features of these models are identified. BNs have also been adapted to fit data taken from several sources as well as data that are dissimilar like boolean values and continuous time variables. Such versatility makes it ideal for interpreting data taken from HTS tools.

Bayesian networks can also have a time component in which case they are considered to by Dynamic Bayesian Networks (DBNs). DBNs have typically have fixed structures but the connections between nodes is fitted as a function of time.

Despite their apparent flexibility and versatility, Bayesian statistic models have their disadvantages. First, large observational datasets are necessary to generate the models, requiring significant computation and parallelization. Second, the model is probabilistic rather than mechanistic and therefore single outcomes are never predicted. Third, the sense of confidence in the model is solely based on probability.

BNs and DBNs can have a significant impact on understanding ES self-renewal and differentiation. Their goal would be to model the effects of external cues on intracellular protein levels that govern ES cell fate. One very recent study has validated the use of this approach [11]-[12]. In this study, ES cells were cultured in the presence of 16 different stimulation conditions and quantitative Western blotting was used to generate the phosphorylation states of 31 intracellular signaling components at three time points. The data was fit into a DBN and the network model was used to validate existing knowledge of the role of LIF and STAT3 on undifferentiated ES cells as well as predict new roles for ERK phosphorylation in ES cell differentiation and RAF phosphorylation in differentiated cell proliferation [12]. This study presents the first use of Bayesian networks in understanding ES cell fate and its success will likely make the use of BNs more prevalent in the future.

VI. Summary

Systems biology as an approach to study ES cell fate is becoming more popular and useful due to its ability to handle the large complexity and number of signals involved in the process. In this paper, three tools of systems biology were discussed in the context of how they can characterize the intrinsic and extrinsic mechanisms governing ES cell fate. First, SAGE was introduced as a means to determine the structure of the network as well as the genetic or intrinsic control of mechanisms involved in self-renewal and differentiation. Second, it was suggested that fluorescent HTS screening could be used to determine the protein level contributions to ES cell fate including synthesis, binding, and trafficking parameters as well as to decipher protein-protein and protein-DNA interactions. Simultaneously, it could be to screen several extrinsic factors on ES cell fate. Third, Bayesian networks and dynamic Bayesian networks were explained in order to show how HTS data could be used to generate network models that describe ES self-renewal and differentiation. Examples were given on how such networks can validate existing knowledge about signal transduction pathways as well as propose new roles for other proteins previously unconsidered. Ultimately as HTS and mathematical modeling tools get more sophisticated, understanding ES cell fate will become more fruitful allowing stem cell biologists to clarify many of the hotly debated topics of today such as plasticity as well as aiding bioengineers to develop ES cell based therapies and tissues.

References

[1] R.E. Davey and P.W. Zandstra, “Signal processing underlying extrinsic control of stem cell fate,” *Curr Opin Hematol*, vol. 11, pp. 95-101, Mar. 2004.

[2] C.H. Song and M. Wyse. Painless Gene Expression Profiling: SAGE (Serial Analysis of Gene Expression) [Online]. Available: http://www.bioteach.ubc.ca/MolecularBiology/PainlessGeneExpressionProfiling

[3] S.V. Anisimov, K.V. Tarasov, D. Tweedie, M.D. Stern, A.M. Wobus, K.R. Boheler, “SAGE Identification of Gene Transcripts with Profiles Unique to Pluripotent Mouse R1 Embryonic Stem Cells,” *Genomics*, vol. 79(2), pp. 169-176, Feb. 2002.

[4] M. Richards, S..P. Tan, J.H. Tan, W.K. Chan, A. Bongso, “The Transcriptome Profile of Human Embryonic Stem Cells by SAGE,” *Stem Cells*, vol. 22, pp. 51-64, 2004.

[5] C.C. Shih, D. DiGiusto, A. Mamelak, T. Lebon, S.J. Forman, “Hematopoietic potential of neural stem cells: plasticity versus heterogeneity,” *Leuk Lymphoma*, vol. 43(12), pp. 2263-2268, Dec. 2002.

[6] R.E. Davey, Stem Cell Laboratories, Toronto, ON, unpublished data, Sept 2004.

[7] K.H. Chang and P.W. Zandstra, “Quantitative screening of embryonic stem cell differentiation: Endoderm formation as a model.,” *Biotechnol Bioeng*, vol. 88(3), pp. 287-298, Nov. 2004.

[8] W.A. Prudhomme, K.H. Duggar, D.A. Lauffenburger, “Cell Population Dynamics Model for Decovolution of Murine Embryonic Stem Cell Self-Renewal and Differentiation Responses to Cytokines and Extracellular Matrix”, *Biotechnol Bioeng*, vol. 88(3), pp. 264-272, Nov. 2004.

[9] K. Sachs, D. Gifford, T. Jaakkola, P. Sorger, and D. A. Lauffenburger, (Sept. 2002) “Bayesian Network Approach to Cell Signaling Pathway Modeling,” Science's STKE [Online] Available: http://www.stke.org/cgi/content/full/sigtrans;2002/148/pe38

[10] K. Murphy, "A Brief Introduction to Graphical Models and Bayesian Networks," [Online], 1998 [cited Oct. 26, 2003], Available: http://www.ai.mit.edu/~murphyk/Bayes/bayes.html

[11] W.A. Prudhomme, G.Q. Daley, .P.W. Zandstra, D.A. Lauffenberger, “Multivariate proteomic analysis of murine embryonic stem cell self-renewal versus differentiation signaling,” *Proc Natl Acad Sci U S A*, vol. 101(9), pp. 2900-2905, Mar. 2004.

[12] P.J. Woolf, W.A. Prudhomme, L. Daheron, G.Q. Daley, D.A. Lauffenburger. (2004, Oct). “Bayesian analysis of signaling networks governing embryonic stem cell fate decisions”. *Bioinformatics* [Online], Available:http://bioinformatics.oupjournals.org/cgi/reprint/bti056v1.pdf